



- Expert Verified, Online, **Free**.



CERTIFICATION TEST

- CertificationTest.net - Cheap & Quality Resources With Best Support

A large mobile network operating company is building a machine learning model to predict customers who are likely to unsubscribe from the service. The company plans to offer an incentive for these customers as the cost of churn is far greater than the cost of the incentive. The model produces the following confusion matrix after evaluating on a test dataset of 100 customers:

n= 100 ACTUAL Churn Yes Actual No	PREDICTED CHURN	
	Yes	No
	10	4
	10	76

Based on the model evaluation results, why is this a viable model for production?

- A. The model is 86% accurate and the cost incurred by the company as a result of false negatives is less than the false positives.
- B. The precision of the model is 86%, which is less than the accuracy of the model.
- C. The model is 86% accurate and the cost incurred by the company as a result of false positives is less than the false negatives.
- D. The precision of the model is 86%, which is greater than the accuracy of the model.

Suggested Answer: A

Community vote distribution

A (73%)

C (27%)

 **tgaos** Highly Voted 3 years, 1 month ago

The Answer is A.

Reasons:

- accurate is 86%
- FN=4, FP= 10. The question is asking why this is a feasible model which means why this is working. So it is not asking the explanation of the unit cost of churn(FN) is greater than cost of incentive(FP). It is asking from the matrix result, the number itself, FN(4) is less than FP(10). The model successfully keep a smaller number of FN regarding of FP.

upvoted 30 times

 **JK_314** Highly Voted 3 years, 8 months ago

Such question cannot be answered because we do not know how much more is greater the cost of churn than the cost of the incentive.

CoC - Cost of Churn

Col - Cost of Incentive

cost incurred by the company as a result of false positives = Col * 10

cost incurred by the company as a result of false negatives = CoC * 4

So is it the case that Col * 10 > CoC * 4 => Col > 0.4 * CoC, or rather Col < 0.4 * CoC? We don't know that because we don't know what does it mean "far greater", is it 100% greater, or is it 500% greater or any other number.

upvoted 9 times

 **Robertwilliamm** Most Recent 4 days, 4 hours ago

Selected Answer: A

A is Correct Option

Thanks to SkillCertExams I successfully cleared my MLS-C01 exam today.


upvoted 1 times

 **zWareZ** 4 weeks, 1 day ago

Selected Answer: C

FP is 10 and FN is 4, however, since the cost of FP is far less than FN, we can use this model. That's exactly what C said.

upvoted 1 times

 **Fa1ve** 1 month, 1 week ago

Selected Answer: C

Options A and C are a tough choice. However, we should also pay attention to the question prompt "Why is this a viable model for production?" This means we're looking for a justification that supports the model's use. From this perspective, only option C provides a correct number and a reason in favor of the model.

upvoted 1 times

🗨️ 👤 **robctsgps** 1 month, 3 weeks ago

Selected Answer: C

the answer is C, but A is tempting and a classic AWS trick question
upvoted 1 times

🗨️ 👤 **sarutc** 2 months, 3 weeks ago

Selected Answer: A

The answer is A
upvoted 2 times

🗨️ 👤 **sfwewv** 4 months, 2 weeks ago

Selected Answer: C

The Answer is c
upvoted 3 times

🗨️ 👤 **6dc4e56** 4 months, 2 weeks ago

Selected Answer: C

Even though there are 10 false positives compared to 4 false negatives, the cost incurred by offering an incentive unnecessarily (false positive) is significantly less than the cost of losing a customer (false negative). This risk management aligns well with the company's strategy to minimize expensive churn events.

Thus, the model is viable for production because it achieves 86% accuracy and, importantly, the cost of false positives (incentives given) is much lower than the cost associated with false negatives (lost customers).

upvoted 4 times

🗨️ 👤 **d2c29a3** 5 months ago

Selected Answer: C

Option C is indeed the correct choice. The model is 86% accurate, and the cost of false positives (offering incentives) is less than the cost of false negatives (losing customers). This makes the model viable for production.

upvoted 3 times

🗨️ 👤 **2bc8f6c** 5 months, 1 week ago

Selected Answer: C

Changing my earlier Answer from A to C. Cost of FP(10) is lower than Cost of FN(4)

upvoted 3 times

🗨️ 👤 **diblas** 5 months, 1 week ago

Selected Answer: C

some people that voted A have the right idea, but they chose the wrong option because they need to read the question again. We all agree that the cost of churn is much higher. So a false-negative means a customer churned and you didn't do anything about it (because your model said "churn=no") . A false positive means you tried to keep a customer that was not going to leave anyway (because your model said "churn=yes"). As you can see, false-negative is way costlier and should be avoided, therefore answer is C.

upvoted 4 times

🗨️ 👤 **2bc8f6c** 5 months, 2 weeks ago

Selected Answer: A

Cost incurred for churn higher than incentive. Cost of FN is higher than FP. And accuracy is 86%.

upvoted 1 times

🗨️ 👤 **4bc91ae** 5 months, 3 weeks ago

Selected Answer: C

what tomatoteacher said

upvoted 2 times

🗨️ 👤 **587df71** 5 months, 4 weeks ago

Selected Answer: C

Accuracy is 86% and it should be A or C. Lost is very high compare to intensive. Means it is Okay to give intensive to customers who are not going to leave. Which means False positives potion.


upvoted 3 times

🗨️ 👤 **Antoh1978** 9 months, 1 week ago

Selected Answer: A

Should be A. Since the cost of churn is much higher, the priority should be focused on minimizing FN and a viable model should be one with $FN < FP$, isn't it?

upvoted 2 times

  **Tomatoteacher** 9 months, 1 week ago

Selected Answer: C

Definitely C. If you look at the same question in <https://aws.amazon.com/blogs/machine-learning/predicting-customer-churn-with-amazon-machine-learning/>. Same question, but the confusion matrix is flipped in this case(TP top left, Tn bottom right) . When you miss an actual churn (FN) this would cost the company more. Therefore the answer is C 100%. I will die on this hill. I spent 20 minutes researching this to be certain. Most people who put A are incorrectly saying FPs are actual churns that are stated as no churn.. that is what a FN is. You can trust me on this.

upvoted 5 times

A Machine Learning Specialist is designing a system for improving sales for a company. The objective is to use the large amount of information the company has on users' behavior and product preferences to predict which products users would like based on the users' similarity to other users.

What should the Specialist do to meet this objective?

- A. Build a content-based filtering recommendation engine with Apache Spark ML on Amazon EMR
- B. Build a collaborative filtering recommendation engine with Apache Spark ML on Amazon EMR.
- C. Build a model-based filtering recommendation engine with Apache Spark ML on Amazon EMR
- D. Build a combinative filtering recommendation engine with Apache Spark ML on Amazon EMR

Suggested Answer: B

Many developers want to implement the famous Amazon model that was used to power the “People who bought this also bought these items” feature on

Amazon.com. This model is based on a method called Collaborative Filtering. It takes items such as movies, books, and products that were rated highly by a set of users and recommending them to other users who also gave them high ratings. This method works well in domains where explicit ratings or implicit user actions can be gathered and analyzed.

Reference:

<https://aws.amazon.com/blogs/big-data/building-a-recommendation-engine-with-spark-ml-on-amazon-emr-using-zeppelin/>


Community vote distribution

B (100%)

  **mlyu**  3 years, 9 months ago

B



see https://en.wikipedia.org/wiki/Collaborative_filtering#Model-based
upvoted 21 times

  **kalyanvarma**  3 years, 7 months ago

Content-based filtering relies on similarities between features of items, whereas collaborative-based filtering relies on preferences from other users and how they respond to similar items.
upvoted 14 times

  **Manju_Bn**  8 months, 2 weeks ago

Answer is B : Build a collaborative filtering recommendation engine with Apache Spark ML on Amazon EMR.
Collaborative filtering focuses on user behavior and preferences therefore it is perfect for predicting products based on user similarities.
upvoted 2 times

  **Ajose0** 9 months, 1 week ago

Selected Answer: B

B. Build a collaborative filtering recommendation engine with Apache Spark ML on Amazon EMR.

Collaborative filtering is a technique used to recommend products to users based on their similarity to other users. It is a widely used method for building recommendation engines. Apache Spark ML is a distributed machine learning library that provides scalable implementations of collaborative filtering algorithms. Amazon EMR is a managed cluster platform that provides easy access to Apache Spark and other distributed computing frameworks.

upvoted 1 times

  **solution123** 9 months, 1 week ago

Selected Answer: B

Build a collaborative filtering recommendation engine with Apache Spark ML on Amazon EMR. (TRUE)

Collaborative filtering is a commonly used method for recommendation systems that aims to predict the preferences of a user based on the behavior of similar users. In the case described, the objective is to use users' behavior and product preferences to predict which products they want, making collaborative filtering a good fit.

Apache Spark ML is a machine learning library that provides scalable, efficient algorithms for building recommendation systems, while Amazon EMR

provides a cloud-based platform for running Spark applications.

You can find more detail in <https://www.udemy.com/course/aws-certified-machine-learning-specialty-2023>

upvoted 2 times

🗳️ 👤 **ychaabane** 9 months, 2 weeks ago

Selected Answer: B

collaborative filtering

upvoted 1 times

🗳️ 👤 **james2033** 1 year, 3 months ago

Selected Answer: B

'Collaborative filtering is a technique that can filter out items that a user might like on the basis of reactions by similar users.'

Source: <https://realpython.com/build-recommendation-engine-collaborative-filtering/#what-is-collaborative-filtering>

upvoted 1 times

🗳️ 👤 **loict** 1 year, 9 months ago

Selected Answer: B

A. NO - content-based filtering looks at similarities with items the user already looked at, not activities of other users

B. YES - state of the art

C. NO - too generic terms, everything is a model

D. NO - combinative filtering does not exist

upvoted 4 times

🗳️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: B

Collaborative filtering is a technique used by recommendation engines to make predictions about the interests of a user by collecting preferences or taste information from many users. The underlying assumption of the collaborative filtering approach is that if a person A has the same opinion as a person B on an issue, A is more likely to have B's opinion on a different issue than that of a randomly chosen person.

upvoted 2 times

🗳️ 👤 **Mickey321** 1 year, 11 months ago

Selected Answer: B

B. Build a collaborative filtering recommendation engine with Apache Spark ML on Amazon EMR.

upvoted 1 times

🗳️ 👤 **Venkatesh_Babu** 1 year, 11 months ago

Selected Answer: B

I think it should be b

upvoted 1 times

🗳️ 👤 **brunokiyoshi** 2 years, 3 months ago

Selected Answer: B

Content-based recommendations rely on product similarity. If a user likes a product, products that are similar to that one will be recommended.

Collaborative recommendations are based on user similarity. If you and other users have given similar reviews to a range of products, the model assumes it is likely that other products those other people have liked but that you haven't purchased should be a good recommendation for you.

upvoted 4 times

🗳️ 👤 **dreswardev** 2 years, 6 months ago

feature engineering is required, use model based

upvoted 1 times

🗳️ 👤 **ryuhei** 2 years, 9 months ago

Selected Answer: B

Answer is "B"

upvoted 1 times

🗳️ 👤 **roytruong** 3 years, 8 months ago

go for B

upvoted 2 times

🗳️ 👤 **cybe001** 3 years, 8 months ago

B is correct

<https://aws.amazon.com/blogs/big-data/building-a-recommendation-engine-with-spark-ml-on-amazon-emr-using-zeppelin/>

upvoted 6 times

A Mobile Network Operator is building an analytics platform to analyze and optimize a company's operations using Amazon Athena and Amazon S3.

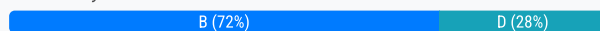
The source systems send data in .CSV format in real time. The Data Engineering team wants to transform the data to the Apache Parquet format before storing it on Amazon S3.

Which solution takes the LEAST effort to implement?

- A. Ingest .CSV data using Apache Kafka Streams on Amazon EC2 instances and use Kafka Connect S3 to serialize data as Parquet
- B. Ingest .CSV data from Amazon Kinesis Data Streams and use Amazon Glue to convert data into Parquet.
- C. Ingest .CSV data using Apache Spark Structured Streaming in an Amazon EMR cluster and use Apache Spark to convert data into Parquet.
- D. Ingest .CSV data from Amazon Kinesis Data Streams and use Amazon Kinesis Data Firehose to convert data into Parquet.

Suggested Answer: B

Community vote distribution



DonaldCMLIN Highly Voted 3 years, 9 months ago

Answer is B

upvoted 31 times

Antriksh 3 years, 8 months ago

you cannot use AWS glue for streaming data. Clearly B is incorrect.

upvoted 3 times

scuzzy2010 3 years, 8 months ago

Even if the exam's answer is based on solution before AWS implemented the capability of AWS glue to process streaming data, this answer is still correct as Kinesis would output the data to S3 and Glue will pick it up from there and covert to parquet. Question does not say data must be converted to parquet in real time, it only says the csv data is received as a stream in real time.

upvoted 2 times

GeeBeeEI 3 years, 8 months ago

Actually question says "The source systems send data in CSV format in real time The Data Engineering team wants to transform the data to the Apache Parquet format before storing it on Amazon S3" same as saying data must be converted real time

upvoted 5 times

zzeng 3 years, 8 months ago

AWS Glue can do it now (2020 May)

<https://aws.amazon.com/jp/blogs/news/new-serverless-streaming-etl-with-aws-glue/>

upvoted 6 times

hamimelon 2 years, 6 months ago

This link is in Japanese

upvoted 3 times

OmarSaadEldien 3 years, 8 months ago

the Approve Of B

<https://aws.amazon.com/blogs/aws/new-serverless-streaming-etl-with-aws-glue/>

upvoted 7 times

vetal Highly Voted 3 years, 9 months ago

D is wrong as kinesis firehose can convert from JSON to parquet but here we have CSV.

B is correct and here is another proof link: <https://medium.com/searce/convert-csv-json-files-to-apache-parquet-using-aws-glue-a760d177b45f>

upvoted 24 times

zzeng 3 years, 8 months ago

<https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html>

You are right.

<https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html>

If you want to convert an input format other than JSON, such as comma-separated values (CSV) or structured text, you can use AWS Lambda to transform it to JSON first

upvoted 8 times

  **samy666** 3 years, 1 month ago

But there is no Lambda in D

upvoted 2 times

  **AdolinKholin** 2 years, 9 months ago

But there's a D in Lambda

upvoted 3 times

  **daveclear** Most Recent 2 months ago

Selected Answer: B

Answer is B

- Firehose cannot convert from csv to parquet without a lambda: <https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html>

- Glue can handle streaming data: <https://docs.aws.amazon.com/glue/latest/dg/add-job-streaming.html>

upvoted 2 times

  **LalBSingh** 4 months, 1 week ago

Selected Answer: D



Kinesis Data Firehose supports real-time streaming ingestion and can automatically convert CSV to Parquet before storing it in S3.

upvoted 3 times

  **daveclear** 2 months ago

It requires a lambda to go from csv to parquet

upvoted 1 times

  **JonSno** 4 months, 2 weeks ago

Selected Answer: D

Amazon Kinesis Data Streams + Amazon Kinesis Data Firehose

Effort: Lowest effort

Why?

Amazon Kinesis Data Firehose natively supports real-time CSV ingestion and automatic conversion to Parquet.

Fully managed, serverless, and directly integrates with Amazon S3.

Requires zero infrastructure management compared to other solutions.



upvoted 1 times

  **JonSno** 4 months, 2 weeks ago

I take this back .. ans shd be B.. on researching further it is JSON or ORC to Parquet that KDS supports.. So answer is B - not optimal but close to suitable

. Amazon Kinesis Data Streams + AWS Glue AWS Glue can batch-process CSV and convert it to Parquet for S3. However, Glue is batch-oriented, not real-time.

upvoted 1 times

  **liquen14** 4 months, 2 weeks ago

Selected Answer: B

Although I'd go with Glue and option B I'm pretty sure that this is one of those "15 unscored questions that do not affect your score. AWS collects information about performance on these unscored questions to evaluate these questions for future use as scored questions"

Just for fun I asked perplexity, chatgpt, gemini, deepseek and claude: all gave D as first response

When I pointed out that "according to this <https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html> Kinesis can't convert directly cvs to parquet. It needs a Lambda" each model responded in a different way (some of them contradictory).

My reasoning is that D (Kinesis + Firehose) is incorrect because Firehose does not support direct CSV-to-Parquet conversion and needs a Lambda not mentioned in the option. But discussing about questions like this one is nothing but a big waste of time ;P

upvoted 2 times

  **AbimbolaOlaniran** 6 months, 1 week ago

Selected Answer: D

D

Kinesis Data Firehose is designed specifically for streaming data delivery to destinations like S3. It has built-in support for data format conversion,

including CSV to Parquet. This eliminates the need for managing separate transformation services like Glue or Spark. The setup is significantly simpler: you configure a Firehose delivery stream, specify the data format conversion, and point it to your S3 bucket.

Therefore, option D requires the least implementation effort because it leverages a fully managed service (Kinesis Data Firehose) with built-in functionality for data format conversion.

upvoted 1 times

🗳️ 👤 **venksters** 6 months, 2 weeks ago

Selected Answer: B

Amazon Kinesis Data Firehose can only convert from JSON to Apache Parquet or Apache ORC before storing the data in Amazon S3.

upvoted 1 times

🗳️ 👤 **TinTinAWS** 9 months ago

Answer B,

Yes, Amazon Kinesis Data Firehose can convert CSV to Apache Parquet, but you need to use a Lambda function to transform the CSV to JSON first: here the question is least effort to build, so B is the right answer with least effort to build the solution

upvoted 1 times

🗳️ 👤 **Keya** 9 months, 1 week ago

Selected Answer: B

Use Amazon Kinesis Data Streams to ingest customer data and configure a Kinesis Data Firehose delivery stream as a consumer to convert the data into Apache Parquet is incorrect. Although this could be a valid solution, it entails more development effort as Kinesis Data Firehose does not support converting CSV files directly into Apache Parquet, unlike JSON.

upvoted 2 times

🗳️ 👤 **geoan13** 9 months, 1 week ago

Selected Answer: B

Amazon Kinesis Data Firehose can convert the format of your input data from JSON to Apache Parquet or Apache ORC before storing the data in Amazon S3. Parquet and ORC are columnar data formats that save space and enable faster queries compared to row-oriented formats like JSON. If you want to convert an input format other than JSON, such as comma-separated values (CSV) or structured text, you can use AWS Lambda to transform it to JSON first.

<https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html>

upvoted 1 times

🗳️ 👤 **rav009** 1 year, 1 month ago

Selected Answer: D

Between B and D chose D.

Because Firehose can't handle csv directly.

upvoted 1 times

🗳️ 👤 **rav009** 1 year, 1 month ago

Between B and D chose B.

Because Firehose can't handle csv directly.

upvoted 1 times

🗳️ 👤 **s_k_aws** 1 year, 3 months ago

Answer is B.

<https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html>

"If you want to convert an input format other than JSON, such as comma-separated values (CSV) or structured text, you can use AWS Lambda to transform it to JSON first."

upvoted 1 times

🗳️ 👤 **chewasa** 1 year, 3 months ago

Selected Answer: B

u need glue to convert to parquet

upvoted 1 times

🗳️ 👤 **0c47783** 1 year, 4 months ago

D for sure, Firehose can convert csv to parquet

upvoted 3 times

🗳️ 👤 **vkbajoria** 1 year, 4 months ago

Answer is unfortunately B. firehose cannot convert coma separated CSV to parquet directly.

upvoted 1 times

🗨️ 👤 **kyuhuck** 1 year, 4 months ago

Selected Answer: D

b is not goog but - >given the context of "finding the solution that requires the least effort to implement," option D is the most suitable choice. Ingesting data from Amazon Kinesis Data Streams and using Amazon Kinesis Data Firehose to convert the data to Parquet format is a serverless approach. It allows for automatic data transformation and storage in Amazon S3 without the need for additional development or management of data conversion logic. Therefore, under the given conditions, option D is considered the solution that requires the "least effort" to implement
upvoted 3 times

🗨️ 👤 **shammous** 10 months, 4 weeks ago

Kinesis Data Firehose doesn't convert anything, it rather calls a lambda function to do so which is the overhead we want to avoid. B is the correct answer.

upvoted 1 times

A city wants to monitor its air quality to address the consequences of air pollution. A Machine Learning Specialist needs to forecast the air quality in parts per million of contaminants for the next 2 days in the city. As this is a prototype, only daily data from the last year is available. Which model is MOST likely to provide the best results in Amazon SageMaker?

- A. Use the Amazon SageMaker k-Nearest-Neighbors (kNN) algorithm on the single time series consisting of the full year of data with a predictor_type of regressor.
- B. Use Amazon SageMaker Random Cut Forest (RCF) on the single time series consisting of the full year of data.
- C. Use the Amazon SageMaker Linear Learner algorithm on the single time series consisting of the full year of data with a predictor_type of regressor.
- D. Use the Amazon SageMaker Linear Learner algorithm on the single time series consisting of the full year of data with a predictor_type of classifier.

Suggested Answer: C

Reference:

<https://aws.amazon.com/blogs/machine-learning/build-a-model-to-predict-the-impact-of-weather-on-urban-air-quality-using-amazon-sagemaker/?ref=Welcome.AI>

Community vote distribution



ozan11 Highly Voted 3 years, 9 months ago

answer should be C
upvoted 17 times

roytruong Highly Voted 3 years, 8 months ago

go for C
upvoted 6 times

robctsgps Most Recent 1 month, 3 weeks ago

Selected Answer: C

kNN is not specifically designed for time series forecasting. The best choice is C. Use the Amazon SageMaker Linear Learner algorithm on the single time series consisting of the full year of data with a predictor_type of regressor.
upvoted 1 times

JonSno 4 months, 2 weeks ago

Selected Answer: C

Amazon SageMaker Linear Learner (Regressor)
Why?

The Linear Learner algorithm can be used for time series regression.

Using predictor_type=regressor, it learns trends and patterns in historical data and extrapolates future values.

Given limited historical data (only 1 year), a simple linear regression model might perform well as a baseline.

While deep learning models (like Amazon Forecast) may be more advanced, Linear Learner is easier to implement and train for a prototype.

upvoted 2 times

loict 9 months, 1 week ago

Selected Answer: C

- A. NO - kNN is not forecasting, it is similarities
 - B. NO - RCF is for anomaly detection
 - C. YES - Linear Regression good for forecasting
 - D. NO - we don't want to classify
- upvoted 3 times

Mickey321 9 months, 1 week ago

Selected Answer: C

The reason for this choice is that the Linear Learner algorithm is a versatile algorithm that can be used for both regression and classification tasks1. Regression is a type of supervised learning that predicts a continuous numeric value, such as the air quality in parts per million2. The predictor_type

parameter specifies whether the algorithm should perform regression or classification³. Since the goal is to forecast a numeric value, the `predictor_type` should be set to `regressor`.

upvoted 3 times

🗨️ **ninomfr64** 1 year ago

Selected Answer: D

- A. Managing Kafka on EC2 is not compatible with least effort requirement
- B. Doable (in 2024) as Glue supports streaming ETL to consumes streams and supports CSV records -> <https://docs.aws.amazon.com/glue/latest/dg/add-job-streaming.html>
- C. Managing an EMR cluster imo is no compatible with least effort requirement
- D. Firehose supports kinesis data stream as source and it can use lambda to convert CSV records into parquet -> <https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html>

I guess this is a bit old question, pre Glue streaming ETL support (2023) -> <https://aws.amazon.com/about-aws/whats-new/2023/03/aws-glue-4-0-streaming-etl/>

Thus I'll go for D

upvoted 1 times

🗨️ **LocalHero** 1 year, 7 months ago

This blog wrote Japanese.

but its said using LinearLearner for air pollution prediction.

<https://aws.amazon.com/jp/blogs/news/build-a-model-to-predict-the-impact-of-weather-on-urban-air-quality-using-amazon-sagemaker/>

upvoted 2 times

🗨️ **jyrajan69** 1 year, 11 months ago

The HyperParameter is . Either "binary_classifier" or "multiclass_classifier" or "regressor"., there is no classifier so the answer is C

upvoted 1 times

🗨️ **Venkatesh_Babu** 1 year, 11 months ago

Selected Answer: C

Ans should be c

upvoted 1 times

🗨️ **ortamina** 1 year, 11 months ago

a kNN will require a large value of k to avoid overfitting and we only have 1 year's worth of data - kNNs also face a difficult time extrapolating if the air quality series contains a trend

If we had assurances there is no trend in the air quality series (no extrapolation), and we had enough data, then kNN should beat a linear model ... I am inclined to go for C just going off of the cue that "only daily data from last year is available"

upvoted 1 times

🗨️ **ninomfr64** 1 year ago

Agree with you analysis, to further expand it: we don't have info about dataset features based on "only daily data from last year is available" this let me think we could be in a situation where our dataset is made up by timestamp and pollution_value so KNN would be pretty useless in this situation.

upvoted 1 times

🗨️ **brunokiyoshi** 2 years, 3 months ago

Selected Answer: C

Random cut forests in timeseries are used for anomaly detection, and not for forecasting. KNN's are classification algorithms. You would use the Linear Learner as a regressor, since forecasting falls into the domain of regression.

upvoted 3 times

🗨️ **brunokiyoshi** 2 years, 3 months ago

I mean, you could use KNN's for regression, but for forecasting I don't think so

upvoted 1 times

🗨️ **Valcilio** 2 years, 3 months ago

Selected Answer: C

KNN isn't for time series predicting, go for A!

upvoted 2 times

🗨️ **Valcilio** 2 years, 3 months ago

Im sorry, I wanted to say go for C!

upvoted 2 times



  **rockyykrish** 2 years, 3 months ago

Creating a machine learning model to predict air quality

To start small, we will follow the second approach, where we will build a model that will predict the NO2 concentration of any given day based on wind speed, wind direction, maximum temperature, pressure values of that day, and the NO2 concentration of the previous day. For this we will use the Linear Learner algorithm provided in Amazon SageMaker, enabling us to quickly build a model with minimal work.

Our model will consist of taking all of the variables in our dataset and using them as features of the Linear Learner algorithm available in Amazon SageMaker

upvoted 1 times

  **Ajose0** 2 years, 4 months ago


Selected Answer: A

Answer should be A.

k-Nearest-Neighbors (kNN) algorithm will provide the best results for this use case as it is a good fit for time series data, especially for predicting continuous values. The predictor_type of regressor is also appropriate for this task, as the goal is to forecast a continuous value (air quality in parts per million of contaminants). The other options are also viable, but may not provide as good of results as the kNN algorithm, especially with limited data.

using the Amazon SageMaker Linear Learner algorithm with a predictor_type of regressor, may still provide reasonable results, but it assumes a linear relationship between the input features and the target variable (air quality), which may not always hold in practice, especially with complex time series data. In such cases, non-linear models like kNN may perform better. Furthermore, the kNN algorithm can handle irregular patterns in the data, which may be present in the air quality data, and provide more accurate predictions.

upvoted 3 times

  **ryuhei** 2 years, 9 months ago

Selected Answer: C

Answer is "C" !!!

upvoted 1 times

  **yemauricio** 2 years, 9 months ago

answer C

upvoted 1 times

A Data Engineer needs to build a model using a dataset containing customer credit card information
How can the Data Engineer ensure the data remains encrypted and the credit card information is secure?

- A. Use a custom encryption algorithm to encrypt the data and store the data on an Amazon SageMaker instance in a VPC. Use the SageMaker DeepAR algorithm to randomize the credit card numbers.
- B. Use an IAM policy to encrypt the data on the Amazon S3 bucket and Amazon Kinesis to automatically discard credit card numbers and insert fake credit card numbers.
- C. Use an Amazon SageMaker launch configuration to encrypt the data once it is copied to the SageMaker instance in a VPC. Use the SageMaker principal component analysis (PCA) algorithm to reduce the length of the credit card numbers.
- D. Use AWS KMS to encrypt the data on Amazon S3 and Amazon SageMaker, and redact the credit card numbers from the customer data with AWS Glue.

Suggested Answer: D

Community vote distribution

D (100%)

vetal **Highly Voted** 3 years, 9 months ago

Why not D? When the data encrypted on S3 and SageMaker uses the same AWS KMS key it can use encrypted data there.
upvoted 36 times

WWODIN 3 years, 9 months ago

should be D
upvoted 12 times

zzeng 3 years, 8 months ago

Should be D.
Use Glue to do ETL to Hash the card number
upvoted 8 times

Antriksh 3 years, 8 months ago

Answer would be D
upvoted 9 times

cybe001 **Highly Voted** 3 years, 9 months ago

D is correct
upvoted 8 times

Ganshank **Most Recent** 4 months ago

Selected Answer: D
<https://aws.amazon.com/blogs/big-data/detect-and-process-sensitive-data-using-aws-glue-studio/>
AWS Glue can be used for detecting and processing sensitive data.
upvoted 2 times

JonSno 4 months, 2 weeks ago

Selected Answer: D
Use AWS KMS for encryption and AWS Glue to redact credit card numbers
Reasoning:
AWS KMS (Key Management Service) encrypts data at rest in Amazon S3 and during processing in Amazon SageMaker.
AWS Glue can be used to redact sensitive data before processing, ensuring that credit card numbers are removed from datasets before being used for ML.
Complies with PCI DSS requirements for handling payment information securely.
upvoted 2 times

Mickey321 9 months, 1 week ago

Selected Answer: D
The reason for this choice is that AWS KMS is a service that allows you to easily create and manage encryption keys and control the use of encryption across a wide range of AWS services and in your applications¹. By using AWS KMS, you can encrypt the data on Amazon S3, which is a

durable, scalable, and secure object storage service², and on Amazon SageMaker, which is a fully managed service that provides every developer and data scientist with the ability to build, train, and deploy machine learning models quickly³. This way, you can protect the data at rest and in transit.

upvoted 2 times

🗨️ 👤 **loict** 1 year, 9 months ago

Selected Answer: D

- A. NO - no need for custom encryption
- B. NO - IAM Policies are not to encrypt
- C. NO - launch configuration is not to encrypt
- D. YES

upvoted 2 times

🗨️ 👤 **Venkatesh_Babu** 1 year, 11 months ago

Selected Answer: D

I think d is correct

upvoted 1 times

🗨️ 👤 **Valcilio** 2 years, 3 months ago

Selected Answer: D

It's D, KMS key can be used for encrypting the data at rest!

upvoted 4 times

🗨️ 👤 **ystotest** 2 years, 7 months ago

Selected Answer: D

agreed with D

upvoted 3 times

🗨️ 👤 **jerto97** 3 years, 7 months ago

IMHO, the problem with the question is that it is not clear whether the credit card number is used in the model. In that case discarding is never a good option. Hashing should be a safe option to keep it in the learning path

upvoted 1 times

🗨️ 👤 **cloud_trail** 3 years, 8 months ago

It's gotta be D but C is a clever fake answer. Use PCA to reduce the length of the credit card number? That's a clever joke, as if reducing the length of a character string is the same as reducing dimensionality in a feature set.

upvoted 3 times

🗨️ 👤 **cnethers** 3 years, 8 months ago

Can Glue do redaction?

upvoted 1 times

🗨️ 👤 **cloud_trail** 3 years, 7 months ago

Just have the Glue job remove the credit card column.

upvoted 1 times

🗨️ 👤 **syu31svc** 3 years, 8 months ago

Encryption on AWS can be done using KMS so D is the answer

upvoted 1 times

🗨️ 👤 **roytruong** 3 years, 8 months ago

D is correct

upvoted 1 times

🗨️ 👤 **PRC** 3 years, 8 months ago

D..KMS fully managed and other options are too whacky..

upvoted 4 times

🗨️ 👤 **AKT** 3 years, 8 months ago

D is correct

upvoted 1 times

🗨️ 👤 **bhavesh0124** 3 years, 9 months ago

Ans D is correct

upvoted 2 times

A Machine Learning Specialist is using an Amazon SageMaker notebook instance in a private subnet of a corporate VPC. The ML Specialist has important data stored on the Amazon SageMaker notebook instance's Amazon EBS volume, and needs to take a snapshot of that EBS volume. However, the ML Specialist cannot find the Amazon SageMaker notebook instance's EBS volume or Amazon EC2 instance within the VPC. Why is the ML Specialist not seeing the instance visible in the VPC?

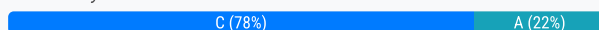
- A. Amazon SageMaker notebook instances are based on the EC2 instances within the customer account, but they run outside of VPCs.
- B. Amazon SageMaker notebook instances are based on the Amazon ECS service within customer accounts.
- C. Amazon SageMaker notebook instances are based on EC2 instances running within AWS service accounts.
- D. Amazon SageMaker notebook instances are based on AWS ECS instances running within AWS service accounts.

Suggested Answer: C

Reference:

<https://docs.aws.amazon.com/sagemaker/latest/dg/gs-setup-working-env.html>

Community vote distribution



mlyu Highly Voted 3 years, 9 months ago

I think the answer should be C
upvoted 24 times

dhs227 Highly Voted 3 years, 8 months ago

The correct answer HAS TO be A

The instances are running in customer accounts but it's in an AWS managed VPC while exposing ENI to customer VPC if it was chosen. See explanation at <https://aws.amazon.com/blogs/machine-learning/understanding-amazon-sagemaker-notebook-instance-networking-configurations-and-advanced-routing-options/>
upvoted 18 times

scuzzy2010 3 years, 8 months ago

Can't be A because A says "but they run outside of VPCs", which is not correct. They are attached to VPC, but it can either be AWS Service VPC or Customer VPC, or Both, as per the explanation url you provided.
upvoted 10 times

cloud_trail 3 years, 7 months ago

This is exactly right. According to that document, if the notebook instance is not in a customer VPC, then it has to be in the Sagemaker managed VPC. See Option 1 in that document.
upvoted 1 times

mawsmann 3 years, 8 months ago

Actually your link says: The notebook instance is running in an Amazon SageMaker managed VPC as shown in the above diagram. That means the correct answer is C. An Amazon SageMaker managed VPC can only be created in an Amazon managed Account.
upvoted 18 times

JonSno Most Recent 4 months, 2 weeks ago

Selected Answer: C

C. Amazon SageMaker notebook instances are based on EC2 instances running within AWS service accounts.
Why?

Amazon SageMaker does use EC2 instances, but they are not directly managed within the customer's AWS account.

Instead, these instances are provisioned within AWS-managed service accounts, which is why they do not appear within the customer's VPC or EC2 console.

The only way to access the underlying EBS volume is via SageMaker APIs, rather than the EC2 console.

upvoted 5 times

liquen14 4 months, 2 weeks ago

Selected Answer: B



Although I'd go with Glue and option B I'm pretty sure that this is one of those "15 unscored questions that do not affect your score. AWS collects information about performance on these unscored questions to evaluate these questions for future use as scored questions"

Just for fun I asked perplexity, chatgpt, gemini, deepseek and claude: all gave D as first response

When I pointed out that "according to this <https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html> Kinesis can't convert directly cvs to parquet. It needs a Lambda" each model responded in a different way (some of them contradictory).

My reasoning is that D (Kinesis + Firehose) is incorrect because Firehose does not support direct CSV-to-Parquet conversion and needs a Lambda not mentioned in the option. But discussing about questions like this one is nothing but I big waste of time ;-P

upvoted 1 times

  **liquen14** 4 months, 2 weeks ago

Forget about this please I posted this here incorrectly. This corresponds to Question 3. Apologies

upvoted 1 times

  **reginav** 6 months, 2 weeks ago

Selected Answer: A

Amazon SageMaker notebook instances are indeed based on EC2 instances, but they are managed by the SageMaker service and do not appear as standard EC2 instances in the customer's VPC. Instead, they run in a managed environment that abstracts away the underlying EC2 instances, which is why the ML Specialist cannot see the instance in the VPC.

upvoted 1 times

  **Mickey321** 9 months, 1 week ago

Selected Answer: C

The explanation for this choice is that Amazon SageMaker notebook instances are fully managed by AWS and run on EC2 instances that are not visible to customers. These EC2 instances are launched in AWS-owned accounts and are isolated from customer accounts by using AWS PrivateLink1. This means that customers cannot access or manage these EC2 instances directly, nor can they see the EBS volumes attached to them.


upvoted 1 times

  **loict** 9 months, 1 week ago

Selected Answer: C

- A. NO - AEC2 instances within the customer account are necessarily in a VPCb
- B. NO - Amazon ECS service is not within customer accounts
- C. YES - EC2 instances running within AWS service accounts are not visible to customer account
- D. NO - SageMaker manages EC2 instance, not ECS

upvoted 6 times



  **ninomfr64** 9 months, 1 week ago

Selected Answer: C

- A. NO. If the EC2 instance of the notebook was in the customer account, customer would be able to see it. Also, "they run outside VPCs" isn't true as they run in service managed VPC or can be also attached to customer provided VPC -> <https://aws.amazon.com/blogs/machine-learning/understanding-amazon-sagemaker-notebook-instance-networking-configurations-and-advanced-routing-options/>
- B. NO, Notebooks are based on EC2 + EBS
- C. YES -> <https://aws.amazon.com/blogs/machine-learning/understanding-amazon-sagemaker-notebook-instance-networking-configurations-and-advanced-routing-options/>
- D. NO, Notebooks are based on EC2 + EBS

I also actually tested it in my account: I created a Notebook and attached it to my VPC, I was not able to see the EC2 instance behind the Notebook but I was able to see the its ENI with the following description "[Do not delete] Network Interface created to access resources in your VPC for SageMaker Notebook Instance ..."



upvoted 1 times

  **Rejju** 1 year, 9 months ago

Selected Answer: A

already given below

upvoted 1 times

  **Rejju** 1 year, 9 months ago

I am pretty sure the answer is A : Amazon SageMaker notebook instances are indeed based on EC2 instances, and these instances are within your AWS customer account. However, by default, SageMaker notebook instances run outside of your VPC (Virtual Private Cloud), which is why they may not be visible within your VPC. SageMaker instances are designed to be easily accessible for data science and machine learning tasks, which is why

they typically do not reside within a VPC. If you need them to operate within a VPC, you can configure them accordingly, but this is not the default behavior.

upvoted 1 times

🗨️ 👤 **Venkatesh_Babu** 1 year, 11 months ago

Selected Answer: C

I think it should be c

upvoted 1 times

🗨️ 👤 **ADVIT** 2 years ago

Per <https://docs.aws.amazon.com/sagemaker/latest/dg/studio-notebooks-and-internet-access.html> it's C

upvoted 1 times

🗨️ 👤 **BeCalm** 2 years, 1 month ago

Selected Answer: C

Notebooks can run inside AWS managed VPC or customer managed VPC

upvoted 1 times

🗨️ 👤 **Maaayaaa** 2 years, 2 months ago

Selected Answer: C

C, check the digram in <https://docs.aws.amazon.com/sagemaker/latest/dg/studio-notebooks-and-internet-access.html>

upvoted 5 times

🗨️ 👤 **oso0348** 2 years, 2 months ago

Selected Answer: A

When a SageMaker notebook instance is launched in a VPC, it creates an Elastic Network Interface (ENI) in the subnet specified, but the underlying EC2 instance is not visible in the VPC. This is because the EC2 instance is managed by AWS, and it is outside of the VPC. The ENI acts as a bridge between the VPC and the notebook instance, allowing network connectivity between the notebook instance and other resources in the VPC.

Therefore, the EBS volume of the notebook instance is also not visible in the VPC, and you cannot take a snapshot of the volume using VPC-based tools. Instead, you can create a snapshot of the EBS volume directly from the SageMaker console, AWS CLI, or SDKs.

upvoted 2 times

🗨️ 👤 **ZSun** 2 years, 1 month ago

what you described is C

"This is because the EC2 instance is managed by AWS, and it is outside of the VPC."

upvoted 1 times

🗨️ 👤 **Valcilio** 2 years, 3 months ago

Selected Answer: C

Notebooks run inside a VPC not outside!

upvoted 1 times

🗨️ 👤 **krzyhoo** 2 years, 4 months ago

Selected Answer: C

Definitely C

upvoted 2 times

A Machine Learning Specialist is building a model that will perform time series forecasting using Amazon SageMaker. The Specialist has finished training the model and is now planning to perform load testing on the endpoint so they can configure Auto Scaling for the model variant. Which approach will allow the Specialist to review the latency, memory utilization, and CPU utilization during the load test?

- A. Review SageMaker logs that have been written to Amazon S3 by leveraging Amazon Athena and Amazon QuickSight to visualize logs as they are being produced.
- B. Generate an Amazon CloudWatch dashboard to create a single view for the latency, memory utilization, and CPU utilization metrics that are outputted by Amazon SageMaker.
- C. Build custom Amazon CloudWatch Logs and then leverage Amazon ES and Kibana to query and visualize the log data as it is generated by Amazon SageMaker.
- D. Send Amazon CloudWatch Logs that were generated by Amazon SageMaker to Amazon ES and use Kibana to query and visualize the log data.

Suggested Answer: B

Reference:

<https://docs.aws.amazon.com/sagemaker/latest/dg/monitoring-cloudwatch.html>


Community vote distribution

B (100%)

 **mlyu** Highly Voted 3 years, 9 months ago

Agreed. Ans is B

upvoted 17 times

 **JonSno** Most Recent 4 months, 2 weeks ago

Selected Answer: B

Generate an Amazon CloudWatch dashboard to create a single view for latency, memory utilization, and CPU utilization
Why?

Amazon SageMaker automatically pushes latency and instance utilization metrics to CloudWatch.

CloudWatch dashboards provide a single real-time view of these key metrics during load testing.

You can configure custom CloudWatch alarms to trigger auto scaling based on the load.

upvoted 2 times

 **teka112233** 9 months, 1 week ago

Selected Answer: B

the question is clear that the specialist is seeking for latency, memory utilization, and CPU utilization during the load test and the ideal answer for all of these is amazon cloud watch which give you all these metrics

<https://docs.aws.amazon.com/sagemaker/latest/dg/monitoring-cloudwatch.html>

upvoted 2 times

 **Mickey321** 9 months, 1 week ago

Selected Answer: B

The reason for this choice is that Amazon CloudWatch is a service that monitors and manages your cloud resources and applications. It collects and tracks metrics, which are variables you can measure for your resources and applications¹. Amazon SageMaker automatically reports metrics such as latency, memory utilization, and CPU utilization to CloudWatch². You can use these metrics to monitor the performance and health of your SageMaker endpoint during the load test.

upvoted 1 times

 **teka112233** 1 year, 11 months ago

the question is clear that the specialist is seeking for latency, memory utilization, and CPU utilization during the load test and the ideal answer for all of these is amazon cloud watch which give you all these metrics

<https://docs.aws.amazon.com/sagemaker/latest/dg/monitoring-cloudwatch.html>

upvoted 1 times

 **Venkatesh_Babu** 1 year, 11 months ago

Selected Answer: B

I think it should be b

upvoted 1 times

🗨️ 👤 **Valcilio** 2 years, 3 months ago

Selected Answer: B

It's B, even the resources that aren't visible in a first try are visible if you use cloudwatch agent.

upvoted 3 times

🗨️ 👤 **DS2021** 2 years, 6 months ago

Selected Answer: B

Should be B

upvoted 2 times

🗨️ 👤 **ystotest** 2 years, 7 months ago

Selected Answer: B

agreed with B

upvoted 2 times

🗨️ 👤 **apprehensive_scar** 3 years, 4 months ago

B is the ans

upvoted 1 times

🗨️ 👤 **anttan** 3 years, 7 months ago

Should be C right, as Cloudwatch does not have metrics for memory utilization.

upvoted 2 times

🗨️ 👤 **anttan** 3 years, 6 months ago

After further research, I think answer is B. While indeed true that Cloudwatch does not have metrics for memory utilization by default, you can achieve by installing CloudWatch agent on the EC2. The EC2 used by Sagemaker is pre-installed with Cloudwatch Agent.

upvoted 2 times

🗨️ 👤 **[Removed]** 3 years, 7 months ago

I do not think that CloudWatch, by default, logs memory utilization. It does log CPU utilization. If memory utilization is required, then a separate agent needs to be installed to watch for memory. Hence, in this case, we have to write an agent if the answer has to be B. Else, C looks to be a better solution.

upvoted 2 times

🗨️ 👤 **Willnguyen22** 3 years, 8 months ago

answer is B

upvoted 1 times

🗨️ 👤 **syu31svc** 3 years, 8 months ago

Answer is B 100%; very straightforward method

upvoted 1 times

🗨️ 👤 **scuzzy2010** 3 years, 8 months ago

B is correct. Don't need to use Kibana or QuickSight.

upvoted 1 times

🗨️ 👤 **roytruong** 3 years, 8 months ago

ans is B

upvoted 3 times

🗨️ 👤 **cybe001** 3 years, 9 months ago

B is correct

upvoted 3 times

A manufacturing company has structured and unstructured data stored in an Amazon S3 bucket. A Machine Learning Specialist wants to use SQL to run queries on this data.

Which solution requires the LEAST effort to be able to query this data?

- A. Use AWS Data Pipeline to transform the data and Amazon RDS to run queries.
- B. Use AWS Glue to catalogue the data and Amazon Athena to run queries.
- C. Use AWS Batch to run ETL on the data and Amazon Aurora to run the queries.
- D. Use AWS Lambda to transform the data and Amazon Kinesis Data Analytics to run queries.

Suggested Answer: B

Community vote distribution

B (100%)

🗳️ 👤 **cybe001** Highly Voted 3 years, 9 months ago

B is correct

upvoted 24 times

🗳️ 👤 **dhs227** Highly Voted 3 years, 8 months ago

The correct answer HAS TO be B

Using Glue Use AWS Glue to catalogue the data and Amazon Athena to run queries against data on S3 are very typical use cases for those services.

D is not ideal, Lambda can surely do many things but it requires development/testing effort, and Amazon Kinesis Data Analytics is not ideal for ad-hoc queries.

upvoted 9 times

🗳️ 👤 **JonSno** Most Recent 4 months, 2 weeks ago

Selected Answer: B

B. Use AWS Glue to catalog the data and Amazon Athena to run queries.

Why is this the best choice?

AWS Glue can automatically catalog both structured and unstructured data in S3.

Amazon Athena is a serverless SQL query service that allows direct SQL queries on S3 data without moving it.

No infrastructure setup is required—just define a Glue Data Catalog and start querying with Athena.

upvoted 2 times

🗳️ 👤 **reginav** 6 months, 2 weeks ago

Selected Answer: B

S3 query == athena , to catalog data glue

upvoted 1 times

🗳️ 👤 **Ajose0** 9 months, 1 week ago

Selected Answer: B

AWS Glue is a fully managed ETL service that makes it easy to move data between data stores. It can automatically crawl, catalogue, and classify data stored in Amazon S3, and make it available for querying and analysis. With AWS Glue, you don't have to worry about the underlying infrastructure and can focus on your data.

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. It integrates with AWS Glue, so you can use the catalogued data directly in Athena without any additional data movement or transformation.

upvoted 3 times

🗳️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: B

The reason for this choice is that AWS Glue is a fully managed service that provides a data catalogue to make your data in S3 searchable and queryable¹. AWS Glue crawls your data sources, identifies data formats, and suggests schemas and transformations¹. You can use AWS Glue to catalogue both structured and unstructured data, such as relational data, JSON, XML, CSV files, images, or media files².

upvoted 1 times

🗳️ 👤 **Venkatesh_Babu** 1 year, 11 months ago

Selected Answer: B

I think it should be b
upvoted 1 times

🗳️ 👤 **SK27** 2 years, 6 months ago

Selected Answer: B

B is the easiest. We can use Glue crawler.
upvoted 2 times

🗳️ 👤 **ryuhei** 2 years, 9 months ago

Selected Answer: B

Answer B
upvoted 2 times

🗳️ 👤 **vetaal** 3 years, 5 months ago

Selected Answer: B

Querying data in S3 with SQL is almost always Athena.
upvoted 3 times

🗳️ 👤 **gcpwhiz** 3 years, 7 months ago

If AWS asks the question of querying unstructured data in an efficient manner, it is almost always Athena
upvoted 2 times

🗳️ 👤 **cloud_trail** 3 years, 7 months ago

B. I don't think that you even need Glue to transform anything. Just use Glue to define the schemas and then use Athena to query based on those schemas.
upvoted 2 times

🗳️ 👤 **Willnguyen22** 3 years, 8 months ago

answer is B
upvoted 1 times

🗳️ 👤 **syu31svc** 3 years, 8 months ago

SQL on S3 is Athena so answer is B for sure
upvoted 1 times

🗳️ 👤 **roytruong** 3 years, 8 months ago

B is right
upvoted 2 times

🗳️ 👤 **Jayraam** 3 years, 8 months ago

Answer is B.

Queries Against an Amazon S3 Data Lake

Data lakes are an increasingly popular way to store and analyze both structured and unstructured data. If you want to build your own custom Amazon S3 data lake, AWS Glue can make all your data immediately available for analytics without moving the data.

<https://aws.amazon.com/glue/>

upvoted 1 times

🗳️ 👤 **PRC** 3 years, 9 months ago

Correct Ans is D...Kinesis Data Analytics can use Lambda to transform and then run the SQL queries..
upvoted 1 times

🗳️ 👤 **Urban_Life** 3 years, 8 months ago

May I know why you are taking complex route?
upvoted 10 times

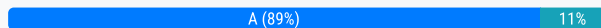
A Machine Learning Specialist is developing a custom video recommendation model for an application. The dataset used to train this model is very large with millions of data points and is hosted in an Amazon S3 bucket. The Specialist wants to avoid loading all of this data onto an Amazon SageMaker notebook instance because it would take hours to move and will exceed the attached 5 GB Amazon EBS volume on the notebook instance.

Which approach allows the Specialist to use all the data to train the model?

- A. Load a smaller subset of the data into the SageMaker notebook and train locally. Confirm that the training code is executing and the model parameters seem reasonable. Initiate a SageMaker training job using the full dataset from the S3 bucket using Pipe input mode.
- B. Launch an Amazon EC2 instance with an AWS Deep Learning AMI and attach the S3 bucket to the instance. Train on a small amount of the data to verify the training code and hyperparameters. Go back to Amazon SageMaker and train using the full dataset
- C. Use AWS Glue to train a model using a small subset of the data to confirm that the data will be compatible with Amazon SageMaker. Initiate a SageMaker training job using the full dataset from the S3 bucket using Pipe input mode.
- D. Load a smaller subset of the data into the SageMaker notebook and train locally. Confirm that the training code is executing and the model parameters seem reasonable. Launch an Amazon EC2 instance with an AWS Deep Learning AMI and attach the S3 bucket to train the full dataset.

Suggested Answer: A

Community vote distribution



JayK Highly Voted 3 years, 8 months ago

Answer is A. The answer to this question is about Pipe mode from S3. The only options are A and C. As AWS Glue cannot be use to create models which is option C.

The correct answer is A

upvoted 31 times

liangfb Highly Voted 3 years, 8 months ago

Answer is A.

upvoted 13 times

JonSno Most Recent 4 months, 2 weeks ago

Selected Answer: A

Training locally on a small dataset ensures the training script and model parameters are working correctly.

Amazon SageMaker training jobs allow direct access to S3 data without downloading everything.

Pipe input mode efficiently streams data from S3 to the training instance, reducing disk space requirements and speeding up training.

upvoted 4 times

reginav 6 months, 2 weeks ago

Selected Answer: A

Only Pipe mode can stream data from S3

upvoted 1 times

Mickey321 9 months, 1 week ago

Selected Answer: A

The reason for this choice is that Pipe input mode is a feature of Amazon SageMaker that allows you to stream data directly from an Amazon S3 bucket to your training instances without downloading it first¹. This way, you can avoid the time and space limitations of loading a large dataset onto your notebook instance. Pipe input mode also offers faster start times and better throughput than File input mode, which downloads the entire dataset before training¹.

upvoted 3 times

loict 9 months, 1 week ago

Selected Answer: A

A. YES - pipe mode is best to start inference before the entire data is transferred; the only drawback is if multiple training jobs are done in sequence (eg. different hyperparameter), the data will be downloaded again

B. NO - we want to use SageMaker first for initial training

C. NO - We first want to test things in SageMaker

D. NO - the SageMaker notebook will not use the AMI so the testing done is useless

upvoted 1 times

🗨️ 👤 **kyuhuck** 1 year, 4 months ago

Selected Answer: B

B. Generate daily precision-recall data in Amazon QuickSight, and publish the results in a dashboard shared with the Business team.

This solution leverages QuickSight's managed service capabilities for both data processing and visualization, which should minimize the coding effort required to provide the Business team with the necessary insights. However, it's important to note that QuickSight's ability to calculate the precision-recall data depends on its support for the necessary statistical functions or the availability of such calculations in the dataset. If QuickSight cannot perform these calculations directly, option C might be necessary, despite the increased effort.

upvoted 1 times

🗨️ 👤 **Venkatesh_Babu** 1 year, 11 months ago

Selected Answer: A

I think it should be a

upvoted 1 times

🗨️ 👤 **Valcilio** 2 years, 3 months ago

Selected Answer: A

It's A, pipe mode is for dealing with very big data.

upvoted 2 times

🗨️ 👤 **yemauricio** 2 years, 6 months ago

Selected Answer: A

A, PIPE is to do that sort of modeling

upvoted 2 times

🗨️ 👤 **Shailendraa** 2 years, 9 months ago

When data is already in S3 and next it should move to Sagemaker.. so option A is suitable

upvoted 1 times

🗨️ 👤 **Huy** 3 years, 8 months ago

Answer is A. B, C & D can be dropped because there is no integration from/to Sage Maker train job (model).

upvoted 1 times

🗨️ 👤 **cloud_trail** 3 years, 8 months ago

Gotta be A. You need to use Pipe mode but Glue cannot train a model.

upvoted 2 times

🗨️ 👤 **bobbylan1** 3 years, 8 months ago

AAAAAAAAAAa

upvoted 1 times

🗨️ 👤 **WillNguyen22** 3 years, 8 months ago

ans is A

upvoted 1 times

🗨️ 👤 **GeeBeeEl** 3 years, 8 months ago

Will you run AWS Deep Learning AMI for all cases where the data is very large in S3? Also what role is Glue playing here? Is there a transformation?

These are the two issues for options B C and D. I believe they do not represent what is required to satisfy the requirements in the question. The

answer definitely requires the pipe mode, but not with Glue. I go with A <https://aws.amazon.com/blogs/machine-learning/using-pipe-input-mode-for-amazon-sagemaker-algorithms/>

upvoted 3 times

🗨️ 👤 **roytruong** 3 years, 8 months ago

go for A

upvoted 2 times

A Machine Learning Specialist has completed a proof of concept for a company using a small data sample, and now the Specialist is ready to implement an end- to-end solution in AWS using Amazon SageMaker. The historical training data is stored in Amazon RDS. Which approach should the Specialist use for training a model using that data?

- A. Write a direct connection to the SQL database within the notebook and pull data in
- B. Push the data from Microsoft SQL Server to Amazon S3 using an AWS Data Pipeline and provide the S3 location within the notebook.
- C. Move the data to Amazon DynamoDB and set up a connection to DynamoDB within the notebook to pull data in.
- D. Move the data to Amazon ElastiCache using AWS DMS and set up a connection within the notebook to pull data in for fast access.

Suggested Answer: B

Community vote distribution

B (77%)

A (23%)

🗨️ 👤 **JayK** Highly Voted 9 months ago

Answer is B as the data for a SageMaker notebook needs to be from S3 and option B is the only option that says it. The only thing with option B is that it is talking of moving data from MS SQL Server not RDS

upvoted 31 times

🗨️ 👤 **mlyu** 3 years, 9 months ago

<https://www.slideshare.net/AmazonWebServices/train-models-on-amazon-sagemaker-using-data-not-from-amazon-s3-aim419-aws-reinvent-2018>

upvoted 2 times

🗨️ 👤 **HaiHN** 3 years, 8 months ago

Please look at the slide 14 of that link, although the data source from DynamoDB or RDS, it is still need to use AWS Glue to move the data to S3 for SageMaker to use.

So, the right answer should be B.

upvoted 2 times

🗨️ 👤 **jasonsunbao** 3 years, 8 months ago

I agree. As from the ML developer guide I just read, it is the MYSQL RDS that can be used as SQL datasource.

upvoted 2 times

🗨️ 👤 **Rama_Adim** Most Recent 1 month, 1 week ago

Selected Answer: A

Sagemaker can read from RDS directly - Maybe this is a new feature. Please check.

<https://aws.amazon.com/blogs/machine-learning/using-pipe-input-mode-for-amazon-sagemaker-algorithms/>

upvoted 1 times

🗨️ 👤 **JonSno** 4 months, 2 weeks ago

Selected Answer: B

Amazon SageMaker does not natively connect to Amazon RDS. Instead, training jobs work best with data stored in Amazon S3.

Amazon S3 is the preferred data source for SageMaker because:

It integrates seamlessly with SageMaker's training job infrastructure.

It supports distributed training for large datasets.

It is cost-effective and decouples storage from compute.

Best practice → Export RDS data to Amazon S3 and train using SageMaker.

upvoted 2 times

🗨️ 👤 **SophieSu** 9 months, 1 week ago

B is the correct answer.

Official AWS Documentation:

"Amazon ML allows you to create a datasource object from data stored in a MySQL database in Amazon Relational Database Service (Amazon RDS). When you perform this action, Amazon ML creates an AWS Data Pipeline object that executes the SQL query that you specify, and places the output into an S3 bucket of your choice. Amazon ML uses that data to create the datasource."

upvoted 2 times

🗨️ 👤 **Ajose0** 9 months, 1 week ago

Selected Answer: B

In Option B approach, the Specialist can use AWS Data Pipeline to automate the movement of data from Amazon RDS to Amazon S3. This allows for the creation of a reliable and scalable data pipeline that can handle large amounts of data and ensure the data is available for training.

In the Amazon SageMaker notebook, the Specialist can then access the data stored in Amazon S3 and use it for training the model. Using Amazon S3 as the source of training data is a common and scalable approach, and it also provides durability and high availability of the data.

upvoted 2 times

🗨️ 👤 **Mickey321** 9 months, 1 week ago

Selected Answer: B

This approach is the most scalable and reliable way to train a model using data stored in Amazon RDS. Amazon S3 is a highly scalable and durable object storage service, and Amazon Data Pipeline is a managed service that makes it easy to move data between different AWS services. By pushing the data to Amazon S3, the Specialist can ensure that the data is available for training the model even if the Amazon RDS instance is unavailable.

upvoted 2 times

🗨️ 👤 **loict** 9 months, 1 week ago

Selected Answer: B

- A. NO - SageMaker can only read from S3
- B. YES - AWS Data Pipeline can moved from SQL Server to S3
- C. NO - SageMaker can only read from S3 and not DynamoDB
- D. NO - SageMaker can only read from S3 and not ElastiCache

upvoted 2 times

🗨️ 👤 **shammous** 9 months, 1 week ago

Selected Answer: B

Option B (exporting to S3) is typically more flexible and cost-effective for large-scale or complex data needs (Which is our case - production), while Option A (direct connection) can be simpler and more immediate for real-time or smaller-scale scenarios like testing.

upvoted 1 times

🗨️ 👤 **ninomfr64** 9 months, 1 week ago

Selected Answer: B

- A. NO. It is doable, but this is not the best approach.
- B. YES
- C. NO. Pushing data to DynamoDB would not make it easier to access data
- D. NO. Pushing data to ElastiCache would not make it easier to access data

upvoted 1 times

🗨️ 👤 **Denise123** 1 year, 2 months ago

Selected Answer: A

For Amazon S3, you can import data from an Amazon S3 bucket as long as you have permissions to access the bucket.

For Amazon Athena, you can access databases in your AWS Glue Data Catalog as long as you have permissions through your Amazon Athena workgroup.

For Amazon RDS, if you have the AmazonSageMakerCanvasFullAccess policy attached to your user's role, then you'll be able to import data from your Amazon RDS databases into Canvas.

<https://docs.aws.amazon.com/sagemaker/latest/dg/canvas-connecting-external.html>

upvoted 4 times

🗨️ 👤 **Aja1** 1 year, 2 months ago

<https://aws.amazon.com/about-aws/whats-new/2024/04/amazon-sagemaker-studio-notebooks-data-sql-query/>

upvoted 1 times

🗨️ 👤 **Venkatesh_Babu** 1 year, 11 months ago

Selected Answer: B

I think it should be b

upvoted 1 times

🗨️ 👤 **Valcilio** 2 years, 3 months ago

Selected Answer: B

It's B, even if Microsoft SQL Server is a strange name for RDS, it's a possible database to use there and the data for sagemaker needs to be in S3!
upvoted 1 times

  **cnethers** 3 years, 8 months ago

While B is a valid answer, It is also possible to make a SQL connection in a notebook and create a data object so A could be a valid answer too
<https://stackoverflow.com/questions/36021385/connecting-from-python-to-sql-server>
<https://www.mssqltips.com/sqlservertip/6120/data-exploration-with-python-and-sql-server-using-jupyter-notebooks/>
upvoted 2 times

  **gcpwhiz** 3 years, 7 months ago

you need to choose the best answer, not any valid answer. Often, many of the answers are valid solutions, but are not best practice.
upvoted 2 times

  **scuzzy2010** 3 years, 8 months ago


B is correct. MS SQL Server is also under RDS.
upvoted 2 times

  **roytruong** 3 years, 8 months ago

B is right
upvoted 2 times

  **bhavesh0124** 3 years, 8 months ago

B it is
upvoted 1 times

  **cybe001** 3 years, 8 months ago

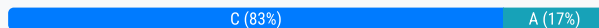
I'll go with B
upvoted 2 times

A Machine Learning Specialist receives customer data for an online shopping website. The data includes demographics, past visits, and locality information. The Specialist must develop a machine learning approach to identify the customer shopping patterns, preferences, and trends to enhance the website for better service and smart recommendations. Which solution should the Specialist recommend?

- A. Latent Dirichlet Allocation (LDA) for the given collection of discrete data to identify patterns in the customer database.
- B. A neural network with a minimum of three layers and random initial weights to identify patterns in the customer database.
- C. Collaborative filtering based on user interactions and correlations to identify patterns in the customer database.
- D. Random Cut Forest (RCF) over random subsamples to identify patterns in the customer database.

Suggested Answer: C

Community vote distribution



👤 **WWODIN** Highly Voted 3 years, 8 months ago

answer should be C
Collaborative filtering is for recommendation, LDA is for topic modeling
upvoted 21 times

👤 **syu31svc** Highly Voted 3 years, 8 months ago

In natural language processing, the latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.

Amazon SageMaker Random Cut Forest (RCF) is an unsupervised algorithm for detecting anomalous data points within a data set

Neural network is used for image detection

Answer is C
upvoted 12 times

👤 **Vernoxx** Most Recent 2 months, 1 week ago

Selected Answer: C
I think it should be c
upvoted 1 times

👤 **JonSno** 4 months, 2 weeks ago

Selected Answer: C
Collab filtering it is..
Collaborative filtering is the most widely used approach for recommendation systems.
It uses customer interactions (purchases, clicks, ratings) to determine preferences based on similar users or items.
Implicit collaborative filtering (based on user behavior) and explicit collaborative filtering (based on ratings) can effectively personalize recommendations.
upvoted 1 times

👤 **loict** 9 months, 1 week ago

Selected Answer: C
A. NO - LDA is for topic modeling
B. NO - NN is a too generic term, you want Neural Collaborative
C. YES - Collaborative filtering best fit
D. NO - Random Cut Forest (RCF) for anomalies
upvoted 3 times

👤 **Mickey321** 9 months, 1 week ago

Selected Answer: C

Collaborative filtering is a machine learning technique that recommends products or services to users based on the ratings or preferences of other users. This technique is well-suited for identifying customer shopping patterns and preferences because it takes into account the interactions between users and products.

upvoted 1 times

🗳️ 👤 **killermouse0** 1 year, 3 months ago

Selected Answer: A

From the doc: "You can use LDA for a variety of tasks, from clustering customers based on product purchases to automatic harmonic analysis in music."

<https://docs.aws.amazon.com/sagemaker/latest/dg/lda-how-it-works.html>

upvoted 1 times

🗳️ 👤 **Venkatesh_Babu** 1 year, 11 months ago

Selected Answer: C

I think it should be c

upvoted 1 times

🗳️ 👤 **Valcilio** 2 years, 3 months ago

Selected Answer: C

C, always when talk about recommendation you can think about collaborative patterns!

upvoted 2 times

🗳️ 👤 **stjokerli** 2 years, 3 months ago

A

LDA used before collaborative filtering is largely adopted.

1) the input data that we have doesn't lend itself to collaborative filtering - it requires a set of items and a set of users who have reacted to some of the items, which is NOT what we have

2) recommendation is just one thing that we want to do. What about trends?

3) collaborative filtering isn't one of the pre-built algorithms (weak argument, admittedly)

upvoted 2 times

🗳️ 👤 **Shailendraa** 2 years, 9 months ago

collaborative

upvoted 1 times

🗳️ 👤 **apprehensive_scar** 3 years, 4 months ago

C. Easy question.

upvoted 1 times

🗳️ 👤 **technoguy** 3 years, 8 months ago

its a appropriate use case of Collaborative filtering

upvoted 1 times

🗳️ 👤 **roytruong** 3 years, 8 months ago

this is C

upvoted 1 times

🗳️ 👤 **sdsfsdsf** 3 years, 8 months ago

I'm thinking that it is A because:

1) the input data that we have doesn't lend itself to collaborative filtering - it requires a set of items and a set of users who have reacted to some of the items, which is NOT what we have

2) recommendation is just one thing that we want to do. What about trends?

3) collaborative filtering isn't one of the pre-built algorithms (weak argument, admittedly)

upvoted 6 times

🗳️ 👤 **cybe001** 3 years, 8 months ago

Answer is C, demographics, past visits, and locality information data, LDA is appropriate

upvoted 3 times

🗳️ 👤 **cybe001** 3 years, 8 months ago

Collaborative filtering is appropriate

upvoted 4 times

🗳️ 👤 **DonaldCMLIN** 3 years, 9 months ago

Answer A might be more suitable than other

https://docs.aws.amazon.com/zh_tw/sagemaker/latest/dg/lda-how-it-works.html

upvoted 4 times

  **rsimham** 3 years, 9 months ago

Not convinced with A. Answer C seems to be a better fit than A for recommendation model (LDA appears to be a topic-based model on unavailable data with similar patterns)

<https://aws.amazon.com/blogs/machine-learning/extending-amazon-sagemaker-factorization-machines-algorithm-to-predict-top-x-recommendations/>

upvoted 10 times

A Machine Learning Specialist is working with a large company to leverage machine learning within its products. The company wants to group its customers into categories based on which customers will and will not churn within the next 6 months. The company has labeled the data available to the Specialist.

Which machine learning model type should the Specialist use to accomplish this task?

- A. Linear regression
- B. Classification
- C. Clustering
- D. Reinforcement learning

Suggested Answer: B

The goal of classification is to determine to which class or category a data point (customer in our case) belongs to. For classification problems, data scientists would use historical data with predefined target variables AKA labels (churner/non-churner) to train an algorithm. With classification, businesses can answer the following questions:

- ⇒ Will this customer churn or not?
- ⇒ Will a customer renew their subscription?
- ⇒ Will a user downgrade a pricing plan?
- ⇒ Are there any signs of unusual customer behavior?

Reference:

<https://www.kdnuggets.com/2019/05/churn-prediction-machine-learning.html>

Community vote distribution

B (100%)

 **rsimham** Highly Voted 3 years, 9 months ago

B seems to be okay
upvoted 14 times

 **JonSno** Most Recent 4 months, 2 weeks ago

Selected Answer: B

CLASSIFICATION - Binary Classification - Supervised Learning to be precise

The company wants to predict customer churn (whether a customer will leave or stay).

The data is labeled, meaning we have historical outcomes (churn or no churn).

The task involves categorizing customers into two groups:

Customers who will churn (leave)

Customers who will not churn (stay)

This means the problem is a Supervised Learning problem, specifically a binary classification problem.

The company wants to predict customer churn (whether a customer will leave or stay).

The data is labeled, meaning we have historical outcomes (churn or no churn).


The task involves categorizing customers into two groups:

Customers who will churn (leave)

Customers who will not churn (stay)

This means the problem is a Supervised Learning problem, specifically a binary classification problem.

upvoted 2 times

 **Mickey321** 9 months, 1 week ago

Selected Answer: B

The reason for this choice is that classification is a type of supervised learning that predicts a discrete categorical value, such as yes or no, spam or not spam, or churn or not churn¹. Classification models are trained using labeled data, which means that the input data has a known target attribute that indicates the correct class for each instance². For example, a classification model that predicts customer churn would use data that has a label indicating whether the customer churned or not in the past.

Classification models can be used for various applications, such as sentiment analysis, image recognition, fraud detection, and customer segmentation². Classification models can also handle both binary and multiclass problems, depending on the number of possible classes in the target attribute³.

upvoted 1 times

🗨️ 👤 **Sharath1783** 9 months, 1 week ago

Selected Answer: B

Option B. This is a scenario for supervised learning model as data is labelled and only A, B are supervised learning algorithms from the options. Linear learning is to predict time series data and distribution is selecting which class the input belongs to. Hence most suitable is to use Binomial distribution model in this case.

upvoted 1 times

🗨️ 👤 **loict** 9 months, 1 week ago

Selected Answer: B

- A. NO - Linear regression is not best for classification
- B. YES - Classification
- C. NO - we want supervised classification
- D. NO - there is nothing to Reinforce from

upvoted 1 times

🗨️ 👤 **mirik** 2 years ago

The question is not clear. Actually we have 2 tasks here - group into categories (clustering) and predict if customers will churn/not churn (classification). If we had to simply do classification, why there was mentioned to group into categories?

upvoted 3 times

🗨️ 👤 **ovokpus** 3 years ago

Selected Answer: B

This is definitely a classification problem

upvoted 4 times

🗨️ 👤 **Sivadharan** 3 years, 1 month ago

Selected Answer: B

B is correct

upvoted 2 times

🗨️ 👤 **FabG** 3 years, 7 months ago

B - it's a Binary Classification problem. Will the customer churn: Yes or No

upvoted 4 times

🗨️ 👤 **syu31svc** 3 years, 7 months ago

100% is B since it is about labelled data

upvoted 1 times

🗨️ 👤 **eji** 3 years, 8 months ago

i think the key is "the company has labeled the data" so this is classification, so it's B

upvoted 3 times

🗨️ 👤 **roytruong** 3 years, 9 months ago

B is okey

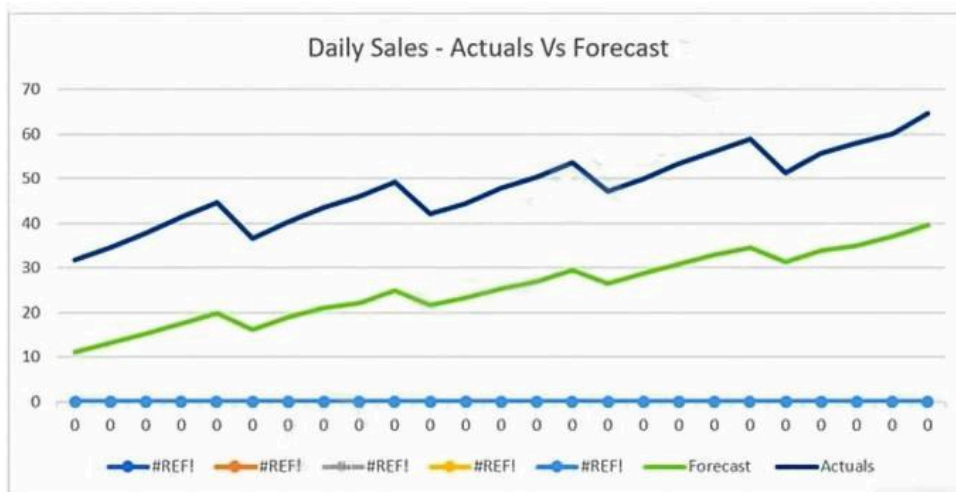
upvoted 2 times

🗨️ 👤 **cybe001** 3 years, 9 months ago

B is correct

upvoted 3 times

The displayed graph is from a forecasting model for testing a time series.



Considering the graph only, which conclusion should a Machine Learning Specialist make about the behavior of the model?

- A. The model predicts both the trend and the seasonality well
- B. The model predicts the trend well, but not the seasonality.
- C. The model predicts the seasonality well, but not the trend.
- D. The model does not predict the trend or the seasonality well.

Suggested Answer: A

Community vote distribution

A (100%)

jeetss1 Highly Voted 3 years, 2 months ago

A is correct answer.

Please Refer: <https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/>
upvoted 16 times

Dr_Kiko Highly Voted 3 years, 1 month ago

A; the problem is bias, not trends
upvoted 8 times

apache007 Most Recent 3 months ago

Selected Answer: B

B. The model predicts the trend well, but not the seasonality.

Here's what we can observe:

The predicted mean line closely follows the general upward trend of the observed line.

The predicted mean line does not capture the high frequency up and down changes of the observed line.

upvoted 1 times

VR10 10 months, 2 weeks ago

agreed, this seems to be A. there is similarity between the blue and green lines as far as capturing trend and seasonality is concerned. It just seems that if assumption is that the model is a linear regression model then just the intercept is off by a few units.
upvoted 2 times

Mickey321 1 year, 4 months ago

Selected Answer: A

A. The model predicts both the trend and the seasonality well
upvoted 1 times

Valcilio 1 year, 9 months ago

Selected Answer: A



The problem is Bias not trends or sesonality!

upvoted 2 times

  **spamicho** 3 years, 2 months ago



A is right, both trend (rising) and seasonality is there

upvoted 3 times

  **btsql** 3 years, 2 months ago

C is correct answer

upvoted 1 times

  **btsql** 3 years, 2 months ago

A is correct answer. Not C

upvoted 2 times

  **Kuntazulu** 3 years, 2 months ago

The trend is up, so isn't it correctly predicted? And the seasonality is also in sync, the amplitude is wrong.

upvoted 3 times

  **georschi** 3 years, 2 months ago



A is right. trend and seasonality are fine, level is the one the model gets wrong

upvoted 4 times

  **NotAnMLProfessional** 3 years, 3 months ago

Should be C

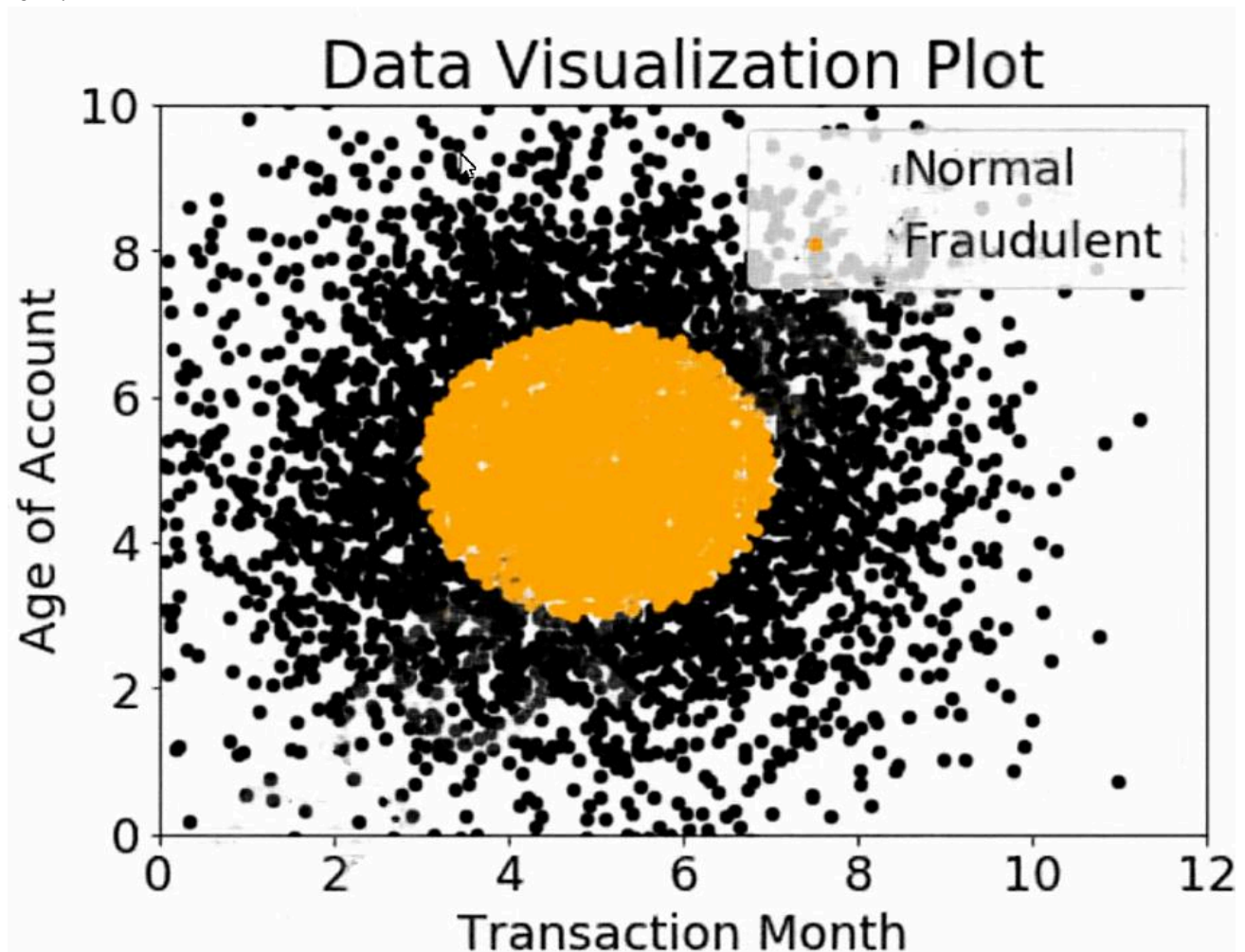
upvoted 1 times

  **ashlash** 3 years, 3 months ago

Should be A

upvoted 2 times

A company wants to classify user behavior as either fraudulent or normal. Based on internal research, a Machine Learning Specialist would like to build a binary classifier based on two features: age of account and transaction month. The class distribution for these features is illustrated in the figure provided.



Based on this information, which model would have the HIGHEST accuracy?

- A. Long short-term memory (LSTM) model with scaled exponential linear unit (SELU)
- B. Logistic regression
- C. Support vector machine (SVM) with non-linear kernel
- D. Single perceptron with tanh activation function

Suggested Answer: C

Community vote distribution

C (100%)

mizuakari Highly Voted 3 years, 9 months ago

Answer is C. SVM sample use case is to put the dimensions into a higher hyperplane that can separate it. Seeing how separable it is, SVM can be used for it.

upvoted 21 times

JonSno Most Recent 4 months, 2 weeks ago

Selected Answer: C

Support Vector Machine (SVM) with Non-Linear Kernel -> Non-linear Data
Why?

SVM is powerful for classification and works well even with small datasets.

If the data has a non-linear decision boundary, using an SVM with a non-linear kernel (like RBF or polynomial) can improve accuracy.

Works well in low-dimensional feature spaces (since we have only 2 features: age of account & transaction month).

Optimal choice if the data has a non-linear decision boundary.

upvoted 1 times

🗨️ 👤 **MultiCloudIronMan** 8 months, 1 week ago

Selected Answer: C

SVMs are particularly effective for binary classification tasks and can handle non-linear relationships between features¹.

upvoted 2 times

🗨️ 👤 **Mickey321** 9 months, 1 week ago

Selected Answer: C

You can use a support vector machine (SVM) when your data has exactly two classes. An SVM classifies data by finding the best hyperplane that separates all data points of one class from those of the other class. The best hyperplane for an SVM means the one with the largest margin between the two classes. Margin means the maximal width of the slab parallel to the hyperplane that has no interior data points.

upvoted 2 times

🗨️ 👤 **kaike_reis** 1 year, 11 months ago

Well, C is the correct answer. This example is a classical one to use SVM.

upvoted 1 times

🗨️ 👤 **Valcilio** 2 years, 3 months ago

Selected Answer: C

SVM for RBF mode is the answer!

upvoted 1 times

🗨️ 👤 **Broncomailo** 3 years, 4 months ago

Selected Answer: C

Answer is C

upvoted 4 times

🗨️ 👤 **Dr_Kiko** 3 years, 7 months ago

Textbook C

upvoted 3 times

🗨️ 👤 **halfway** 3 years, 8 months ago

C. more reading for using non-linear kernel and separate samples with a hyperplane in a higher dimension space:

<https://medium.com/pursuitnotes/day-12-kernel-svm-non-linear-svm-5fdefe77836c>

upvoted 2 times

🗨️ 👤 **spamicho** 3 years, 8 months ago

C seems right

upvoted 1 times

🗨️ 👤 **Juka3lj** 3 years, 8 months ago

answer is C

upvoted 2 times

🗨️ 👤 **omar_bahrain** 3 years, 9 months ago

Agree. The answer is A.

<https://www.surveypractice.org/article/2715-using-support-vector-machines-for-survey-research>

upvoted 1 times

🗨️ 👤 **cnethers** 3 years, 9 months ago

This is a good explanation of SVM

<https://uk.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html>

upvoted 3 times

A Machine Learning Specialist at a company sensitive to security is preparing a dataset for model training. The dataset is stored in Amazon S3 and contains

Personally Identifiable Information (PII).

The dataset:

- ⇒ Must be accessible from a VPC only.
- ⇒ Must not traverse the public internet.

How can these requirements be satisfied?

- A. Create a VPC endpoint and apply a bucket access policy that restricts access to the given VPC endpoint and the VPC.
- B. Create a VPC endpoint and apply a bucket access policy that allows access from the given VPC endpoint and an Amazon EC2 instance.
- C. Create a VPC endpoint and use Network Access Control Lists (NACLs) to allow traffic between only the given VPC endpoint and an Amazon EC2 instance.
- D. Create a VPC endpoint and use security groups to restrict access to the given VPC endpoint and an Amazon EC2 instance

Suggested Answer: A

Community vote distribution

A (83%)

B (17%)

🗨️ 👤 **rajs** Highly Voted 3 years, 9 months ago

Important things to note here is that

1. "The Data in S3 Needs to be Accessible from VPC"
2. "Traffic should not Traverse internet"

To fulfill Requirement #2 we need a VPC endpoint

To RESTRICT the access to S3/Bucket

- Access allowed only from VPC via VPC Endpoint

Even though Sagemaker uses EC2 - we are NOT asked to secure the EC2 :)

So the answer is A

upvoted 41 times

🗨️ 👤 **sdsfdsf** Highly Voted 3 years, 8 months ago

Between A & B, the answer should be A. From here:

<https://docs.aws.amazon.com/vpc/latest/userguide/vpc-endpoints-s3.html#vpc-endpoints-s3-bucket-policies>

We can see that we restrict access using DENY if sourceVpce (vpc endpoint), or sourceVpc (vpc) is not equal to our VPCe/VPC. So we are using a DENY (choice A) and not an ALLOW policy (choice B).

Choices C, D we eliminate because they don't address S3 access at all.

upvoted 12 times

🗨️ 👤 **JonSno** Most Recent 4 months, 2 weeks ago

Selected Answer: A

Create a VPC endpoint and apply a bucket access policy that restricts access to the given VPC endpoint and the VPC.

Why is this correct?

VPC endpoint for S3 allows private connectivity between Amazon S3 and the VPC without using the public internet.

Bucket access policy can be written to allow access only from this VPC endpoint.

This ensures maximum security by:

Preventing access from outside the VPC.

Blocking public access.

upvoted 2 times

🗨️ 👤 **Ajose0** 9 months, 1 week ago

Selected Answer: A

In Option A, the Machine Learning Specialist would create a VPC endpoint for Amazon S3, which would allow traffic to flow directly between the VPC and Amazon S3 without traversing the public internet. Access to the S3 bucket containing PII can then be restricted to the VPC endpoint and the VPC using a bucket access policy. This would ensure that only instances within the VPC can access the data, and that the data does not traverse the public internet.

Option B and D, allowing access from an Amazon EC2 instance, would not meet the requirement of not traversing the public internet, as the EC2 instance would be accessible from the internet. Option C, using Network Access Control Lists (NACLs) to allow traffic between only the VPC endpoint and an EC2 instance, would also not meet the requirement of not traversing the public internet, as the EC2 instance would still be accessible from the internet.

upvoted 1 times

🗳️ 👤 **loict** 9 months, 1 week ago

Selected Answer: A

A. YES - We first create a S3 endpoint in the VPC subnet so traffic does not flow through the Internet, then on the S3 bucket create an access policy that restricts access to the given VPC based on its ID

B. NO - we don't want to be specific to an instance

C. NO - the S3 bucket is on AWS network, you cannot change the NACL for it

D. NO - not all instances in a VPC will necessarily have the same principal that can be specified in the policy

upvoted 2 times

🗳️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: A

Definetly A

upvoted 1 times

🗳️ 👤 **kaike_reis** 1 year, 11 months ago

Selected Answer: A

Well, but removing methodology, only A remains: The question never cited EC2

upvoted 3 times

🗳️ 👤 **ADVIT** 2 years ago

Per <https://docs.aws.amazon.com/AmazonS3/latest/userguide/example-bucket-policies-vpc-endpoint.html> it's A

upvoted 1 times

🗳️ 👤 **exam887** 3 years, 1 month ago

Selected Answer: A

The question do not mention EC2 at all, so should be A

upvoted 4 times

🗳️ 👤 **dunhill** 3 years, 6 months ago

I think it should be B. Traning instance is a EC2 instance and need to be set an endpoint to load the data from S3.

upvoted 1 times

🗳️ 👤 **[Removed]** 3 years, 7 months ago

Selected Answer: B

AWS security is a conservative security model, which implies that access are denied by default rather than granted by default. We have to explicitly allow access to a AWS resource. Additionally, B talks about allowing access FROM the VPC to S3 while A talks about allowing access from S3 to VPC (which is not what we need).

So, B.

upvoted 2 times

🗳️ 👤 **cpal012** 2 years, 3 months ago

Um, no. A VPC endpoint is outbound from the VPC to a supported AWS service.

upvoted 1 times

🗳️ 👤 **technoguy** 3 years, 7 months ago

Will go with B

upvoted 1 times

🗳️ 👤 **spamicho** 3 years, 7 months ago

Betting on B here, we should control access from VPC, not to VPC.

upvoted 1 times

🗳️ 👤 **achiko** 3 years, 8 months ago

A!

Restricting access to a specific VPC endpoint

The following is an example of an Amazon S3 bucket policy that restricts access to a specific bucket, `awsexamplebucket1`, only from the VPC endpoint with the ID `vpce-1a2b3c4d`. The policy denies all access to the bucket if the specified endpoint is not being used. The `aws:SourceVpce` condition is used to specify the endpoint.



<https://docs.aws.amazon.com/AmazonS3/latest/userguide/example-bucket-policies-vpc-endpoint.html>

upvoted 2 times

  **senseikimoji** 3 years, 8 months ago

Can't be B. You simple cannot enable access to an endpoint to some selected instance. So A.

upvoted 1 times

  **Huy** 3 years, 7 months ago



We shouldn't use private IP in bucket policy.

upvoted 1 times

  **cloud_trail** 3 years, 8 months ago

B does not say enable access TO the VPC endpoint. It says to allow access FROM the endpoint. So B is the correct answer. A talks about restricting access TO the VPC endpoint, so that option is irrelevant. We're worried about access TO the S3 bucket, not access to the VPC. The question is not poorly-worded, but it is tricky and you need to read it carefully.

upvoted 1 times

  **yeetusdeleetus** 3 years, 8 months ago

I also vote A.

upvoted 1 times

  **Thai_Xuan** 3 years, 8 months ago

A

found here

"You can control which VPCs or VPC endpoints have access to your buckets by using Amazon S3 bucket policies. For examples of this type of bucket policy access control, see the following topics on restricting access."

<https://docs.aws.amazon.com/AmazonS3/latest/dev/example-bucket-policies-vpc-endpoint.html>

upvoted 3 times

During mini-batch training of a neural network for a classification problem, a Data Scientist notices that training accuracy oscillates. What is the MOST likely cause of this issue?

- A. The class distribution in the dataset is imbalanced.
- B. Dataset shuffling is disabled.
- C. The batch size is too big.
- D. The learning rate is very high.


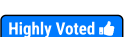
Suggested Answer: D

Reference:

<https://towardsdatascience.com/deep-learning-personal-notes-part-1-lesson-2-8946fe970b95>

Community vote distribution

D (100%)

 **gaku1016**  3 years, 9 months ago

Answer is D.

Should the weight be increased or reduced so that the error is smaller than the current value? You need to examine the amount of change to know that. Therefore, we differentiate and check whether the slope of the tangent is positive or negative, and update the weight value in the direction to reduce the error. The operation is repeated over and over so as to approach the optimal solution that is the goal. The width of the update amount is important at this time, and is determined by the learning rate.

upvoted 17 times

 **ozan11**  3 years, 9 months ago

maybe D ?

upvoted 8 times

 **JonSno**  4 months, 1 week ago


Selected Answer: D

D. The learning rate is very high.

Explanation:

When the learning rate is too high, the optimization process may overshoot the optimal weights in parameter space. Instead of gradually converging, the model updates weights in a highly unstable manner, causing fluctuations in training accuracy. The network fails to settle into a minimum because the updates are too aggressive.


upvoted 2 times

 **Ajose0** 9 months, 1 week ago

Selected Answer: D

A high learning rate can cause oscillations in the training accuracy because the optimizer makes large updates to the model parameters in each iteration, which can cause overshooting the optimal values. This can result in the model oscillating back and forth across the optimal solution.

upvoted 3 times

 **Mickey321** 9 months, 1 week ago

Selected Answer: D

If the learning rate is too high, the model weights may overshoot the optimal values and bounce back and forth around the minimum of the loss function. This can cause the training accuracy to oscillate and prevent the model from converging to a stable solution. The training accuracy is the proportion of correct predictions made by the model on the training data.

upvoted 2 times

 **Reju** 9 months, 1 week ago

When the learning rate is set too high, it can lead to oscillations or divergence during training. Here's why:

High Learning Rate: A high learning rate means that the model's parameters are updated by a large amount in each training step. This can cause the model to overshoot the optimal parameter values, leading to instability in training.

Oscillations: If the learning rate is excessively high, the model's updates can become unstable, causing it to oscillate back and forth between

parameter values. This oscillation can prevent the model from converging to an optimal solution.

To address this issue, you can try reducing the learning rate. It's often necessary to experiment with different learning rates to find the one that works best for your specific problem and dataset. Learning rate scheduling techniques, such as reducing the learning rate over time, can also help stabilize training.

upvoted 2 times

🗨️ **CKS1210** 2 years ago

Answer is A.

A high learning rate means that the model parameters are being updated by large magnitudes in each iteration. As a result, the optimization process may struggle to converge to the optimal solution, leading to erratic behavior and fluctuations in training accuracy.

upvoted 1 times

🗨️ **soonmo** 2 years ago

Selected Answer: D

If learning rate is high, the accuracy is fluctuated because the value of loss function moves back and forth over the global minimum.

upvoted 1 times

🗨️ **Valcilio** 2 years, 3 months ago

Selected Answer: D

The big learning rating overshoot in true minima.

upvoted 2 times

🗨️ **Tomatoteacher** 2 years, 5 months ago

Selected Answer: D

D Learning rate is too high. Textbook example of learning rate being too high. Lower Learning_rate will take more iterations, or longer to train, but will settle in place.

upvoted 1 times

🗨️ **Shailendraa** 2 years, 9 months ago

12-sep exam

upvoted 1 times

🗨️ **Sam1610** 2 years, 11 months ago

D: per supuesto

upvoted 1 times

🗨️ **missionml** 3 years, 3 months ago

A company sells thousands of products on a public website and wants to automatically identify products with potential durability problems. The company has 1.000 reviews with date, star rating, review text, review summary, and customer email fields, but many reviews are incomplete and have empty fields. Each review has already been labeled with the correct durability result.

A machine learning specialist must train a model to identify reviews expressing concerns over product durability. The first model needs to be trained and ready to review in 2 days.

What is the MOST direct approach to solve this problem within 2 days?

A.

Train a custom classifier by using Amazon Comprehend.

B.

Build a recurrent neural network (RNN) in Amazon SageMaker by using Gluon and Apache MXNet.

C.

Train a built-in BlazingText model using Word2Vec mode in Amazon SageMaker.

D.

Use a built-in seq2seq model in Amazon SageMaker.

upvoted 1 times

🗨️ **missionml** 3 years, 3 months ago

Is A valid option?

upvoted 1 times

🗨️ **btsql** 3 years, 7 months ago

D is correct. big batch size make local minia.

upvoted 1 times

🗨️ 👤 **jeetss1** 3 years, 8 months ago

it is a multiple answer question and answer should be both A and D
upvoted 1 times

🗨️ 👤 **syu31svc** 3 years, 8 months ago

Answer is D 100%; learning rate too high will cause such an event
upvoted 3 times

🗨️ 👤 **deep_n** 3 years, 8 months ago

The answer is D, from the Coursera deep learning specialization (course 2 - improving Deep NN)
upvoted 2 times

An employee found a video clip with audio on a company's social media feed. The language used in the video is Spanish. English is the employee's first language, and they do not understand Spanish. The employee wants to do a sentiment analysis. What combination of services is the MOST efficient to accomplish the task?

- A. Amazon Transcribe, Amazon Translate, and Amazon Comprehend
- B. Amazon Transcribe, Amazon Comprehend, and Amazon SageMaker seq2seq
- C. Amazon Transcribe, Amazon Translate, and Amazon SageMaker Neural Topic Model (NTM)
- D. Amazon Transcribe, Amazon Translate and Amazon SageMaker BlazingText

Suggested Answer: A

Community vote distribution

A (100%)

🗳️ 👤 **DonaldCMLIN** Highly Voted 3 years, 9 months ago
 the MOST efficient means to you don't need to coding, building infra
 All of services are manage by AWS is good,
 Transcribe, Amazon Translate, and Amazon Comprehend

Answer is A
 upvoted 44 times

🗳️ 👤 **WVODIN** 3 years, 9 months ago
 Agree, Answer is A
 upvoted 9 times

🗳️ 👤 **Pg690** Highly Voted 2 years, 7 months ago
 A is not 100% correct. You don't need to translate Spanish. Amazon Comprehend supports Spanish.
 upvoted 8 times

🗳️ 👤 **cpal012** 2 years, 3 months ago
 Arguably, you still need a translation since the person doesn't speak Spanish.
 upvoted 2 times

🗳️ 👤 **tonton3** 2 years ago
 I think there is no need to use Amazon translate because sometimes the translation is not accurate.
 It means some information gets lost.
 upvoted 1 times

🗳️ 👤 **kaike_reis** 1 year, 11 months ago
 Given the question, I believe that is necessary: look at the enphase of not understanding spanish. besides that, even with some information lost, you will at least understand something.
 upvoted 1 times

🗳️ 👤 **JonSno** Most Recent 4 months, 1 week ago
Selected Answer: A
 A. Amazon Transcribe, Amazon Translate, and Amazon Comprehend
 Explanation of the Process:
 Amazon Transcribe – Converts the Spanish audio in the video into text.
 Amazon Translate – Translates the Spanish text to English.
 Amazon Comprehend – Performs sentiment analysis on the translated English text.
 upvoted 3 times

🗳️ 👤 **ADVIT** 9 months, 1 week ago
 It's A:
 1.Amazon Transcribe - to convert Spanish speech to Spanish text.
 2.Amazon Translate - to translate Spanish text to English text
 3.Amazon Comprehend - to analyze text for sentiments

upvoted 2 times

🗳️ 👤 **loict** 9 months, 1 week ago

Selected Answer: A

- A. YES - Comprehend is supervised so user must understand through Translate
- B. NO - seq2seq is for generation and not classification
- C. NO - Amazon SageMaker Neural Topic Model is unsupervised topic extraction, will not give sentiment against user-defined classes
- D. NO - BlazingText is word2vec, does not give sentiment classes

upvoted 1 times

🗳️ 👤 **Mickey321** 9 months, 1 week ago

Selected Answer: A

It's A:

1. Amazon Transcribe - to convert Spanish speech to Spanish text.
2. Amazon Translate - to translate Spanish text to English text
3. Amazon Comprehend - to analyze text for sentiments

upvoted 2 times

🗳️ 👤 **DavidRou** 1 year, 11 months ago

It's A 100%

upvoted 1 times

🗳️ 👤 **CKS1210** 2 years ago

Transcribe: Speech to text

Translate: Any language to any language

Comprehend: offers a range of capabilities for extracting insights and meaning from unstructured text data. Ex: Sentiment analysis, entity recognition, KeyPhrase Extraction, Language Detection, Document Classification

upvoted 1 times

🗳️ 👤 **soonmo** 2 years ago

absolutely need STT(transcribe), translation(translate), and sentimental analysis(comprehend)

upvoted 1 times

🗳️ 👤 **gnolam** 2 years, 9 months ago

Selected Answer: A

A - confirmed by ACG

upvoted 2 times

🗳️ 👤 **KM226** 3 years, 6 months ago

Selected Answer: A

I agree that the answer is A

upvoted 1 times

🗳️ 👤 **in4976** 3 years, 6 months ago

Selected Answer: A

answer is a

upvoted 1 times

🗳️ 👤 **Dr_Kiko** 3 years, 7 months ago

A; D is wrong because The Amazon SageMaker BlazingText algorithm provides highly optimized implementations of the Word2vec and text classification algorithms.

upvoted 1 times

🗳️ 👤 **harmanbirstudy** 3 years, 8 months ago

The Question/Answer is not poorly as someone mentioned.

--Even though Comprehend can do the analysis directly on Spanish (no need of translate) but if comprehend does analysis and the resulting words are still in Spanish, it will not help the employee as he doesn't know Spanish. So the translate after transcribe will help Employee understand what is being analyzed by Comprehend in next step.



So read the question carefully before jumping to conclusions. it will save you an Exam :)

upvoted 1 times

🗳️ 👤 **senseikimoji** 3 years, 8 months ago

I don't get this question. Comprehend supports Spanish natively. There is no need for Translate, and translate would actually reduce effectiveness of sentimental analysis. However, BCD are all invalid choices.

upvoted 3 times

  **ybad** 3 years, 8 months ago

A

because Comprehend can provide sentiment analysis

upvoted 2 times

  **FastTrack** 3 years, 8 months ago

A,

<https://aws.amazon.com/getting-started/hands-on/analyze-sentiment-comprehend/>

upvoted 4 times

A Machine Learning Specialist is packaging a custom ResNet model into a Docker container so the company can leverage Amazon SageMaker for training. The Specialist is using Amazon EC2 P3 instances to train the model and needs to properly configure the Docker container to leverage the NVIDIA GPUs.

What does the Specialist need to do?

- A. Bundle the NVIDIA drivers with the Docker image.
- B. Build the Docker container to be NVIDIA-Docker compatible.
- C. Organize the Docker container's file structure to execute on GPU instances.
- D. Set the GPU flag in the Amazon SageMaker CreateTrainingJob request body.

Suggested Answer: A

Community vote distribution

B (100%)

  **vetal** Highly Voted 3 years, 9 months ago

<https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf>
page 55:

If you plan to use GPU devices, make sure that your containers are nvidia-docker compatible. Only the CUDA toolkit should be included on containers. Don't bundle NVIDIA drivers with the image. For more information about nvidia-docker, see NVIDIA/nvidia-docker.

So the answer is B
upvoted 52 times

  **devsean** 3 years, 9 months ago

Yeah, it's B. But the page in the developer guide is page number 201 (209 in pdf). Second bullet point at the top.
upvoted 6 times

  **AKT** Highly Voted 3 years, 9 months ago

Answer is B. below is from AWS documentation,

If you plan to use GPU devices for model training, make sure that your containers are nvidia-docker compatible. Only the CUDA toolkit should be included on containers; don't bundle NVIDIA drivers with the image.

<https://docs.aws.amazon.com/sagemaker/latest/dg/your-algorithms-training-algo-dockerfile.html>
upvoted 14 times


  **JonSno** Most Recent 4 months, 1 week ago

Selected Answer: B

When using Amazon SageMaker with GPU-based EC2 instances (e.g., P3 instances), you must ensure that your custom Docker container can leverage NVIDIA GPUs. NVIDIA-Docker (now part of Docker with nvidia-container-runtime) allows containers to access GPU resources without needing to bundle NVIDIA drivers inside the container.

To make a custom Docker container GPU-compatible, the Machine Learning Specialist should:

Use NVIDIA CUDA and cuDNN in the Dockerfile.
Ensure the container is built using the NVIDIA Container Toolkit (nvidia-docker).
Use nvidia-container-runtime as the runtime.
upvoted 2 times

  **Ajose0** 9 months, 1 week ago

Selected Answer: B

To leverage the NVIDIA GPUs on Amazon EC2 P3 instances for training with Amazon SageMaker, the Docker container must be built to be compatible with NVIDIA-Docker.

NVIDIA-Docker is a wrapper around Docker that makes it easier to use GPUs in containers by providing GPU-aware functionality.

To build a Docker container that is compatible with NVIDIA-Docker, the Specialist should install the NVIDIA GPU drivers in the Docker container and install the NVIDIA-Docker runtime on the EC2 instances.

upvoted 1 times

🗳️ 👤 **bakarys** 9 months, 1 week ago

Selected Answer: B

NVIDIA-Docker is a Docker container runtime plugin that allows the Docker container to access the GPU resources on the host machine. By building the Docker container to be NVIDIA-Docker compatible, the Docker container will have access to the NVIDIA GPU resources on the Amazon EC2 P3 instances, allowing for accelerated training of the ResNet model.

upvoted 1 times

🗳️ 👤 **Mickey321** 9 months, 1 week ago

Selected Answer: B

The reason for this choice is that NVIDIA-Docker is a tool that enables GPU-accelerated containers by automatically configuring the container runtime to use NVIDIA GPUs¹. NVIDIA-Docker allows you to build and run Docker containers that can fully access the GPUs on your host system. This way, you can run GPU-intensive applications, such as deep learning frameworks, inside containers without any performance loss or compatibility issues.

upvoted 1 times

🗳️ 👤 **loict** 9 months, 1 week ago

Selected Answer: B

- A. NO - the drivers are not necessary (<https://docs.aws.amazon.com/sagemaker/latest/dg/your-algorithms-training-algo-dockerfile.html>)
- B. YES - it is about using the CUDA library, need to use proper base image (<https://medium.com/@jgleeee/building-docker-images-that-require-nvidia-runtime-environment-1a23035a3a58>)
- C. NO - file structure irrelevant to GPU
- D. NO - SageMaker config, irrelevant to Docker

upvoted 2 times

🗳️ 👤 **6ff83cb** 1 year, 4 months ago

Selected Answer: B

<https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf>
page 55

upvoted 1 times

🗳️ 👤 **iambasspaul** 1 year, 2 months ago

page 570

On a GPU instance, the image is run with the --gpus option. Only the CUDA toolkit should be included in the image not the NVIDIA drivers. For more information, see NVIDIA User Guide.

upvoted 1 times

🗳️ 👤 **Crypt0zknight** 1 year, 9 months ago

Answer B

Load the CUDA toolkit only, not the drivers. Ref GPU section : <https://docs.aws.amazon.com/sagemaker/latest/dg/studio-byoi-specs.html>

upvoted 1 times

🗳️ 👤 **Venkatesh_Babu** 1 year, 11 months ago

Selected Answer: B

I think it should be b

upvoted 1 times

🗳️ 👤 **jackzhao** 2 years, 3 months ago

B is correct!

upvoted 1 times

🗳️ 👤 **Sylzys** 2 years, 3 months ago

Selected Answer: B

As per aws documentation, answer is B, and A is even explicitly not recommended

upvoted 1 times

🗳️ 👤 **Sorrybutnotsorry** 3 years, 5 months ago

Selected Answer: B

As referred in other comments ans is B

upvoted 1 times

🗳️ 👤 **hussamS** 3 years, 6 months ago

Selected Answer: B

ANS B

As mentioned by other users

upvoted 1 times

🗨️ 👤 **sachin80** 3 years, 8 months ago

As per me answer is B

upvoted 1 times

🗨️ 👤 **konradL** 3 years, 8 months ago

The answer is for sure B - as mentioned by others. And this is clearly stated in the docs

upvoted 1 times

🗨️ 👤 **takahirokoyama** 3 years, 8 months ago

Ans. is B.

upvoted 1 times

A Machine Learning Specialist is building a logistic regression model that will predict whether or not a person will order a pizza. The Specialist is trying to build the optimal model with an ideal classification threshold.

What model evaluation technique should the Specialist use to understand how different classification thresholds will impact the model's performance?

- A. Receiver operating characteristic (ROC) curve
- B. Misclassification rate
- C. Root Mean Square Error (RMSE)
- D. L1 norm


Suggested Answer: A

Reference:

<https://docs.aws.amazon.com/machine-learning/latest/dg/binary-model-insights.html>

Community vote distribution

A (100%)

  **rsimham** Highly Voted 3 years, 9 months ago

Ans. A is correct

upvoted 20 times

  **AKT** Highly Voted 3 years, 9 months ago

Answer is A.

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds
upvoted 9 times

  **JonSno** Most Recent 4 months, 1 week ago

Selected Answer: A

Explanation:

The ROC curve is the best technique to evaluate how different classification thresholds impact the model's performance. It plots True Positive Rate (TPR) against False Positive Rate (FPR) at various threshold values.

Why the ROC Curve?

Logistic regression outputs probabilities, and we need to select a classification threshold to decide between "order pizza" (1) and "not order pizza" (0).

Changing the threshold impacts the trade-off between sensitivity (recall) and specificity.

The ROC curve helps visualize this trade-off and select the best threshold based on the business goal (e.g., maximizing recall vs. minimizing false positives).

The Area Under the ROC Curve (AUC-ROC) is a useful metric to measure the model's discrimination ability.

upvoted 2 times

  **GeeBeeEI** 9 months, 1 week ago

A is indeed correct see <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc>

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.

This curve plots two parameters:

- True Positive Rate
- False Positive Rate

True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:

$TPR = TP / (TP + FN)$

False Positive Rate (FPR) is defined as follows:

$FPR = FP / (FP + TN)$

upvoted 4 times

  **Mickey321** 9 months, 1 week ago

Selected Answer: A

The reason for this choice is that a ROC curve is a graphical plot that illustrates the performance of a binary classifier across different values of the classification threshold¹. A ROC curve plots the true positive rate (TPR) or sensitivity against the false positive rate (FPR) or 1-specificity for various threshold values². The TPR is the proportion of positive instances that are correctly classified, while the FPR is the proportion of negative instances that are incorrectly classified.

upvoted 2 times

🗨️ 👤 **Valcilio** 2 years, 3 months ago

Selected Answer: A

ROC curve is for defining the threshold.

upvoted 2 times

🗨️ 👤 **spamicho** 3 years, 7 months ago

A surely

upvoted 1 times

🗨️ 👤 **syu31svc** 3 years, 7 months ago

Question is about classification so confusion matrix would come into mind; A is the answer

upvoted 1 times

🗨️ 👤 **hans1234** 3 years, 8 months ago

It is A.

upvoted 1 times

🗨️ 👤 **roytruong** 3 years, 8 months ago

obviously A

upvoted 1 times

🗨️ 👤 **bitiyaha** 3 years, 9 months ago

Root Mean Square Error (RMSE) Ans. c

upvoted 1 times

🗨️ 👤 **bzhao** 3 years, 8 months ago

I think RMSE is for regression model

upvoted 5 times

An interactive online dictionary wants to add a widget that displays words used in similar contexts. A Machine Learning Specialist is asked to provide word features for the downstream nearest neighbor model powering the widget. What should the Specialist do to meet these requirements?

- A. Create one-hot word encoding vectors.
- B. Produce a set of synonyms for every word using Amazon Mechanical Turk.
- C. Create word embedding vectors that store edit distance with every other word.
- D. Download word embeddings pre-trained on a large corpus.

Suggested Answer: A

Reference:

<https://aws.amazon.com/blogs/machine-learning/amazon-sagemaker-object2vec-adds-new-features-that-support-automatic-negative-sampling-and-speed-up-training/>


Community vote distribution

D (100%)

 **JayK** Highly Voted 3 years, 8 months ago

the solution is word embedding. As it is a interactive online dictionary, we need pre-trained word embedding thus the answer is D. In addition, there is no mention that the online dictionary is unique and does not have a pre-trained word embedding.

Thus I strongly feel the answer is D
upvoted 31 times

 **cybe001** Highly Voted 3 years, 8 months ago

D is correct. It is not a specialized dictionary so use the existing word corpus to train the model
upvoted 16 times

 **JonSno** Most Recent 4 months, 1 week ago

Selected Answer: D

D. Download word embeddings pre-trained on a large corpus.

Reason :

For a nearest neighbor model that finds words used in similar contexts, word embeddings are the best choice. Pre-trained word embeddings capture semantic relationships and contextual similarity between words based on a large text corpus (e.g., Wikipedia, Common Crawl).


The Specialist should:

Use pre-trained word embeddings like Word2Vec, GloVe, or FastText.

Load the embeddings into the model for efficient similarity comparisons.

Use a nearest neighbor search algorithm (e.g., FAISS, k-d tree, Annoy) to quickly find similar words.

upvoted 2 times

 **Ajose0** 9 months, 1 week ago

Selected Answer: D

D. Download word embeddings pre-trained on a large corpus.

Word embeddings are a type of dense representation of words, which encode semantic meaning in a vector form. These embeddings are typically pre-trained on a large corpus of text data, such as a large set of books, news articles, or web pages, and capture the context in which words are used. Word embeddings can be used as features for a nearest neighbor model, which can be used to find words used in similar contexts.

Downloading pre-trained word embeddings is a good way to get started quickly and leverage the strengths of these representations, which have been optimized on a large amount of data. This is likely to result in more accurate and reliable features than other options like one-hot encoding, edit distance, or using Amazon Mechanical Turk to produce synonyms.

upvoted 6 times

 **loict** 9 months, 1 week ago

Selected Answer: D

- A. NO - one-hot encoding is a very early featurization stage
- B. NO - we don't want human labelling
- C. NO - too costly to do from scratch
- D. YES - leverage exiting training; the word embeddings will provide vectors than be used to measure distance in the downstream nearest neighbor model

upvoted 3 times

🗨️ 👤 **game_changer** 9 months, 1 week ago

Selected Answer: D

Pre-trained word embeddings, such as Word2Vec, GloVe, or FastText, capture the semantic and contextual meaning of words based on a large corpus of text data. By downloading pre-trained word embeddings, the Specialist can leverage the semantic relationships between words to provide meaningful word features for the nearest neighbor model powering the widget. Utilizing pre-trained word embeddings allows the model to understand and display words used in similar contexts effectively.

upvoted 2 times

🗨️ 👤 **game_changer** 9 months, 1 week ago

Selected Answer: D

A. One-hot word encoding vectors: These vectors represent words by marking them as present or absent in a fixed-length binary vector. However, they don't capture relationships between words or their meanings.

B. Producing synonyms: This would involve generating similar words for each word manually, which could be time-consuming and might not cover all possible contexts.

C. Word embedding vectors based on edit distance: This approach focuses on how similar words are in terms of their spelling or characters, not necessarily their meaning or context in sentences.

D. Downloading pre-trained word embeddings: These are vectors that represent words based on their contextual usage in a large dataset, capturing relationships between words and their meanings.

upvoted 5 times

🗨️ 👤 **elvin_ml_qayiran25091992razor** 1 year, 7 months ago

Selected Answer: D

correct D ay tupoy

upvoted 1 times

🗨️ 👤 **sonoluminescence** 1 year, 8 months ago

Selected Answer: D

words that are used in similar contexts will have vectors that are close in the embedding space

upvoted 1 times

🗨️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: D

D is correct

upvoted 1 times

🗨️ 👤 **DavidRou** 1 year, 11 months ago

I also believe that D is the correct answer. No reason to create word embeddings from scratch

upvoted 1 times

🗨️ 👤 **ortamina** 1 year, 11 months ago

Selected Answer: D

1. One-hot encoding will blow up the feature space - it is not recommended for a high cardinality problem domain.

2. One still needs to train the word features on large bodies of text to map context to each word

upvoted 1 times

🗨️ 👤 **Shailendraa** 2 years, 9 months ago

12-sep exam

upvoted 1 times

🗨️ 👤 **helpaws** 2 years, 10 months ago

Selected Answer: D

DDDDDDDDDDDDDD

upvoted 3 times

🗨️ 👤 **engomaradel** 3 years, 7 months ago

D for sure

upvoted 2 times

🗨️ 👤 **yeetusdeleetus** 3 years, 7 months ago

Definitely D.

upvoted 3 times

🗨️ 👤 **weslleylc** 3 years, 8 months ago

A)It requires that document text be cleaned and prepared such that each word is one-hot encoded.

Ref:<https://machinelearningmastery.com/what-are-word-embeddings/>

upvoted 1 times

A Machine Learning Specialist is configuring Amazon SageMaker so multiple Data Scientists can access notebooks, train models, and deploy endpoints. To ensure the best operational performance, the Specialist needs to be able to track how often the Scientists are deploying models, GPU and CPU utilization on the deployed SageMaker endpoints, and all errors that are generated when an endpoint is invoked. Which services are integrated with Amazon SageMaker to track this information? (Choose two.)

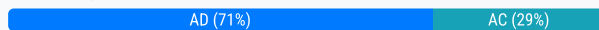
- A. AWS CloudTrail
- B. AWS Health
- C. AWS Trusted Advisor
- D. Amazon CloudWatch
- E. AWS Config

Suggested Answer: AD

Reference:

<https://aws.amazon.com/sagemaker/faqs/>

Community vote distribution



rsimham Highly Voted 3 years, 2 months ago

AD is correct

upvoted 27 times

eji Highly Voted 3 years, 2 months ago

CloudTrail is use to track scientist how ofthe they deploy a model

CloudWatch for monitoring GPU and CPU

so answer is A & D

upvoted 10 times

JonSno Most Recent 4 months, 1 week ago

Selected Answer: AD

To monitor SageMaker model deployments, track resource utilization, and log errors, the Machine Learning Specialist should use:

AWS CloudTrail – Tracks API activity, such as:

Model deployments (e.g., CreateModel, CreateEndpoint)

Notebook access and actions

SageMaker job executions

Amazon CloudWatch – Monitors and logs operational metrics, such as:

CPU & GPU utilization of SageMaker endpoints

Invocation errors and latencies

Custom metrics from deployed models

Logs from training jobs and inference endpoints (via CloudWatch Logs)

upvoted 2 times

VR10 10 months, 2 weeks ago

I think AWS Config is still not the service designed to track how often Data Scientists are deploying models, nor does it track operational performance metrics like GPU and CPU utilization or the invocation errors of SageMaker endpoints.

and AWS CloudTrail continues to be the service that will track and record user activity and API usage, which includes deploying models in Amazon SageMaker.

So the answers are still A and D - CloudTrail and CloudWatch.

upvoted 1 times

elvin_ml_qayiran25091992razor 1 year, 1 month ago

Selected Answer: AD

AD is correct

upvoted 2 times

🗨️ 👤 **loict** 1 year, 3 months ago

Selected Answer: AC

- A. YES - to track deployments
- B. NO - AWS Health is to track AWS Cloud itself (eg. is a zone down ?)
- C. NO - AWS Trusted Advisor to give recommendations on infra
- D. YES - for errors
- E. AWS Config

upvoted 2 times

🗨️ 👤 **DavidRou** 1 year, 3 months ago

I also believe that A and D are correct. Can someone please explain to me the main differences between CloudWatch and CloudTrail? I find the documentation a bit confusing about it

upvoted 1 times

🗨️ 👤 **CKS1210** 1 year, 6 months ago

Option E AWS Config to record all resource types, then the new resources will be automatically recorded in your account.

Option A CloudTrail is use to track scientist how of the they deploy a model

Option D CloudWatch for monitoring GPU and CPU

upvoted 1 times

🗨️ 👤 **joe3232** 1 year, 11 months ago

Log Amazon Sagemaker API Calls with AWS CloudTrail - <https://docs.aws.amazon.com/sagemaker/latest/dg/logging-using-cloudtrail.html>

upvoted 2 times

🗨️ 👤 **f4bi4n** 2 years, 6 months ago

I wouldn't be so sure about CloudTrail, AWS Configs also tracks Sagemaker and the resource "AWS::Sagemaker::Model"

upvoted 1 times

🗨️ 👤 **f4bi4n** 2 years, 6 months ago

just seen, this was release 4 days ago...

<https://aws.amazon.com/about-aws/whats-new/2022/06/aws-config-15-new-resource-types/>

upvoted 2 times

🗨️ 👤 **yogesh1** 2 years, 10 months ago

Selected Answer: AD

A&D

CloudWatch and ClouTrail

upvoted 3 times

🗨️ 👤 **hess** 2 years, 11 months ago

AD Are Correct.

upvoted 1 times

🗨️ 👤 **Urban_Life** 3 years, 2 months ago

absolutely

upvoted 1 times

🗨️ 👤 **roytruong** 3 years, 2 months ago

cloudtrail and cloudwatch, no thinking

upvoted 6 times

A retail chain has been ingesting purchasing records from its network of 20,000 stores to Amazon S3 using Amazon Kinesis Data Firehose. To support training an improved machine learning model, training records will require new but simple transformations, and some attributes will be combined. The model needs to be retrained daily.




Given the large number of stores and the legacy data ingestion, which change will require the LEAST amount of development effort?

- A. Require that the stores to switch to capturing their data locally on AWS Storage Gateway for loading into Amazon S3, then use AWS Glue to do the transformation.
- B. Deploy an Amazon EMR cluster running Apache Spark with the transformation logic, and have the cluster run each day on the accumulating records in Amazon S3, outputting new/transformed records to Amazon S3.
- C. Spin up a fleet of Amazon EC2 instances with the transformation logic, have them transform the data records accumulating on Amazon S3, and output the transformed records to Amazon S3.
- D. Insert an Amazon Kinesis Data Analytics stream downstream of the Kinesis Data Firehose stream that transforms raw record attributes into simple transformed values using SQL.

Suggested Answer: D

Community vote distribution

D (100%)

  **cybe001**  2 years, 9 months ago

D is correct. Question has "simple transformations, and some attributes will be combined" and Least development effort. Kinesis analytics can get data from Firehose, transform and write to S3

<https://docs.aws.amazon.com/kinesisanalytics/latest/java/examples-s3.html>

upvoted 49 times

  **mawsmann** 2 years, 9 months ago

Best explanation here, kudos.

upvoted 4 times

  **kakalotka** 2 years, 8 months ago

I can't find any information that indicate Kinesis data analytics taking data from firehose

upvoted 2 times

  **Huy**  2 years, 7 months ago

The best way to transform data is before it arrives to S3 so D should be best answer. But D is not completed. It should have another Firehose to deliver results to S3.

upvoted 9 times

  **JonSno**  4 months, 1 week ago

Selected Answer: D

D. Insert an Amazon Kinesis Data Analytics stream downstream of the Kinesis Data Firehose stream that transforms raw record attributes into simple transformed values using SQL.

Explanation:

Since the data is already flowing through Amazon Kinesis Data Firehose, the least development effort solution is to use Amazon Kinesis Data Analytics, which supports SQL-based transformations on streaming data without requiring new infrastructure.

Why is this the best choice?



No major architectural changes – Data continues flowing from stores into Kinesis Data Firehose and then to Amazon S3.

Simple SQL transformations – Since the changes are simple (e.g., attribute combinations), SQL is sufficient.

Low operational overhead – No need to manage clusters or instances.

Real-time processing – Transformed records immediately enter Amazon S3 for training.

upvoted 2 times

  **CKS1210** 1 year ago

Ans is D

Amazon Kinesis Data Analytics provides a serverless option for real-time data processing using SQL queries. In this case, by inserting a Kinesis Data

Analytics stream downstream of the Kinesis Data Firehose stream, the retail chain can easily perform the required simple transformations on the ingested purchasing records.

upvoted 1 times

🗳️ 👤 **Valcilio** 1 year, 3 months ago

Selected Answer: D

The best answer is to use a lambda, but the letter D can do it very good too in the absence of the lambda option.

upvoted 2 times

🗳️ 👤 **cloud_trail** 2 years, 8 months ago

I go with D. A tough question, though. And C are definitely out. The key to the question is that it does not say that the transformed data needs to be stored again in S3. It just needs to be sent to the model for training after being transformed. So a Kinesis Data Analytics stream is appropriate to do the transformation.

upvoted 1 times

🗳️ 👤 **harmanbirstudy** 2 years, 8 months ago

Legacy data -- Firehose -- Kinesis Analytics -- S3. This happens in near real time before the data ends up in S3.

--Legacy data -- Firehose -- S3 is already happening (mentioned in first line in question), adding Kinesis Data Analytics to do simple transformation joins using SQL on the incoming data is the LEAST amount of work needed.

Kinesis Data analytics can write to S3. here is the AWS link with working example. Even Though Udemy tutorial said it cannot write directly to S3 :).

<https://docs.aws.amazon.com/kinesisanalytics/latest/java/examples-s3.html>

upvoted 6 times

🗳️ 👤 **gamaX** 2 years, 8 months ago

It seems that LEAST development effort:

<https://aws.amazon.com/fr/blogs/big-data/preprocessing-data-in-amazon-kinesis-analytics-with-aws-lambda/>

and GREATEST development effort:

<https://aws.amazon.com/fr/blogs/big-data/optimizing-downstream-data-processing-with-amazon-kinesis-data-firehose-and-amazon-emr-running-apache-spark/>

upvoted 1 times

🗳️ 👤 **HaiHN** 2 years, 8 months ago

It's D

<https://aws.amazon.com/blogs/big-data/preprocessing-data-in-amazon-kinesis-analytics-with-aws-lambda/>

"In some scenarios, you may need to enhance your streaming data with additional information, before you perform your SQL analysis. Kinesis Analytics gives you the ability to use data from Amazon S3 in your Kinesis Analytics application, using the Reference Data feature. However, you cannot use other data sources from within your SQL query."

upvoted 1 times

🗳️ 👤 **h_sahu** 2 years, 8 months ago

I believe, kinesis should be used only in case of live data stream and this is not the case here. So as per me D shouldn't be the answer. I think A should be the answer as AWS storage gateway is something which is used along with on premise applications to move data to S3. Then glue can be used to transform the data.

upvoted 1 times

🗳️ 👤 **cloud_trail** 2 years, 8 months ago

With option A, you would be changing the legacy data ingestion, a huge development effort. Remember, you're talking about 20,000 stores.

upvoted 2 times

🗳️ 👤 **hans1234** 2 years, 8 months ago

It is D.

upvoted 1 times

🗳️ 👤 **dikers** 2 years, 8 months ago

I think the answer is D, because require the LEAST amount of development effort.

upvoted 1 times

🗳️ 👤 **roytruong** 2 years, 8 months ago

It's D, kinesis analytic can easily connect with firehose

upvoted 2 times

🗨️ 👤 **dreemswang** 2 years, 9 months ago

why not A. it seems good to me
upvoted 2 times

🗨️ 👤 **ExamTaker123456789** 2 years, 8 months ago

"require stores to capture data locally using S3 gateway" - for 20k stores this creates a HUUUGE operational overhead and development effort, definitely wrong
upvoted 3 times

🗨️ 👤 **PRC** 2 years, 9 months ago

D is correct...rest all need some kind of manual intervention as well as they are not simple..Firehose allows transformation as well as moving into S3
upvoted 6 times

🗨️ 👤 **devsean** 2 years, 9 months ago

I think the answer is B. D would be correct if they didn't want to transform the legacy data from before the switch, but it seems like they do. Choosing D would mean that you'd have to use an EC2 instance or something else to transform the legacy data along with adding the Kinesis data analytics functionality. Also, there is no real-time requirement so daily transformation is fine.
upvoted 3 times

🗨️ 👤 **hailiang** 2 years, 8 months ago

Its D, because with KDA you can transform the data with SQL while with EMR you need to write code, considering the requirement of "least development effort", so D
upvoted 3 times

🗨️ 👤 **devsean** 2 years, 9 months ago

I think the answer is B. D would be correct if they didn't want to transform the legacy data from before the switch, but it seems like they do. Choosing D would mean that you'd have to use an EC2 instance or something else to transform the legacy data along with adding the Kinesis data analytics functionality. Also, there is no real-time requirement so daily transformation is fine.
upvoted 7 times

🗨️ 👤 **ADVIT** 1 year ago

"LEAST amount of development effort", EMR is no complicated to LEAST
upvoted 1 times

🗨️ 👤 **ZSun** 1 year, 2 months ago

If the question is "least cost" then B, but the question is "least developement effort, then you want to keep original architecture. I agree that for daily ETL instead of real-time, and large dataset, B is better option.
upvoted 1 times

🗨️ 👤 **HaiHN** 2 years, 8 months ago

You can use Lambda instead of EC2. So D should be OK.
<https://aws.amazon.com/blogs/big-data/preprocessing-data-in-amazon-kinesis-analytics-with-aws-lambda/>
upvoted 1 times

🗨️ 👤 **am7** 2 years, 9 months ago

can be B
upvoted 1 times

A Machine Learning Specialist is building a convolutional neural network (CNN) that will classify 10 types of animals. The Specialist has built a series of layers in a neural network that will take an input image of an animal, pass it through a series of convolutional and pooling layers, and then finally pass it through a dense and fully connected layer with 10 nodes. The Specialist would like to get an output from the neural network that is a probability distribution of how likely it is that the input image belongs to each of the 10 classes.

Which function will produce the desired output?

- A. Dropout
- B. Smooth L1 loss
- C. Softmax
- D. Rectified linear units (ReLU)

Suggested Answer: C

Community vote distribution

C (100%)

🗳️ **DonaldCMLIN** Highly Voted 2 years, 9 months ago

C might be much suitable

softmax is to turn numbers into probabilities.

<https://medium.com/data-science-bootcamp/understand-the-softmax-function-in-minutes-f3a59641e86d>

upvoted 30 times

🗳️ **rsinham** Highly Voted 2 years, 9 months ago

C is right. Softmax function is used for multi-class predictoins

upvoted 14 times

🗳️ **JonSno** Most Recent 4 months, 1 week ago

Selected Answer: C

In a multiclass classification problem (such as classifying an image into one of 10 animal categories), the model should output a probability distribution over the classes. The Softmax function achieves this by:

Taking the raw scores (logits) from the final dense layer (10 nodes, one per class).

Exponentializing each score and normalizing them so they sum to 1, effectively turning them into probabilities.

upvoted 1 times

🗳️ **loict** 9 months, 2 weeks ago

Selected Answer: C

A. NO - Dropout is to prevent overfitting

B. NO - L1 regularization is to prevent overfitting

C. YES - Softmax will give probabilities for each class

D. NO - Rectified linear units (ReLU) is an activation function

upvoted 2 times

🗳️ **DavidRou** 9 months, 3 weeks ago

Softmax is the correct answer.

upvoted 1 times

🗳️ **Valcilio** 1 year, 3 months ago

Selected Answer: C

Multiclassification with probabilities is about softmax!

upvoted 1 times

🗳️ **vbal** 1 year, 5 months ago

Softmax is for probability distribution

upvoted 1 times

🗳️ **technoguy** 2 years, 7 months ago

it should be C. Softmax

Softmax converts outputs to Probabilites of each classification

upvoted 3 times

🗳️ 👤 **omar8024** 2 years, 8 months ago

absolutely C

upvoted 1 times

🗳️ 👤 **takahirokoyama** 2 years, 8 months ago

Absolute C.

upvoted 4 times

🗳️ 👤 **cloud_trail** 2 years, 8 months ago

This is as easy a question as you will likely see on the exam, Everyone has the right answer here.

upvoted 3 times

🗳️ 👤 **felbuch** 2 years, 8 months ago

C --> Softmax.

Let's go over the alternatives:

A. Dropout --> Not really a function, but rather a method to avoid overfitting. It consists of dropping some neurons during the training process, so that the performance of our algorithm does not become very dependent on any single neuron.

B. Smooth L1 loss --> It's a loss function, thus a function to be minimized by the entire neural network. It's not an activation function.

C. Softmax --> This is the traditional function used for multi-class classification problems (such as classifying an animal into one of 10 categories)

D. Rectified linear units (ReLU) --> This activation function is often used on the first and intermediate (hidden) layers, not on the final layer. In any case, it wouldn't make sense to use it for classification because its values can exceed 1 (and probabilities can't)

upvoted 11 times

🗳️ 👤 **MOMoez** 2 years, 8 months ago

C, Softmax is the best suitable answer

Ref: The softmax function, also known as softargmax[1]:184 or normalized exponential function,[2]:198 is a generalization of the logistic function to multiple dimensions. It is used in multinomial logistic regression and is often used as the last activation function of a neural network to normalize the output of a network to a probability distribution over predicted output classes, based on Luce's choice axiom.

upvoted 2 times

🗳️ 👤 **ybad** 2 years, 8 months ago

You guys are right, the answer is C since it automatically provides the output with a confidence interval...

Relu could be used as well but it needs to be coded in to provide the probabilities

<https://medium.com/@himanshuxd/activation-functions-sigmoid-relu-leaky-relu-and-softmax-basics-for-neural-networks-and-deep-8d9c70eed91e>

upvoted 1 times

🗳️ 👤 **yeetusdeleetus** 2 years, 8 months ago

Definitely C

upvoted 1 times

🗳️ 👤 **bidbs** 2 years, 8 months ago

Definitely softmax.

upvoted 1 times

🗳️ 👤 **hans1234** 2 years, 8 months ago

Are you sure it is C?

The output should be "[the probability that] the input image belongs to each of the 10 classes." And not the most likely class with the highest probability, which would be the result of softmax layer.

upvoted 1 times

🗳️ 👤 **hans1234** 2 years, 8 months ago

Yes, softmax returns indeed a vector of probabilities.

upvoted 1 times

A Machine Learning Specialist trained a regression model, but the first iteration needs optimizing. The Specialist needs to understand whether the model is more frequently overestimating or underestimating the target.

What option can the Specialist use to determine whether it is overestimating or underestimating the target value?

- A. Root Mean Square Error (RMSE)
- B. Residual plots
- C. Area under the curve
- D. Confusion matrix

Suggested Answer: B

Community vote distribution

B (100%)

🗳️ **vetal** Highly Voted 2 years, 3 months ago

RMSE says about the error value but not the sign of error. The question is to find whether the model overestimates or underestimates - I guess residual plots clearly show that

answer B

upvoted 37 times

🗳️ **rsimham** Highly Voted 2 years, 3 months ago

Answer is B. Residual plot distribution indicates over or under-estimations

upvoted 14 times

🗳️ **JonSno** Most Recent 4 months, 1 week ago

Selected Answer: B

A residual plot helps determine whether a regression model is overestimating or underestimating the target value.

Residual = Actual Value - Predicted Value

Positive residual → The model underestimated the target.

Negative residual → The model overestimated the target.

By plotting residuals, the Machine Learning Specialist can see patterns that indicate bias:

More positive residuals → The model is underestimating.

More negative residuals → The model is overestimating.

Randomly scattered residuals around zero → The model is well-calibrated.

upvoted 2 times

🗳️ **Valcilio** 9 months, 3 weeks ago

Selected Answer: B

Residual plots shows mistake by mistake!

upvoted 1 times

🗳️ **vetaal** 1 year, 11 months ago

Selected Answer: B

B - Residual plots it is - <https://docs.aws.amazon.com/machine-learning/latest/dg/regression-model-insights.html>

upvoted 4 times

🗳️ **felbuch** 2 years, 1 month ago

Residual Plots (B).

AUC and Confusion Matrices are used for classification problems, not regression.

And RMSE does not tell us if the target is being over or underestimated, because residuals are squared! So we actually have to look at the residuals themselves. And that's B.

upvoted 7 times

🗳️ **cnethers** 2 years, 1 month ago

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

- 1) Squaring the residuals.
- 2) Finding the average of the residuals.
- 3) Taking the square root of the result.

upvoted 3 times

  **cnethers** 2 years, 1 month ago

Residual Plots (B). would have to be my answer

upvoted 1 times

  **Thai_Xuan** 2 years, 1 month ago

residual plot

<https://docs.aws.amazon.com/machine-learning/latest/dg/regression-model-insights.html>

upvoted 5 times

  **syu31svc** 2 years, 2 months ago

[https://stattrek.com/statistics/dictionary.aspx?](https://stattrek.com/statistics/dictionary.aspx?definition=residual%20plot#:~:text=A%20residual%20plot%20is%20a,nonlinear%20model%20is%20more%20appropriate.)

[definition=residual%20plot#:~:text=A%20residual%20plot%20is%20a,nonlinear%20model%20is%20more%20appropriate.](https://stattrek.com/statistics/dictionary.aspx?definition=residual%20plot#:~:text=A%20residual%20plot%20is%20a,nonlinear%20model%20is%20more%20appropriate.)



Answer is B

upvoted 2 times

  **Antriksh** 2 years, 2 months ago

without a second thought residual plot



upvoted 2 times

  **qururu** 2 years, 2 months ago

The answer is B. Refer to Exercise 7.2.1.A

[https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_OpenIntro_Statistics_\(Diez_et_al\)./07%3A_Introduction_to_Linear_Regression/7.0](https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_OpenIntro_Statistics_(Diez_et_al)./07%3A_Introduction_to_Linear_Regression/7.0)

upvoted 1 times

  **C10ud9** 2 years, 2 months ago



Residual plot it is Option B

upvoted 1 times

  **roytruong** 2 years, 2 months ago

Residual plot

upvoted 2 times



  **deep_n** 2 years, 2 months ago

B is the correct answer!!!!

RMSE has the S in it that is square... that vanishes the above below factor of the prediction.

Answers C and D are for other type of problems

upvoted 4 times

  **swagy** 2 years, 2 months ago

It should be B. The residual plot will be give whether the target value is overestimated or underestimated.

upvoted 1 times

  **Jayraam** 2 years, 2 months ago

Answer is C.

<https://www.youtube.com/watch?v=MrjWcywVEiU>



upvoted 2 times

  **ExamTaker123456789** 2 years, 2 months ago

Answer is B.



Your vid shows a technique that is useful for defining integrals and has NOTHING to do linear regression. Also, it over-/underestimates the area under the curve, NOT the target value.

upvoted 2 times

  **cloud_trail** 2 years, 1 month ago

Good grief, AUC is used for classification not regression.

upvoted 1 times

  **PRC** 2 years, 2 months ago

B..Residual helps to find out whether the model is underestimating or overestimating

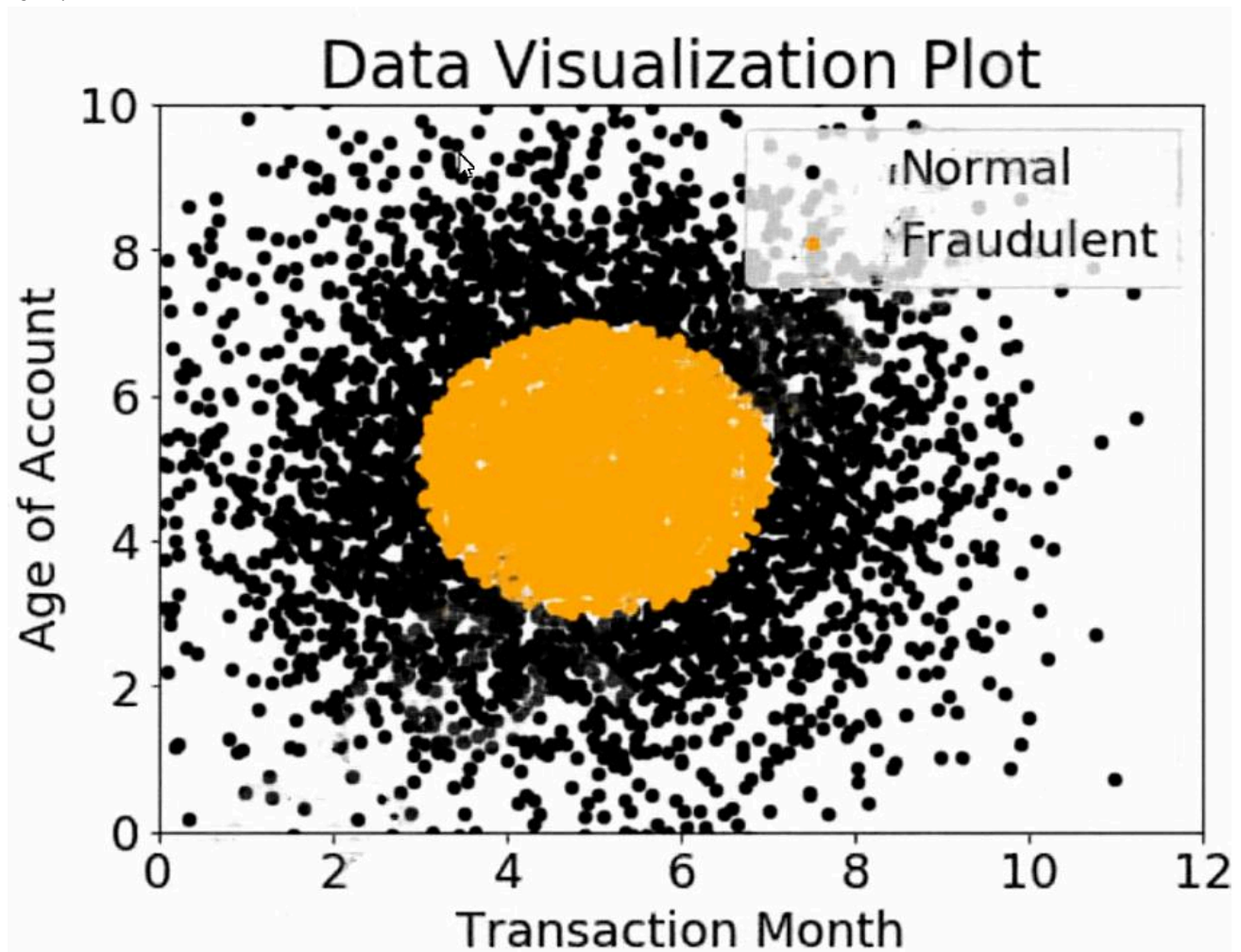
upvoted 3 times

  **AKT** 2 years, 2 months ago

answer is B

upvoted 2 times

A company wants to classify user behavior as either fraudulent or normal. Based on internal research, a Machine Learning Specialist would like to build a binary classifier based on two features: age of account and transaction month. The class distribution for these features is illustrated in the figure provided.

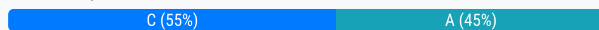


Based on this information, which model would have the HIGHEST recall with respect to the fraudulent class?

- A. Decision tree
- B. Linear support vector machine (SVM)
- C. Naive Bayesian classifier
- D. Single Perceptron with sigmoidal activation function

Suggested Answer: C

Community vote distribution



E_aws Highly Voted 3 years, 8 months ago

C is the correct answer because gaussian naive Bayes can do this nicely.
upvoted 12 times

E_aws 3 years, 8 months ago

of course it doesn't mention the gaussian here and refers to naive bayes in general, but I'm still positive with C.
upvoted 1 times

blubb Highly Voted 3 years, 8 months ago

Answer should be A:.

B: LINEAR SVM is a linear classifier

-> All of these have a linear decision boundary (so it's just a line $y = mx+b$). This leads to a bad recall and so A must be the right choice.
upvoted 9 times

JonSno Most Recent 4 months, 1 week ago

Selected Answer: A

Decision Tree (Best Choice) ✓

Highly flexible: Can capture non-linear decision boundaries, making it effective when the class distribution is not linearly separable.

Maximizes recall: A decision tree can prioritize minimizing false negatives by adjusting its splits.

Handles imbalanced classes well using class weighting or pruning techniques.

upvoted 3 times

MVAS 4 months, 3 weeks ago

Selected Answer: C

Gaussian naive Bayes is correct one

upvoted 1 times

MintTeaClarity 7 months, 2 weeks ago

Selected Answer: A

A non-linear problem would be a case where linear classifiers, such as naive Bayes, would not be suitable since the classes are not linearly separable.

In such a scenario, non-linear classifiers (e.g., instance-based nearest neighbour classifiers) should be preferred.

upvoted 1 times

egorkrash 8 months, 1 week ago

Selected Answer: A

decision tree can effectively maximize the recall by drawing a square ($3 \leq \text{month} \leq 7, 3 \leq \text{age} \leq 7$)

upvoted 2 times

MultiCloudIronMan 8 months, 1 week ago

Selected Answer: A

Option C. Naive Bayesian classifier is not the best choice for achieving the highest recall for the fraudulent class because it makes strong assumptions about the independence of features. In many real-world scenarios, especially with complex data like user behavior, these assumptions do not hold true, which can lead to suboptimal performance.

In contrast, a Decision tree (Option A) can handle feature interactions and is more flexible in capturing the relationships between features, making it more effective in identifying fraudulent behavior and achieving higher recall

upvoted 1 times

ML_2 10 months, 2 weeks ago

Selected Answer: A

Answer in my opinion is A

A Decision Tree Classifier can handle complex decision boundaries and does not assume any particular distribution of data. It is well-suited for cases like this where the decision boundary is non-linear, as seen with the clear separation between the normal and fraudulent transactions.

A Naive Bayesian classifier, on the other hand, assumes independence among features and typically performs better when data is normally distributed, which might not be the case here given the data's clustering pattern.

upvoted 1 times

ninomfr64 1 year ago

Selected Answer: C

From Claude 3 Haiku:

A. NO, decision trees may struggle to capture the linear separability of the classes.

B. NO, Linear SVM may not be able to fully exploit the class separation due to its linear decision boundary.

C. YES, The Naive Bayesian classifier tends to perform well in situations where the classes are linearly separable. This model requires the features are independent and this is the case

D. The single Perceptron with a sigmoidal activation function may not be able to capture the complex class distributions as effectively as the Naive Bayesian classifier.

upvoted 1 times

GrumpyApple 7 months, 1 week ago

Funny that if you ask Haiku to explain its reason step by step, it will chose A instead of C

...

Based on the information provided, the model that is likely to have the highest recall with respect to the fraudulent class is the **Decision Tree

(Most Voted)**.

...

upvoted 1 times

🗳️ 👤 **iambasspaul** 1 year, 2 months ago

Selected Answer: C

Answer by Claude3:

In contrast, the Decision Tree (A) and Linear SVM (B) models are generally more robust to overfitting and can achieve a better balance between recall and precision, but they may not necessarily have the highest recall for the minority class.

Considering the importance of maximizing recall for the fraudulent class in this use case, the Naive Bayesian Classifier (C) could be a valid choice, although it may come with the trade-off of lower precision and potentially higher false positive rates.

upvoted 1 times

🗳️ 👤 **rav009** 1 year, 4 months ago

highest recall.

So A

upvoted 1 times

🗳️ 👤 **notbother123** 1 year, 4 months ago

Selected Answer: A

Only A (DT) is non-linear among the mentioned algorithms.

upvoted 1 times

🗳️ 👤 **kyuhuck** 1 year, 4 months ago

Selected Answer: A

Given the visualized data, the Decision tree (Option A) is likely the best model to achieve the highest recall for the fraudulent class. It can handle complex patterns and create rules that are more suited for clustered and potentially non-linearly separable classes. Recall is a measure of a model's ability to capture all actual positives, and a decision tree can be tuned to prioritize capturing more of the fraudulent cases at the expense of making more false-positive errors on the normal cases.

upvoted 1 times

🗳️ 👤 **phdykd** 1 year, 5 months ago

if it was highest precision:

Given these considerations, the best model for precision would likely be a Support Vector Machine with a non-linear kernel, such as the RBF (Radial Basis Function) kernel. This model can tightly fit the boundary around the fraudulent class, minimizing the inclusion of normal transactions in the fraudulent prediction space, and thus potentially achieving high precision. Precision is sensitive to the false positives, and the flexibility of SVMs with non-linear kernels to create a tight and precise boundary can help to minimize these.

upvoted 1 times

🗳️ 👤 **phdykd** 1 year, 5 months ago

GPT 4 Answer is Decision Tree.

Considering the goal is to achieve the highest recall for the fraudulent class, which means we aim to capture as many fraudulent cases as possible even if it means getting more false positives, a Decision Tree would likely be the best option. This is because it can adapt to the complex shape of the class distribution and encapsulate the majority of the fraudulent class within its decision boundaries. Recall is a measure of a model's ability to capture all actual positives, and the decision tree's complex boundary setting capabilities make it well-suited for maximizing recall in this case.

upvoted 2 times

🗳️ 👤 **taustin2** 1 year, 7 months ago

Selected Answer: A

I'm going with A. As pointed out in this article, Naive Bayes performs poorly with non-linear classification problems. The picture shows a case where the classes are not linearly separable. Decision Tree will probably give better results.

https://sebastianraschka.com/Articles/2014_naive_bayes_1.html

upvoted 3 times

🗳️ 👤 **akgarg00** 1 year, 7 months ago

Selected Answer: A

Highest recall for fraudulent class means that Precision for Fraudulent predictions can be low. So basically just two conditions Transaction Month nearly greater than 8 and age of accounts greater than 8 can help identify the fraudulent class but it will classify most of non-fraudulent cases as fraudulent.

upvoted 2 times

A Machine Learning Specialist kicks off a hyperparameter tuning job for a tree-based ensemble model using Amazon SageMaker with Area Under the ROC Curve (AUC) as the objective metric. This workflow will eventually be deployed in a pipeline that retrains and tunes hyperparameters each night to model click-through on data that goes stale every 24 hours.

With the goal of decreasing the amount of time it takes to train these models, and ultimately to decrease costs, the Specialist wants to reconfigure the input hyperparameter range(s).

Which visualization will accomplish this?

- A. A histogram showing whether the most important input feature is Gaussian.
- B. A scatter plot with points colored by target variable that uses t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the large number of input variables in an easier-to-read dimension.
- C. A scatter plot showing the performance of the objective metric over each training iteration.
- D. A scatter plot showing the correlation between maximum tree depth and the objective metric.

Suggested Answer: B

Community vote distribution

D (75%)

C (25%)

  **cloud_trail** Highly Voted 3 years, 8 months ago

This is a very tricky question. The idea is to reconfigure the ranges of the hyperparameters. A refers to a feature, not a hyperparameter. A is out. C refers to training the model, not optimizing the range of hyperparameters. C is out. Now it gets tricky. D will let you find determine what the approximately best tree depth is. That's good. That's what you're trying to do but it's only one of many hyperparameters. It's the best choice so far. B is tricky. t-SNE does help you visualize multidimensional data but option B refers to input variables, not hyperparameters. For this very tricky question, I would do with D. It's the only one that accomplishes the task of limiting the range of a hyperparameter, even if it is only one of them.

upvoted 50 times

  **cnethers** 3 years, 8 months ago


It's good to see someone keeping a thoughtful and curious mind to this question. I too have the same conclusion, not an easy question.

upvoted 3 times

  **ovokpus** 3 years ago

But, how do you optimize hyperparameters without training experiments? That is why C is the best option. You get a value for each unique combination of hyperparameters.

upvoted 1 times

  **Dr_Kiko** 3 years, 7 months ago

B is also wrong as t-SNE picture is not actionable - good visual but ... that's it. try pictures here <https://lvdmaaten.github.io/tsne/>

upvoted 1 times

  **AddiWei** 3 years, 4 months ago


When you are tuning hyperparameters you are literally training multiple models and searching for the best ones.

upvoted 2 times

  **heihei** Highly Voted 3 years, 9 months ago

B doesn't make sense
I think it's D

upvoted 14 times

  **JonSno** Most Recent 4 months, 1 week ago

Selected Answer: D

The goal is to reduce training time and costs by optimizing the hyperparameter tuning process. In tree-based ensemble models (e.g., XGBoost, Random Forest, or Gradient Boosting), tree depth is one of the most influential hyperparameters affecting:

Model complexity: Deeper trees increase complexity but may lead to overfitting.

Training time: More depth means more splits, significantly increasing computation.

Performance (AUC score in this case): There is typically an optimal depth that balances underfitting and overfitting.

A scatter plot showing the correlation between tree depth and the AUC metric will allow the ML Specialist to:

Identify whether increasing depth leads to diminishing returns.

Choose an optimal tree depth that balances performance with training efficiency.

Reduce the search space of hyperparameters, speeding up tuning and lowering costs.

upvoted 1 times

🗨️ **ninomfr64** 1 year ago

A. No, doesn't help to set/reduce hyperparameter value/range

B. No, honestly this is gibberish to me

C. No, doesn't help to reduce hyperparameter value range

D. YES, this help me understand how to set max tree depth hyperparameter

upvoted 1 times

🗨️ **VR10** 1 year, 4 months ago

Option C.

See it is doing a scatter plot on the metric for each iteration.

Each iteration is running with a certain set of hyper parameters.

So if I plot this. and I find which iteration has the best metric, I could simply pick up those set of hyperparameters.

D will only led to the tuning of maximum tree depth.

I am not sure which option would satisfy the goal to decrease cost but just looking at maximum tree depth doesnt seem right to me. It might be a way to just look at the tree depth and tune just that parameter and since you are only tuning 1 paramter, it may be cheaper, but would that lead to a usable model?

I think it should be option C.

upvoted 1 times

🗨️ **Regu7** 1 year, 5 months ago

On what basis the correct answers are provided in this platform? Are they assuming this is the correct answer or it is taken from somewhere ?

upvoted 1 times

🗨️ **elvin_ml_qayiran25091992razor** 1 year, 7 months ago

Selected Answer: D

D IS THE CORRECT

upvoted 1 times

🗨️ **Reju** 1 year, 9 months ago

Selected Answer: C

Option D, can also be useful in hyperparameter tuning for tree-based ensemble models, especially if the maximum tree depth is one of the hyperparameters you want to optimize.

However, when the goal is to decrease training time and costs by reconfiguring input hyperparameter ranges, a scatter plot showing the performance of the objective metric over each training iteration (Option C) is generally more directly related to the hyperparameter tuning process. It helps you track how the model's performance changes during hyperparameter tuning, which is critical for making decisions about which hyperparameter ranges to explore further.

Option D is valuable for understanding the relationship between maximum tree depth and the objective metric, but it might not provide as comprehensive insights into the overall hyperparameter tuning process compared to Option C.

upvoted 1 times

🗨️ **loict** 1 year, 9 months ago

Selected Answer: D

A. NO - it is about data discovery

B. NO - it is about data discovery

C. MIGHT - (NO) is a training iteration the overnight training the question is referring to ? (YES) Or each HPO training within each night ?

D. YES - the less ambiguous answer

upvoted 1 times

🗨️ **DavidRou** 1 year, 9 months ago

I think that C should be the right answer. The specialist can monitor how the model works by changing hyperparameters' values in each training iteration.

upvoted 1 times

🗨️ **Mickey321** 1 year, 10 months ago

Selected Answer: D

Option D

upvoted 1 times

🗨️ 👤 **kaike_reis** 1 year, 11 months ago

Selected Answer: D

A and B are wrong, because is totally out of question context. C is for monitoring a model, it doesn't help to change your HP range. D is the only answer that applies to the question.

upvoted 3 times

🗨️ 👤 **Venkatesh_Babu** 1 year, 11 months ago

Selected Answer: C

I think it should be c

upvoted 1 times

🗨️ 👤 **CKS1210** 2 years ago

Selected Answer: C

By plotting the performance of the objective metric (AUC) over each training iteration, the Specialist can analyze how different hyperparameter configurations affect the model's performance. This visualization helps in understanding which hyperparameter combinations lead to better results and allows the Specialist to identify areas of improvement.

upvoted 1 times

🗨️ 👤 **mirik** 2 years ago

D: By analyzing this relationship, the Specialist can adjust the range of maximum tree depth values used during hyperparameter tuning to decrease training time and costs.

upvoted 1 times

🗨️ 👤 **earthMover** 2 years, 1 month ago

Selected Answer: D

D Seems like the best answer. When answer is considered correct who is making that call an is there any justification provided for us to learn from?

upvoted 2 times

🗨️ 👤 **Valcilio** 2 years, 3 months ago

Selected Answer: D

It's about parameters, not about dimensionality.

upvoted 2 times

A Machine Learning Specialist is creating a new natural language processing application that processes a dataset comprised of 1 million sentences. The aim is to then run Word2Vec to generate embeddings of the sentences and enable different types of predictions.

Here is an example from the dataset:

"The quck BROWN FOX jumps over the lazy dog."

Which of the following are the operations the Specialist needs to perform to correctly sanitize and prepare the data in a repeatable manner? (Choose three.)

- A. Perform part-of-speech tagging and keep the action verb and the nouns only.
- B. Normalize all words by making the sentence lowercase.
- C. Remove stop words using an English stopword dictionary.
- D. Correct the typography on "quck" to "quick".
- E. One-hot encode all words in the sentence.
- F. Tokenize the sentence into words.

Suggested Answer: BCF

Community vote distribution

BCF (100%)

 **ozan11** Highly Voted 2 years, 9 months ago

B C F should be correct.

upvoted 35 times

 **BigEv** Highly Voted 2 years, 9 months ago

I will select B, C, F

- 1- Apply words stemming and lemmatization
- 2- Remove Stop words
- 3- Tokensize the sentences

<https://towardsdatascience.com/nlp-extracting-the-main-topics-from-your-dataset-using-lda-in-minutes-21486f5aa925>

upvoted 26 times

 **Togy** Most Recent 3 months, 2 weeks ago

Selected Answer: BDF

B. Normalize all words by making the sentence lowercase:

Word2Vec treats words as distinct entities. If you don't convert everything to lowercase, "The" and "the" will be considered different words, which is generally not what you want. Lowercasing ensures consistency.


D. Correct the typography on "quck" to "quick":

Misspellings need to be corrected. Word2Vec learns embeddings based on the words it encounters. If "quck" remains, it will be treated as a separate word from "quick," and you'll lose the relationship between them. Correcting typos is crucial for data quality.

F. Tokenize the sentence into words:

Tokenization is the process of breaking down the sentence into individual words (or tokens). Word2Vec operates on individual words, so you need to split the sentence into its constituent parts. This is a fundamental step in NLP.

upvoted 1 times

 **JonSno** 4 months, 1 week ago

Selected Answer: BDF

While C - is debatable - not always necessary to remove stop words in Word2Vec - as sometimes the stop words do provide context

=====

For Word2Vec training, data preprocessing is essential to ensure that words are correctly represented, consistent, and free from unnecessary noise. The key steps are:

Lowercasing the text (B)

Word embeddings treat "FOX" and "fox" as different words. To avoid redundancy, lowercasing the text ensures consistency.

Correcting typos (D)

"quck" should be corrected to "quick" to prevent incorrect word representations in Word2Vec. Misspelled words can create meaningless embeddings.

Tokenizing the sentence into words (F)

Word2Vec operates at the word level, so breaking the sentence into individual tokens (words) is necessary.

upvoted 2 times

🗳️ 👤 **loict** 9 months, 2 weeks ago

Selected Answer: BCF

A. NO - word2vec works on raw data

B. YES - case here is not significant

C. YES - will help reduce dimensionality

D. NO - word2vec will do it by itself

E. NO - One-hot encoding is for classification

F. YES - word2vec takes tokens as input

upvoted 1 times

🗳️ 👤 **Valcilio** 1 year, 3 months ago

Selected Answer: BCF

Data need to be tokenized and cleaned!

upvoted 2 times

🗳️ 👤 **Aninina** 1 year, 6 months ago

Selected Answer: BCF

B, C F is the correct

upvoted 2 times

🗳️ 👤 **SophieSu** 2 years, 8 months ago

BCF correct. D is not correct (Pay attention to "in a repeatable manner" in the question.)

upvoted 2 times

🗳️ 👤 **cloud_trail** 2 years, 8 months ago

B/C/F. D should not be performed because spell check is a subjective thing. You don't know for sure what the word was supposed to be if you have a typo.

upvoted 2 times

🗳️ 👤 **harmanbirstudy** 2 years, 8 months ago

I saw this exact question on "whizlabs" practice exam and correct options were B/C/F

upvoted 1 times

🗳️ 👤 **GeeBeeEl** 2 years, 8 months ago

<https://towardsdatascience.com/an-implementation-guide-to-word2vec-using-numpy-and-google-sheets-13445eebd281>

Data Preparation — Define corpus, clean, normalise and tokenise words

To begin, we start with the following corpus:

"natural language processing and machine learning is fun and exciting"

For simplicity, we have chosen a sentence without punctuation and capitalization. Also, we did not remove stop words "and" and "is".

In reality, text data are unstructured and can be "dirty". Cleaning them will involve steps such as

o removing stop words,

o removing punctuations,

o convert text to lowercase (actually depends on your use-case),

o replacing digits, etc.

o After preprocessing, we then move on to tokenising the corpus

Answer: B, C, F

upvoted 8 times

🗳️ 👤 **cnethers** 2 years, 8 months ago

BCF is 100% correct

upvoted 2 times

🗨️ 👤 **Antriksh** 2 years, 8 months ago

Correct answers are B, C and F
upvoted 2 times

🗨️ 👤 **TuanAnh** 2 years, 8 months ago

The correct answer is B, C and F
A: POS tagging has nothing to do with word2vec
D: fixing "quck" to "quick" only works for that specific word
F: word2vec can use CBOW or skipgram, so no need to have one-hot decoding here
upvoted 4 times

🗨️ 👤 **TuanAnh** 2 years, 8 months ago

sorry E: word2vec can use CBOW or skipgram, so no need to have one-hot decoding here
upvoted 4 times

🗨️ 👤 **PRC** 2 years, 8 months ago

BCF is correct
upvoted 2 times

🗨️ 👤 **AKT** 2 years, 9 months ago

B, C F correct
upvoted 2 times

🗨️ 👤 **Phong** 2 years, 9 months ago

B, C, and F are correct answers. I have done this question many times in many practice tests.
upvoted 12 times

🗨️ 👤 **tap123** 2 years, 9 months ago

B, C, F are my choice. D is also possible but not as widely used as others.
upvoted 3 times

A company is using Amazon Polly to translate plaintext documents to speech for automated company announcements. However, company acronyms are being mispronounced in the current documents.
How should a Machine Learning Specialist address this issue for future documents?

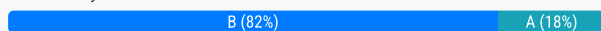
- A. Convert current documents to SSML with pronunciation tags.
- B. Create an appropriate pronunciation lexicon.
- C. Output speech marks to guide in pronunciation.
- D. Use Amazon Lex to preprocess the text files for pronunciation

Suggested Answer: A

Reference:

<https://docs.aws.amazon.com/polly/latest/dg/ssml.html>

Community vote distribution



VB Highly Voted 3 years, 3 months ago

SSML is specific to that particular document, like W3C can be pronounced as "World Wide Web Consortium" using `_{W3C}` in that specific document and when you create a new document, you need to format again. But with LEXICONS, you can upload a lexicon file once and ALL the FUTURE documents can just have W3C and that will be pronounced as "World Wide Web Consortium".. so answer is B, because the question asks for "future" documents.

upvoted 44 times

khchan123 1 year, 1 month ago

The correct answer is B, as explained by VB.

upvoted 1 times

cloud_trail 3 years, 2 months ago

For the exact reason you state, the correct answer is A. For every different document, a particular acronym may mean something different so you must have a solution that is document-specific.

upvoted 3 times

LeoD 6 months, 2 weeks ago

As the question stated "address this issue FOR FUTURE DOCUMENTS". B addresses for future. A only address the issue in a case-by-case manner.

upvoted 1 times

ovokpus 2 years, 6 months ago

It is the same business, so the acronyms are not expected to change from document to document

upvoted 3 times

VR10 10 months, 2 weeks ago

absolutely, B is the correct choice.

upvoted 1 times

Madwyn 3 years, 2 months ago

A.The document section for "Pronouncing Acronyms and Abbreviations".

Source: <https://docs.aws.amazon.com/polly/latest/dg/supportedtags.html>

upvoted 3 times

cybe001 Highly Voted 3 years, 3 months ago

I think the answer is B.

<https://docs.aws.amazon.com/polly/latest/dg/managing-lexicons.html>

<https://www.smashingmagazine.com/2019/08/text-to-speech-aws/>

upvoted 18 times

Tony_1406 2 years, 3 months ago

Lifted from the above link - "Your text might include an acronym, such as W3C. You can use a lexicon to define an alias for the word W3C so that it is read in the full, expanded form (World Wide Web Consortium)."

Clearly this is the same use case.

upvoted 1 times

🗨️ **JonSno** Most Recent 4 months, 1 week ago

Selected Answer: B

Explanation:

Amazon Polly sometimes mispronounces acronyms because it reads them as regular words. The best way to correct mispronunciations in future documents is to create a pronunciation lexicon. This allows you to define how specific words, acronyms, or abbreviations should be pronounced.

How to Use a Pronunciation Lexicon in Amazon Polly?

Define the correct pronunciation of acronyms in a Lexicon XML file.

Use phonetic notation (e.g., IPA or Speech Synthesis Markup Language (SSML) Phoneme tags).

Upload the lexicon to Polly via the AWS Management Console or AWS SDK.

Reference the lexicon in Polly API requests.

upvoted 3 times

🗨️ **VasuKVK** 8 months, 3 weeks ago

Answer : B

<https://aws.amazon.com/blogs/machine-learning/customize-pronunciation-using-lexicons-in-amazon-polly/>

Use <phoneme> SSML tag which is great for inserting one-off customizations or testing purposes. We recommend using Lexicon to create a consistent set of pronunciations for frequently used words across your organization. This enables your content writers to spend time on writing instead of the tedious task of adding phonetic pronunciations in the script repetitively.

upvoted 1 times

🗨️ **WTSpl** 10 months, 1 week ago

SSML supports phonetic pronunciation. Seems to me A is correct too.

<https://docs.aws.amazon.com/polly/latest/dg/supportedtags.html#phoneme-tag>

upvoted 1 times

🗨️ **phdykd** 11 months, 4 weeks ago

B IS ANSWER

upvoted 1 times

🗨️ **elvin_ml_qayiran25091992razor** 1 year, 1 month ago

Selected Answer: B

B is the correct, A hardan cixdi debil?

upvoted 1 times

🗨️ **DavidRou** 1 year, 3 months ago

This issue can be faced with both methods described in A and B. Though the answer A refers to the "current" document while the question regards "future" documents, so I think the right answer is B.

upvoted 1 times

🗨️ **kaike_reis** 1 year, 5 months ago

Selected Answer: B

Letter B is correct to ensure that acronyms or terms are pronounced correctly. Letter A works, but look at the catch: It's asked for future documents, but it mentions converting only current ones to SSML format, while future ones would be in plaintext.

upvoted 2 times

🗨️ **ADVIT** 1 year, 6 months ago

Company using plaintext and Future document means plaintext!

So only Custom Lexicon will help.

upvoted 1 times

🗨️ **soonmo** 1 year, 6 months ago

Selected Answer: B

<https://docs.aws.amazon.com/polly/latest/dg/managing-lexicons.html>

this explains acronym.



upvoted 1 times

🗨️ **earthMover** 1 year, 7 months ago

Selected Answer: B

I believe it should be lexicon. Can you share how you tag the correct answer?

upvoted 1 times

  **Sylzys** 1 year, 9 months ago

Selected Answer: B



Key here being "for future documents", answer should be B as SSML is for a specific document only

upvoted 3 times

  **Chelseajcole** 1 year, 10 months ago

this should be multiple choice question which answer is a AND b

upvoted 1 times

  **bakarys** 1 year, 10 months ago

Selected Answer: B

response B

A pronunciation lexicon is a list of words and their correct phonetic pronunciation that can be used to improve the accuracy of text-to-speech conversion. In this case, the Machine Learning Specialist can create a custom lexicon for the company's acronyms and upload it to Amazon Polly. This will ensure that the acronyms are pronounced correctly in the future announcements.

upvoted 2 times

  **SK27** 2 years ago

Selected Answer: B

Should be B

upvoted 2 times

  **masoa3b** 2 years, 1 month ago

Selected Answer: B

With Amazon Polly's custom lexicons or vocabularies, you can modify the pronunciation of particular words, such as company names, acronyms, foreign words, etc. To customize these pronunciations, you upload an XML file with lexical entries. `{rpmimcoatopm ;exocpms}` enable you to customize the pronunciation of words. Amazon Polly provides API operations that you can use to store lexicons in an AWS region. Those lexicons are then specific to that particular region.

References:

<https://docs.aws.amazon.com/polly/latest/dg/managing-lexicons.html>

<https://aws.amazon.com/blogs/machine-learning/create-accessible-training-with-initiafy-and-amazon-polly/>

upvoted 3 times

An insurance company is developing a new device for vehicles that uses a camera to observe drivers' behavior and alert them when they appear distracted. The company created approximately 10,000 training images in a controlled environment that a Machine Learning Specialist will use to train and evaluate machine learning models.

During the model evaluation, the Specialist notices that the training error rate diminishes faster as the number of epochs increases and the model is not accurately inferring on the unseen test images.

Which of the following should be used to resolve this issue? (Choose two.)

- A. Add vanishing gradient to the model.
- B. Perform data augmentation on the training data.
- C. Make the neural network architecture complex.
- D. Use gradient checking in the model.
- E. Add L2 regularization to the model.

Suggested Answer: BE



Community vote distribution

BE (100%)

  **vetal**  3 years, 3 months ago

The model must have been overfitted. Regularization helps to solve the overfitting problem in machine learning (as well as data augmentation). Correct answers should be BE.

upvoted 36 times

  **rajs** 3 years, 2 months ago



Agreed 100%

upvoted 5 times

  **jasonsunbao** 3 years, 3 months ago

agree on BE

upvoted 3 times

  **benson2021**  3 years, 2 months ago

Answer: BE

<https://www.kdnuggets.com/2019/12/5-techniques-prevent-overfitting-neural-networks.html>

5 techniques to prevent overfitting:

1. Simplifying the model
2. Early stopping
3. Use data argumentation
4. Use regularization
5. Use dropouts

upvoted 15 times

  **JonSno**  4 months, 1 week ago

Selected Answer: BE

he issue described suggests that the model is overfitting to the training data:

Training error decreases quickly, meaning the model is learning the training set very well.

Poor performance on unseen test data, indicating overfitting.

To resolve overfitting, the Machine Learning Specialist should:

Perform Data Augmentation (B)

Expands the training dataset artificially by applying transformations (e.g., rotations, flips, brightness changes, cropping).

Helps the model generalize better by exposing it to more diverse variations of the same class.

Add L2 Regularization (E)

Also known as weight decay, it penalizes large weights, preventing the model from memorizing the training data. Encourages simpler models, which reduces variance and improves generalization.

upvoted 2 times

🗳️ 👤 **delfoxete** 10 months, 3 weeks ago

Selected Answer: BE

agreed with vetal

upvoted 1 times

🗳️ 👤 **loict** 1 year, 3 months ago

Selected Answer: BE

A. NO - vanishing gradient is somebody bad they might happen and prevent convergence, we don't want that or something we can add explicitly. it is a result of the learning

B. YES - we have a overfitting problem so more training examples will help

C. NO - we already have good accuracy on the training set

D. NO - gradient checking is to find bugs in model implementation

E. YES - we have a overfitting problem

upvoted 2 times

🗳️ 👤 **John_Pongthorn** 2 years, 10 months ago

B. Perform data augmentation on the training data. (it should add validation data as well)

data should be distributed among train validation and test.

upvoted 1 times

🗳️ 👤 **KM226** 2 years, 12 months ago

Selected Answer: BE

Answer B&E looks good

upvoted 2 times

🗳️ 👤 **engomaradel** 3 years, 2 months ago

B & E is the correct ans

upvoted 1 times

🗳️ 👤 **roytruong** 3 years, 2 months ago

BE is exact

upvoted 3 times

🗳️ 👤 **stamarpadar** 3 years, 2 months ago

BE are the correct answers

upvoted 4 times

🗳️ 👤 **VB** 3 years, 2 months ago

Looks like B and D are correct.. For D -> <https://www.youtube.com/watch?v=P6EtCVrvYPU>

upvoted 3 times

🗳️ 👤 **C10ud9** 3 years, 2 months ago

gradient checking doesn't resolve the issue, but adding it will confirm / deny the issue. So, it helps to validate the issue but not resolve. I would say

B, E are correct

upvoted 3 times

🗳️ 👤 **VB** 3 years, 2 months ago

L2 regularization tries to reduce the possibility of overfitting by keeping the values of the weights and biases small.

upvoted 3 times

🗳️ 👤 **hughhughhugh** 3 years, 2 months ago

why not because of vanishing gradient?

upvoted 1 times

🗳️ 👤 **lt626** 1 year, 11 months ago

Vanishing gradients are a problem when training a NN. Answer A mentions that the solution should be to add that, which is not possible. Correct solution is BE.

<https://www.kdnuggets.com/2022/02/vanishing-gradient-problem.html>

upvoted 1 times

🗳️ 👤 **PRC** 3 years, 2 months ago

This is L2 Regularization....Do you think this is the right answer?

upvoted 1 times

  **WWODIN** 3 years, 3 months ago

agree BE

upvoted 3 times

When submitting Amazon SageMaker training jobs using one of the built-in algorithms, which common parameters **MUST** be specified? (Choose three.)

- A. The training channel identifying the location of training data on an Amazon S3 bucket.
- B. The validation channel identifying the location of validation data on an Amazon S3 bucket.
- C. The IAM role that Amazon SageMaker can assume to perform tasks on behalf of the users.
- D. Hyperparameters in a JSON array as documented for the algorithm used.
- E. The Amazon EC2 instance class specifying whether training will be run using CPU or GPU.
- F. The output path specifying where on an Amazon S3 bucket the trained model will persist.

Suggested Answer: AEF

Community vote distribution



🗳️ 👤 **DonaldCMLIN** Highly Voted 3 years, 9 months ago

THE ANSWER SHOULD BE CEF

IAM ROLE, INSTANCE TYPE, OUTPUT PATH

upvoted 29 times

🗳️ 👤 **hamimelon** 2 years, 6 months ago

Why not A? You don't need to tell Sagemaker where the training data is located?

upvoted 3 times

🗳️ 👤 **ZSun** 2 years, 2 months ago

You need to specify the InputDataConfig, but it does not need to be "S3"

I think the reason why A and B are wrong, not because data location is not required, but because it doesn't need to be S3, it can be Amazon S3, EFS, or FSx location

upvoted 1 times

🗳️ 👤 **HaiHN** 3 years, 8 months ago

Should be C, E, F

From the SageMaker notebook example:

https://github.com/aws/amazon-sagemaker-examples/blob/master/introduction_to_amazon_algorithms/semantic_segmentation_pascalvoc/semantic_segmentation_pascalvoc.ipynb

Create the sagemaker estimator object.

```
ss_model = sagemaker.estimator.Estimator(training_image,
role,
train_instance_count = 1,
train_instance_type = 'ml.p3.2xlarge',
train_volume_size = 50,
train_max_run = 360000,
output_path = s3_output_location,
base_job_name = 'ss-notebook-demo',
sagemaker_session = sess)
```

upvoted 12 times

🗳️ 👤 **unitit** 2 years, 4 months ago

It says InstanceClass - CPU/GPU in the question, not InstanceType

upvoted 6 times

🗳️ 👤 **mirik** 2 years ago

instance type has default value.

upvoted 3 times

VB Highly Voted 3 years, 8 months ago

From here https://docs.aws.amazon.com/zh_tw/sagemaker/latest/dg/API_CreateTrainingJob.html .. the only "Required: Yes" attributes are:

1. AlgorithmSpecification (in this TrainingInputMode is Required - i.e. File or Pipe)
2. OutputDataConfig (in this S3OutputPath is Required - where the model artifacts are stored)
3. ResourceConfig (in this EC2 InstanceType and VolumeSizeInGB are required)
4. RoleArn (..The Amazon Resource Name (ARN) of an IAM role that Amazon SageMaker can assume to perform tasks on your behalf...the caller of this API must have the iam:PassRole permission.)
5. StoppingCondition
6. TrainingJobName (The name of the training job. The name must be unique within an AWS Region in an AWS account.)

From the given options in the questions.. we have 2, 3, and 4 above. so, the answer is CEF.

upvoted 27 times

cloud_trail 3 years, 8 months ago

This is the best explanation that CEF is the right answer, IMO. The document at that url is very informative. It also specifically states that InputDataConfig is NOT required. Having said that, I have no idea how the model will train if it doesn't know where to find the training data, but that is what the document says. If someone can explain that, I'd like to hear the explanation.

upvoted 7 times

cloud_trail 3 years, 8 months ago

If I see this question on the actual exam, I'm going with AEF. The model absolutely must know where the training data is. I have seen other documentation that does confirm that you need the location of the input data, the compute instance and location to output the model artifacts.

upvoted 3 times

CloudGuru_ZA 3 years, 7 months ago

but you also need to specify the service role sagemaker should use otherwise it will not be able to perform actions on your behalf like provisioning the training instances.

upvoted 2 times

rafaelo 3 years, 6 months ago

Perfect explanation. It is CEF

upvoted 1 times

JK1977 2 years, 1 month ago

The question is asking about built in algorithms. It should be ADE. See

https://docs.aws.amazon.com/zh_tw/sagemaker/latest/dg/API_CreateTrainingJob.html

upvoted 1 times

OAmine 1 year, 9 months ago

for "3. ResourceConfig", only VolumeSizeInGB is required. So, it's not about the instance type.

Check: https://docs.aws.amazon.com/zh_tw/sagemaker/latest/APIReference/API_ResourceConfig.html

upvoted 1 times

JonSno Most Recent 4 months, 1 week ago

Selected Answer: ACF

Reason:

When submitting Amazon SageMaker training jobs using built-in algorithms, the following parameters must be specified:

Training Data Location (A)

SageMaker requires the training dataset's location in Amazon S3.

Provided as a channel input in the training job.

IAM Role (C)

SageMaker needs IAM permissions to access data from S3 and execute tasks on behalf of the user.

Model Output Path (F)

Specifies the S3 bucket location where the trained model artifacts will be stored.

upvoted 2 times

AbhayD 5 months, 1 week ago

Selected Answer: ACF

Instance type is required but not specific class CPU/GPU. Sagamkaer can handle that.

upvoted 1 times

🗨️ 👤 **MultiCloudIronMan** 8 months ago

Selected Answer: ACF

These parameters ensure that the training job has access to the necessary data, permissions, and storage locations to function correctly.

upvoted 1 times

🗨️ 👤 **MultiCloudIronMan** 8 months ago

Selected Answer: ACF

Options B, D, and E are important but not always mandatory for every training job. For example, validation data (Option B) is not always required, and hyperparameters (Option D) and instance types (Option E) can have default values or be optional depending on the specific algorithm and setup.

upvoted 1 times

🗨️ 👤 **amlgeek** 8 months, 4 weeks ago

```
import boto3
```

```
import sagemaker
```

```
sess = sagemaker.Session()
```

```
# Example for the linear learner
```

```
linear = sagemaker.estimator.Estimator(
```

```
container,
```

```
role, # role (c)
```

```
instance_count=1,
```

```
instance_type="ml.c4.xlarge", # instance type (e)
```

```
output_path=output_location, # output path (f)
```

```
sagemaker_session=sess,
```

```
)
```

upvoted 1 times

🗨️ 👤 **kiran15789** 10 months, 1 week ago

Selected Answer: CEF

Going with cef

upvoted 1 times

🗨️ 👤 **ML_2** 10 months, 2 weeks ago

Selected Answer: CEF

ANSWER IS CEF

Here from Amazon docs

InputDataConfig

An array of Channel objects. Each channel is a named input source. InputDataConfig describes the input data and its location.

Required: No

OutputDataConfig

Specifies the path to the S3 location where you want to store model artifacts. SageMaker creates subfolders for the artifacts.

Required: Yes

ResourceConfig - Identifies the resources, ML compute instances, and ML storage volumes to deploy for model training. In distributed training, you specify more than one instance.

Required: Yes

upvoted 1 times

🗨️ 👤 **RathanKalluri** 11 months, 3 weeks ago

CEF

https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_CreateTrainingJob.html#API_CreateTrainingJob_RequestParameters

upvoted 1 times

🗨️ 👤 **ninomfr64** 1 year ago

Based on https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_CreateTrainingJob.html

Required parameters are:

- AlgorithmSpecification (registry path of the Docker image with the training algorithm)

- OutputDataConfig (path to the S3 location where you want to store model artifacts)
- ResourceConfig (resources, including the ML compute instances and ML storage volumes, to use for model training)
- RoleArn
- StoppingCondition (time limit for training job)
- TrainingJobName

Thus, the answer is: C E F

wording for option E is inaccurate "EC2 instance class specifying whether training will be run using CPU or GPU" but they do it on purpose
upvoted 1 times

🗨️ **rookiee1111** 1 year, 2 months ago

Selected Answer: ACF

The input channel and output channel are mandatory, as the training job needs to know where to get the input data from and where to publish the model artifact. IAM role is also needed, for AWS services. others are not mandatory, validation channel is not mandatory for instance in case of unsupervised learning, likewise hyper params can be auto tuned for as well as the ec2 instance types can be default ones that will be picked
upvoted 2 times

🗨️ **Denise123** 1 year, 2 months ago

As they narrowed it to S3, A is incorrect BUT when submitting Amazon SageMaker training jobs using one of the built-in algorithms, it is a MUST to identify the location of training data. While Amazon S3 is commonly used for storing training data, other sources like Docker containers, DynamoDB, or local disks of training instances can also be used. Therefore, specifying the location of training data is essential for SageMaker to know where to access the data during training.

So the right answer is CEF for me for this case... However if A was saying identify the location of training data, I think option A would be included in the MUST parameter.

upvoted 1 times

🗨️ **sachin80** 1 year, 2 months ago

InputDataConfig is optional in create_training_job. Please check the parameters that are required.

So answer is CEF: https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_CreateTrainingJob.html

upvoted 1 times

🗨️ **sachin80** 1 year, 2 months ago

InputDataConfig is optional in create_training_job. Please check the parameters that are required.

So answer is SEF: https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_CreateTrainingJob.html

upvoted 1 times

🗨️ **vkajoria** 1 year, 2 months ago

Selected Answer: CEF

Input is required only when calling Fit method. When initializing the Estimator, we do not need input

upvoted 1 times

🗨️ **rav009** 1 year, 3 months ago

Selected Answer: ACF

I open the sagemaker and tested. A C F

B is not needed for non-supervised algorithm.

upvoted 2 times

A monitoring service generates 1 TB of scale metrics record data every minute. A Research team performs queries on this data using Amazon Athena. The queries run slowly due to the large volume of data, and the team requires better performance. How should the records be stored in Amazon S3 to improve query performance?

- A. CSV files
- B. Parquet files
- C. Compressed JSON
- D. RecordIO

Suggested Answer: B

Community vote distribution


B (100%)

  **gaku1016** Highly Voted 2 years, 9 months ago

Answer is B. Athena is best in Parquet format.
upvoted 24 times

  **emailtorajivk** Highly Voted 2 years, 8 months ago

You can improve the performance of your query by compressing, partitioning, or converting your data into columnar formats. Amazon Athena supports open source columnar data formats such as Apache Parquet and Apache ORC. Converting your data into a compressed, columnar format lowers your cost and improves query performance by enabling Athena to scan less data from S3 when executing your query
upvoted 14 times

  **JonSno** Most Recent 4 months, 1 week ago

Selected Answer: B

Amazon Athena performs best when querying columnar storage formats like Apache Parquet. Given that 1 TB of data is generated every minute, optimizing storage format is critical for query performance and cost efficiency.

Why Parquet (B) is the Best Choice?
Columnar Storage:

Parquet stores data by columns instead of rows, allowing Athena to scan only the needed columns, reducing the amount of data read.
Compression Efficiency:

Parquet automatically compresses data more efficiently than CSV or JSON.
Smaller file sizes = Faster queries + Lower costs.
Efficient Query Performance:

Parquet supports predicate pushdown, meaning queries can skip irrelevant rows without scanning the entire dataset.
Optimized for Big Data & Athena:

Designed for big data workloads in Athena, Redshift Spectrum, and Presto.
Works well with S3 partitioning to improve query speed.
upvoted 2 times

  **loict** 9 months, 2 weeks ago

Selected Answer: B

- A. NO - slower
 - B. YES - Parquet native in Aethena/Presto
 - C. NO - Compressed JSON
 - D. NO - no built-in support
- upvoted 2 times

  **teka112233** 10 months, 1 week ago

Selected Answer: B

according to:

<https://dzone.com/articles/how-to-be-a-hero-with-powerful-parquet-google-and>

the query run time over parquet file was 6.78 seconds while it was 236 seconds on the same data but stored on csv file which mean that parquet file is 34x faster than csv file



upvoted 1 times

  **apprehensive_scar** 2 years, 4 months ago

Selected Answer: B

B it is



upvoted 3 times

  **benson2021** 2 years, 8 months ago

Answer is B. <https://aws.amazon.com/tw/blogs/big-data/top-10-performance-tuning-tips-for-amazon-athena/>

But why does this question relate to Machine Learning?

upvoted 3 times

  **AddiWei** 2 years, 4 months ago

Because you must explore data very quickly using SQL in order to run EDA / analyze data for ML purposes. Those explorations can inform on selecting features that can be used for modeling purposes.

upvoted 5 times

Machine Learning Specialist is working with a media company to perform classification on popular articles from the company's website. The company is using random forests to classify how popular an article will be before it is published. A sample of the data being used is below.

Article_Title	Author	Top_Keywords	Day_Of_Week	URL_of_Article	Page_Views
Building a Big Data Platform	Jane Doe	Big Data, Spark, Hadoop	Tuesday	http://examplecorp.com/data_platform.html	1300456
Getting Started with Deep Learning	John Doe	Deep Learning, Machine Learning, Spark	Tuesday	http://examplecorp.com/started_deep_learning.html	1230661
MXNet ML Guide	Jane Doe	Machine Learning, MXNet, Logistic Regression	Thursday	http://examplecorp.com/mxnet_guide.html	937291
Intro to NoSQL Databases	Mary Major	NoSQL, Operations, Database	Monday	http://examplecorp.com/nosql_intro_guide.html	407812

Given the dataset, the Specialist wants to convert the Day_Of_Week column to binary values.

What technique should be used to convert this column to binary values?

- A. Binarization
- B. One-hot encoding
- C. Tokenization
- D. Normalization transformation

Suggested Answer: B

Community vote distribution

B (100%)

🗳️ **omar_bahrain** Highly Voted 2 years, 8 months ago

I choose b
upvoted 17 times

🗳️ **Juka3lj** Highly Voted 2 years, 8 months ago

Correct answer is B.
Example:
Mon | Tue | Wed
1 0 0
0 1 0
upvoted 9 times

🗳️ **kaike_reis** Most Recent 11 months, 1 week ago

Selected Answer: B
Easy Peasy
upvoted 2 times

🗳️ **earthMover** 1 year, 1 month ago

Selected Answer: B
Any categorical feature needs to be converted using One Hot Encoding and NOT label encoding.
upvoted 1 times

🗳️ **Tomatoteacher** 1 year, 5 months ago

Originally I put A, (believing to be able to format it as (0,1,2,3,4,5,6), or something as it mentioned it to convert the column, but later I realized Binarization is only designed for continuous or numerical data. Even though one-hot encoding will create 6 more columns it is correct. B is correct.
upvoted 1 times

🗳️ **Peeking** 1 year, 6 months ago

B
1000000 = Mon
0100000 = Tue
0010000 = Wed
0001000 = Thur
0000100 = Fri
0000010 = Sat
0000001 = Sun
upvoted 2 times

🗨️ 👤 **benliu1974** 1 year, 9 months ago
why not A? 001 010, 011
upvoted 2 times

🗨️ 👤 **JDKJDKJDK** 8 months, 2 weeks ago
i thought of this at first, but chatgpt's explanation changed my mind

In summary, if the names of days represent nominal categorical variables, one-hot encoding is generally the preferred choice. It maintains distinctiveness, is interpretable, and ensures that each day is clearly represented as a separate binary feature. Binary encoding may be considered for memory efficiency, especially when dealing with a large number of ordinal categories, but it should be used with caution as it introduces an ordinal relationship between categories, which may or may not align with the nature of the data. Ultimately, the choice between the two methods should align with the specific needs of your analysis and the data's characteristics.
upvoted 1 times

🗨️ 👤 **apprehensive_scar** 2 years, 4 months ago
B is the obvious answer
upvoted 2 times

🗨️ 👤 **bitsplease** 2 years, 5 months ago
Binary encoding would've been a correct answer but it is not here & Binarization is used for continuous variables. leaving w/ option B
upvoted 1 times

🗨️ 👤 **Zhubajie** 2 years, 7 months ago
B is wrong. You do not need to one hot encode the variable in random trees. If you do so, your tree must be very deep, which is not efficient. The correct answer is C!
upvoted 1 times

🗨️ 👤 **gmnk999** 2 years, 2 months ago
"The Specialist want to convert the Day Of Week column in the dataset to binary values." You are misreading the question. The answer is B.
upvoted 4 times

🗨️ 👤 **zach288** 2 years, 7 months ago
Stop misleading people, the question already asked to convert the data into binary. C is not even remotely close to be correct
upvoted 13 times

A gaming company has launched an online game where people can start playing for free, but they need to pay if they choose to use certain features. The company needs to build an automated system to predict whether or not a new user will become a paid user within 1 year. The company has gathered a labeled dataset from 1 million users.

The training dataset consists of 1,000 positive samples (from users who ended up paying within 1 year) and 999,000 negative samples (from users who did not use any paid features). Each data sample consists of 200 features including user age, device, location, and play patterns. Using this dataset for training, the Data Science team trained a random forest model that converged with over 99% accuracy on the training set. However, the prediction results on a test dataset were not satisfactory

Which of the following approaches should the Data Science team take to mitigate this issue? (Choose two.)

- A. Add more deep trees to the random forest to enable the model to learn more features.
- B. Include a copy of the samples in the test dataset in the training dataset.
- C. Generate more positive samples by duplicating the positive samples and adding a small amount of noise to the duplicated data.
- D. Change the cost function so that false negatives have a higher impact on the cost value than false positives.
- E. Change the cost function so that false positives have a higher impact on the cost value than false negatives.

Suggested Answer: CD

Community vote distribution

CD (100%)

 **Phong** Highly Voted 3 years, 9 months ago

I think it should be CD

C: because we need a balance dataset

D: The number of positive samples is large so model tends to predict 0 (negative) for all cases leading to False Negative problem. We should minimize that.

My opinion

upvoted 30 times

 **Phong** Highly Voted 3 years, 9 months ago

I think it should be CD

C: because we need a balance dataset

D: The number of negative samples is large so model tends to predict 0 (negative) for all cases leading to False Negative problem. We should minimize that.

My opinion

upvoted 24 times

 **JonSno** Most Recent 4 months, 1 week ago

Selected Answer: CD

Why These Are the Best Choices?

C. Generate more positive samples by duplicating the positive samples and adding a small amount of noise to the duplicated data.

Balances the dataset by increasing the number of positive samples.

Adding noise prevents overfitting and helps the model generalize better.

Alternative: Use SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic positive examples.

D. Change the cost function so that false negatives have a higher impact on the cost value than false positives.

Since missing a potential paying user (false negative) is more critical than misclassifying a non-paying user, adjusting the cost function to penalize false negatives more will improve recall for paid users.

Methods:

Use weighted loss functions (e.g., weighted cross-entropy).

Adjust class weights in random forest or another algorithm.

Use AUC-ROC or F1-score instead of accuracy for evaluation.

upvoted 2 times

 **dinITExam** 8 months ago

Think C and D
upvoted 1 times

🗳️ 👤 **John_Pongthorn** 3 years, 4 months ago

Selected Answer: CD

C,D is correct (percentage of the positive class is key to decide which case we are interested in)

This question, positive class (Pay) is 0.01% as compared to 99.99(not pay) , as a result, we have to pay attention to Pay because if we miss 0.01% out, we didn't get revenue. it is a false negative.

In contrast to these questions, if positive class (Pay) is 40% as compared to negative class (60% not pay), it is avoidable to emphasize on 40% (if model predict as payment but in reality customer neglect), we won't get revenue the amount from false positive)

upvoted 5 times

🗳️ 👤 **apprehensive_scar** 3 years, 4 months ago

I think is CD

upvoted 1 times

🗳️ 👤 **cloud_trail** 3 years, 7 months ago

C and D. Hopefully, no one honestly thinks that B is a good answer. Never expose test data to the training set or vice versa. C is right because of the highly imbalanced training set. D is right because you want to minimize false negatives, maximize true positives, maximize recall of the positive class. I'm not sure why anyone's worried about precision in this case.

upvoted 4 times

🗳️ 👤 **felbuch** 3 years, 7 months ago

CD

The model has 99% accuracy because it's simply predicting that everyone's a negative. Since almost everyone's a negative, it will get almost everyone right.

So we need to penalize the model for predicting that someone is a negative when it is not (i.e. penalize false negatives). So that's D.

Also, it would be really nice to have more positives -- one way to do that is to follow option C.

upvoted 8 times

🗳️ 👤 **engomaradel** 3 years, 8 months ago

CD 100%

upvoted 1 times

🗳️ 👤 **ybad** 3 years, 8 months ago

CD

C: imbalance of test (1000 positive, 999000 negative = 0.1% positive) thus C to increase that

D :also to reduce generalizing, since everyone says no, the model would generalize to no, but increasing the penalty of a false negative would reduce generalizing..

upvoted 2 times

🗳️ 👤 **Omar_Cascudo** 3 years, 8 months ago

It is needed to diminish the FP, because they are player predicted to pay and in reality will not pay. So FP should impact the cost metric more. CE should be the answer.

upvoted 2 times

🗳️ 👤 **bidds** 3 years, 8 months ago

CD are correct for sure.

upvoted 3 times

🗳️ 👤 **hans1234** 3 years, 8 months ago

It is C,E... we want to find all paying customers, which are positives, so we have to punish incorrectly finding negatives, which is E

upvoted 2 times

🗳️ 👤 **Wira** 3 years, 8 months ago

CD

although i am worried about the noise being introduced as it could skew the data nevertheless no better answer is given

upvoted 2 times

🗨️ 👤 **aws_razor** 3 years, 8 months ago

CD

We need high recall so that we do not miss many Positive cases. In that case we need to have less False Negative(FN) therefore it should have high impact on cost function.

upvoted 3 times

🗨️ 👤 **roytruong** 3 years, 8 months ago

in my view, CD are answers

C: of course, handle the imbalanced dataset

D: right now, model accuracy is 99%, it means model predict everything is negative leading to FN problem, so we need to minimize it more in cost function

upvoted 3 times

🗨️ 👤 **wuha5086** 3 years, 8 months ago

CD, FN are valuable players, we should care more on FN

upvoted 8 times

A Data Scientist is developing a machine learning model to predict future patient outcomes based on information collected about each patient and their treatment plans. The model should output a continuous value as its prediction. The data available includes labeled outcomes for a set of 4,000 patients. The study was conducted on a group of individuals over the age of 65 who have a particular disease that is known to worsen with age.

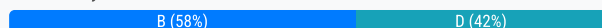
Initial models have performed poorly. While reviewing the underlying data, the Data Scientist notices that, out of 4,000 patient observations, there are 450 where the patient age has been input as 0. The other features for these observations appear normal compared to the rest of the sample population

How should the Data Scientist correct this issue?

- A. Drop all records from the dataset where age has been set to 0.
- B. Replace the age field value for records with a value of 0 with the mean or median value from the dataset
- C. Drop the age feature from the dataset and train the model using the rest of the features.
- D. Use k-means clustering to handle missing features

Suggested Answer: B

Community vote distribution



rajs Highly Voted 3 years, 9 months ago

Dropping the Age feature is a NOT ATOLL a good idea - as age plays a critical role in this disease as per the question

Dropping 10% of data is NOT a good idea considering the fact that the number of observations is already low.

The Mean or Median are a potential solutions

But the question says that "Disease worsens after age 65 so there is a correlation between age and other symptoms related feature" So that means that using Unsupervised Learning we can make pretty good prediction of "Age"

So the answer is D Use K-Means clustering

upvoted 39 times

L2007 3 years, 9 months ago

<https://www.displayr.com/5-ways-deal-missing-data-cluster-analysis/>

B is correct

upvoted 7 times

Shakespeare 6 months, 2 weeks ago

If it was KNN it would be more accurate, but we don't have that option.

upvoted 1 times

vetal Highly Voted 3 years, 9 months ago

Replacing the age with mean or median might bring a bias to the dataset. Use k-means clustering to estimate the missing age based on other features might get better results. Removing 10% available data looks odd. Why not D?

upvoted 20 times

606a82e Most Recent 3 weeks, 6 days ago

Selected Answer: B

Not D because k-means is used for clustering or grouping, not imputation.

upvoted 1 times

JonSno 4 months, 1 week ago

Selected Answer: B

The issue arises from incorrect age values (age = 0) in a dataset where all patients are supposed to be over 65 years old. Since age is an important predictor for the disease's progression, removing or ignoring this feature may negatively impact model performance.

The best approach is imputing missing or incorrect values with a reasonable estimate (e.g., mean or median age of the dataset), ensuring that:

The dataset remains intact without losing valuable patient records.

The model still benefits from age as a feature.

The imputed values are realistic and do not introduce bias.

upvoted 2 times

🗳️ 👤 **growe** 6 months ago

Selected Answer: B

Preserves data, maintains model integrity, and corrects anomalies effectively.

upvoted 1 times

🗳️ 👤 **imymoco** 12 months ago

B. Replace the age field value for records with a value of 0 with the mean or median value from the dataset: This method allows for retaining all patient records while addressing the anomaly. It is a standard approach for dealing with missing or incorrect values in a way that preserves the integrity of the dataset.

B. GPT answer

upvoted 1 times

🗳️ 👤 **pn12345** 1 year, 1 month ago

B-chatgpt

upvoted 1 times

🗳️ 👤 **rookiee1111** 1 year, 2 months ago

The question tries to mislead by adding information around the feature correlation. K-means clustering is not meant for imputing data. Hence answer should be B, that would be the right way of handling the missing value.

upvoted 1 times

🗳️ 👤 **3eb0542** 1 year, 2 months ago

Selected Answer: B

Using k-means clustering to handle missing features is not directly applicable to this scenario. K-means clustering is a method for grouping data points into clusters based on similarity, and it's not typically used for imputing missing values.

upvoted 4 times

🗳️ 👤 **kyuhuck** 1 year, 4 months ago

Selected Answer: B

add/ comment why? b ? - >replacing the age field value for records with a value of 0 with the mean or median value from the dataset, is generally the best approach among the given options. It allows the preservation of the dataset size and leverages the remaining correct data points, assuming age is a crucial predictor in this context. However, it's vital to perform this imputation carefully to avoid introducing bias. Median is often preferred in this scenario to mitigate the impact of outliers.

upvoted 3 times

🗳️ 👤 **kyuhuck** 1 year, 4 months ago

Selected Answer: B

The best way to handle the missing values in the patient age feature is to replace them with the mean or median value from the dataset. This is a common technique for imputing missing values that preserves the overall distribution of the data and avoids introducing bias or reducing the sample size.

Dropping the records or the feature would result in losing valuable information and reducing the accuracy of the model. Using k-means clustering would not be appropriate for handling missing values in a single feature, as it is a method for grouping similar data points based on multiple

upvoted 2 times

🗳️ 👤 **Topg4u** 1 year, 4 months ago

mean or median is for outliers so D

upvoted 1 times

🗳️ 👤 **endeesa** 1 year, 7 months ago

Selected Answer: B

Obviously B, why would you use a clustering algorithm to predict a value? D just doesn't make sense

upvoted 4 times

🗳️ 👤 **geoan13** 1 year, 7 months ago

B is correct.K-means is unsupervised and used mainly for clustering. KNN would have been more accurate. It can be used to predict a value. since knn is not present i think it is mean median value

upvoted 4 times

🗨️ 👤 **elvin_ml_qayiran25091992razor** 1 year, 7 months ago

Selected Answer: B

B is correct or KNN, but dont K means

upvoted 4 times

🗨️ 👤 **loict** 1 year, 9 months ago

Selected Answer: D

A. NO - unless we want to loose 10% of the data

B. NO - age is predictive, so using the means we would introduce a bias

C. NO - age is predictive

D. YES - better quality than B, it is likely that other physiological values can help predict the age

upvoted 2 times

🗨️ 👤 **FloKo** 1 year, 11 months ago

Selected Answer: D

k-means should give the best estimation of the age. Using mean would reduce the correlation between outcome and age for the model.

upvoted 1 times

A Data Science team is designing a dataset repository where it will store a large amount of training data commonly used in its machine learning models. As Data Scientists may create an arbitrary number of new datasets every day, the solution has to scale automatically and be cost-effective. Also, it must be possible to explore the data using SQL.

Which storage scheme is MOST adapted to this scenario?

- A. Store datasets as files in Amazon S3.
- B. Store datasets as files in an Amazon EBS volume attached to an Amazon EC2 instance.
- C. Store datasets as tables in a multi-node Amazon Redshift cluster.
- D. Store datasets as global tables in Amazon DynamoDB.

Suggested Answer: A

Community vote distribution

A (100%)

🗳️ **rsimham** Highly Voted 3 years, 3 months ago

Ans: A (S3) is most cost effective

upvoted 15 times

🗳️ **sonalev419** Highly Voted 3 years, 1 month ago

A : S3 cost effective + athena (not c redshift dont support unstructured data)

upvoted 7 times

🗳️ **JonSno** Most Recent 4 months, 1 week ago

Selected Answer: A

Amazon S3 (Simple Storage Service) is the best choice because it:

Scales automatically to store an arbitrary number of datasets.

Is cost-effective, as S3 charges only for storage used, unlike provisioned databases.

Supports querying datasets with SQL using Amazon Athena.

Is highly durable (99.999999999% durability) and optimized for large datasets.

How It Works in This Scenario?

Store datasets in S3 as files in Parquet, ORC, or CSV format.

Use AWS Glue Data Catalog to create table metadata.

Query the datasets using Amazon Athena (serverless SQL querying on S3).

Automatically scale without worrying about storage limits.

upvoted 1 times

🗳️ **james2033** 9 months, 3 weeks ago

Selected Answer: A

'cost effective' --> AWS S3

upvoted 1 times

🗳️ **loict** 1 year, 3 months ago

Selected Answer: A

A. YES - S3 + Athena/Presto

B. NO - no SQL support

C. NO - expensive to scale

D. NO - DynamoDB is NoSQL

upvoted 1 times

🗳️ **DavidRou** 1 year, 3 months ago

Selected Answer: A

AWS S3 + Athena will do it

upvoted 1 times

🗨️ 👤 **Ajose0** 1 year, 10 months ago

Selected Answer: A

The most appropriate storage scheme for this scenario is option A: Store datasets as files in Amazon S3.

Amazon S3 is a highly scalable and cost-effective object storage service that can store a large amount of data. S3 can scale automatically to accommodate a large number of datasets, making it a good option for storing the training data used in machine learning models. Additionally, S3 supports SQL querying through Amazon Athena or Amazon Redshift Spectrum, allowing data scientists to easily explore the data.

upvoted 2 times

🗨️ 👤 **harmanbirstudy** 3 years, 2 months ago

"store a large amount of training data commonly used in its machine learning models".. well it cannot be anything other than S3. Athena can query S3 cataloged data with SQL commands.

Answer is A

upvoted 2 times

🗨️ 👤 **Stephen_C** 3 years, 2 months ago

Amazon Redshift is not cost-effective.

upvoted 1 times

🗨️ 👤 **syu31svc** 3 years, 2 months ago

I would say C

<https://docs.aws.amazon.com/redshift/latest/mgmt/working-with-clusters.html>

"For workloads that require ever-growing storage, managed storage lets you automatically scale your data warehouse storage capacity without adding and paying for additional nodes."

upvoted 3 times

🗨️ 👤 **HaiHN** 3 years, 2 months ago

Data warehouse is not needed. For exploring data using SQL, you can use Athena

upvoted 5 times

🗨️ 👤 **kwangje** 3 years, 2 months ago

Amazon Redshift is a fast, fully managed data warehouse that makes it simple and cost-effective to analyze all your data using standard SQL and your existing Business Intelligence (BI) tools. It allows you to run complex analytic queries against petabytes of structured data using sophisticated query optimization, columnar storage on high-performance storage, and massively parallel query execution. Most results come back in seconds.

upvoted 1 times

🗨️ 👤 **roytruong** 3 years, 2 months ago

s3 is right

upvoted 1 times

🗨️ 👤 **cybe001** 3 years, 3 months ago

A, S3 is most cost effective

upvoted 3 times

A Machine Learning Specialist deployed a model that provides product recommendations on a company's website. Initially, the model was performing very well and resulted in customers buying more products on average. However, within the past few months, the Specialist has noticed that the effect of product recommendations has diminished and customers are starting to return to their original habits of spending less. The Specialist is unsure of what happened, as the model has not changed from its initial deployment over a year ago.

Which method should the Specialist try to improve model performance?

- A. The model needs to be completely re-engineered because it is unable to handle product inventory changes.
- B. The model's hyperparameters should be periodically updated to prevent drift.
- C. The model should be periodically retrained from scratch using the original data while adding a regularization term to handle product inventory changes
- D. The model should be periodically retrained using the original training data plus new data as product inventory changes.

Suggested Answer: D

Community vote distribution

D (100%)

 **rsimham** Highly Voted 3 years, 3 months ago

Ans: D

upvoted 27 times

 **JonSno** Most Recent 4 months, 1 week ago

Selected Answer: D

The model performance degradation over time suggests concept drift—the relationship between input features and the target variable has changed. Since product recommendations depend on customer behavior, preferences, and product inventory, periodic retraining with updated data ensures the model adapts to these changes.

Why Periodic Retraining?

Customer preferences evolve:

Buying patterns change over time due to seasons, trends, and external factors.

New products get added, and old ones are discontinued:

The model must learn about new items and stop recommending outdated ones.

The dataset needs to reflect recent trends:

Using new and historical data together ensures the model retains useful past knowledge while learning new patterns.

upvoted 1 times

 **james2033** 9 months, 3 weeks ago

Selected Answer: D

'retrained using the original training data plus new data'

upvoted 1 times

 **VR10** 10 months, 2 weeks ago

I believe it should be B

1. The model performance has diminished gradually over the past few months, indicating the data distribution may have changed since initial deployment over a year ago. This is a classic sign of concept drift.

2. The model architecture and training procedure have remained unchanged since initial deployment. Updating the hyperparameters is a lighter approach than retraining the model from scratch, and can help prevent further performance deterioration if done periodically to adapt to changes in user preferences and product inventory.

upvoted 1 times

 **[Removed]** 12 months ago

Selected Answer: D

Answer is D

upvoted 2 times

 **Valcilio** 1 year, 9 months ago

Selected Answer: D


D is the answer!

upvoted 1 times

  **Peeking** 2 years ago

D is the answer. There has been a data drift resulting from new customer segment visiting the site. So, the model needs to be updated periodically with new data from the website.

upvoted 3 times

  **apprehensive_scar** 2 years, 11 months ago

Selected Answer: D



DDDDD. D :D

upvoted 1 times

  **cloud_trail** 3 years, 1 month ago

Incremental training. D.


upvoted 3 times

  **gamaX** 3 years, 2 months ago

Periodically Re-Fit



D

upvoted 1 times

  **eji** 3 years, 2 months ago

agree with D

upvoted 3 times

  **C10ud9** 3 years, 2 months ago

D is correct

upvoted 3 times

A Machine Learning Specialist working for an online fashion company wants to build a data ingestion solution for the company's Amazon S3-based data lake.

The Specialist wants to create a set of ingestion mechanisms that will enable future capabilities comprised of:

- ⇒ Real-time analytics
- ⇒ Interactive analytics of historical data
- ⇒ Clickstream analytics
- ⇒ Product recommendations

Which services should the Specialist use?

- A. AWS Glue as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for real-time data insights; Amazon Kinesis Data Firehose for delivery to Amazon ES for clickstream analytics; Amazon EMR to generate personalized product recommendations
- B. Amazon Athena as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for near-real-time data insights; Amazon Kinesis Data Firehose for clickstream analytics; AWS Glue to generate personalized product recommendations
- C. AWS Glue as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for historical data insights; Amazon Kinesis Data Firehose for delivery to Amazon ES for clickstream analytics; Amazon EMR to generate personalized product recommendations
- D. Amazon Athena as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for historical data insights; Amazon DynamoDB streams for clickstream analytics; AWS Glue to generate personalized product recommendations

Suggested Answer: A

Community vote distribution

A (100%)

🗳️ 👤 **rsimham** Highly Voted 2 years, 9 months ago

Ans: A seems to be reasonable

upvoted 38 times

🗳️ 👤 **cybe001** Highly Voted 2 years, 9 months ago

A looks correct but it is missing for "Interactive analytics of historical data"

upvoted 13 times

🗳️ 👤 **ZSun** 1 year, 2 months ago

AWS Glue as data catalog, then you can analyze historical data, such as running sql with Athena.

upvoted 1 times

🗳️ 👤 **planhanasan** 2 years, 7 months ago

Once you insert real-time data to ES, you can see historical data from Kibana dashboard.

upvoted 1 times

🗳️ 👤 **ejj** 2 years, 8 months ago

but C is missing for "real-time analytics"

upvoted 1 times

🗳️ 👤 **ejj** 2 years, 8 months ago

and also C is saying historical data analytics for Kinesis Data analytics which is real-time analytics not historical, so the answer might not C but the answer is A

upvoted 1 times

🗳️ 👤 **loict** Most Recent 9 months, 2 weeks ago

Selected Answer: A

A. YES - Amazon Kinesis Data Analytics is for real-time data insights

B. NO - Amazon Athena has no data catalog

C. NO - Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics is not for historical data insights

D. NO - Amazon Athena has no data catalog

upvoted 3 times

🗳️ 👤 **kaike_reis** 11 months ago

Selected Answer: A

Athena can not be used for data catalog, so B and D are wrong. A and C are equals, but it's well known that Kinesis DS and Analytics are used together for real time solutions, which is mentioned in the question / answer, but lack on C.

upvoted 2 times

🗳️ 👤 **Valcilio** 1 year, 3 months ago

Selected Answer: A

All are bad options, but A can do it.

upvoted 2 times

🗳️ 👤 **hug_c0sm0s** 1 year, 4 months ago

Selected Answer: A

AWS Glue is a fully managed extract, transform, and load (ETL) service that makes it easy to move data between data stores. It can be used as a data catalog to store metadata information about the data in the data lake. Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics can be used together to collect, process, and analyze real-time streaming data. Amazon Kinesis Data Firehose can be used to deliver streaming data to destinations such as Amazon ES for clickstream analytics. Finally, Amazon EMR can be used to run big data frameworks such as Apache Spark and Apache Hadoop to generate personalized product recommendations.

upvoted 2 times

🗳️ 👤 **gamaX** 2 years, 8 months ago

A or C

<https://aws.amazon.com/blogs/big-data/retaining-data-streams-up-to-one-year-with-amazon-kinesis-data-streams/>

upvoted 2 times

🗳️ 👤 **harmanbirstudy** 2 years, 8 months ago

Athena can do Interactive analytics on Historical data, but here its only use is "Athena as the data catalog" and this is the work of Glue data catalog using its crawlers, so it cannot be B or D.

--So its either A or C

-- Now Kinesis data Streams/Analytics is know for real time data analytics but if it is reading from data already stored in S3 using DMS then we can say it is getting historical data.

-- Here I am not very clear if Kinesis part will happen on incoming data before S3 or After data persists to S3 and Kinesis reads it through S3-->DMS-- Kinesis data stream -- Kinesis analytics-->Firehose.

But still insights are always on real-time/current data based on historical data trends , so the statement in C "Analytics for historical data insights" is in-correct in general .

Hence ANSWER is :A

upvoted 5 times

🗳️ 👤 **ybad** 2 years, 8 months ago

A is correct,

for those asking the difference between A and D, D talks about using kinesis stream and data analytics to create historical analysis.... waste of money no?

upvoted 2 times

🗳️ 👤 **Th3Dud3** 2 years, 8 months ago

Answer = A

upvoted 4 times

🗳️ 👤 **C10ud9** 2 years, 8 months ago

A it is

upvoted 2 times

🗳️ 👤 **roytruong** 2 years, 8 months ago

it's A, ES can perform clickstream analytics and EMR can handle spark job recommendation at scale

upvoted 3 times

🗳️ 👤 **BigPlums** 2 years, 8 months ago

Only C and D mention interactive analytics of historical data.

Glue won't provide personalised recommendation so it is C

upvoted 1 times


🗳️ 👤 **BigEv** 2 years, 8 months ago

What is the difference between the solution in A or C ????

upvoted 2 times

🗳️ 👤 **JayK** 2 years, 8 months ago

A is real time data analytics with Kinesis Data analytics and C is saying historical data which is wrong
upvoted 6 times

  **ComPah** 2 years, 9 months ago

Looks like C Amazon ES has Kibana which supports click stream
upvoted 2 times

  **ComPah** 2 years, 9 months ago

A is Correct
upvoted 4 times

A company is observing low accuracy while training on the default built-in image classification algorithm in Amazon SageMaker. The Data Science team wants to use an Inception neural network architecture instead of a ResNet architecture.

Which of the following will accomplish this? (Choose two.)

- A. Customize the built-in image classification algorithm to use Inception and use this for model training.
- B. Create a support case with the SageMaker team to change the default image classification algorithm to Inception.
- C. Bundle a Docker container with TensorFlow Estimator loaded with an Inception network and use this for model training.
- D. Use custom code in Amazon SageMaker with TensorFlow Estimator to load the model with an Inception network, and use this for model training.
- E. Download and apt-get install the inception network code into an Amazon EC2 instance and use this instance as a Jupyter notebook in Amazon SageMaker.

Suggested Answer: CD

Community vote distribution

CD (100%)

🗳️ **DonaldCMLIN** Highly Voted 2 years, 3 months ago

You might be spent a lot of money for ask AWS A.CHANGE built-in image OR B.Create a support case.

The effectual way BOTH RELATIVE TO SageMaker Estimator

C.DOCKER

OR BRING YOUR CODE BY

D.SageMaker with TensorFlow Estimator

THE BEAUTYFUL ANSWER ARE C AND D

upvoted 33 times

🗳️ **Phong** Highly Voted 2 years, 2 months ago

I will go for C & D

upvoted 10 times

🗳️ **hug_c0sm0s** Most Recent 10 months ago

Selected Answer: CD

Option A is not possible because the built-in image classification algorithm cannot be customized. Option B is not feasible because it is not possible to change the default image classification algorithm through a support case. Option E is also not a recommended approach because it involves manually installing software on an EC2 instance rather than using the managed services provided by SageMaker.

upvoted 4 times

🗳️ **sqavi** 10 months, 3 weeks ago

Selected Answer: CD

The effectual way BOTH RELATIVE TO SageMaker Estimator

C.DOCKER

OR BRING YOUR CODE BY

D.SageMaker with TensorFlow Estimator

upvoted 2 times

🗳️ **Huy** 2 years, 2 months ago

This question ask for 2 ways not a set of actions. So may be confused.

upvoted 1 times

🗳️ **ahquiceno** 2 years, 2 months ago

Answers AD go to: <https://docs.aws.amazon.com/sagemaker/latest/dg/docker-containers.html>

upvoted 2 times

🗳️ **ybad** 2 years, 2 months ago

CD and also A says it but in a more general term....

upvoted 1 times

🗨️ 👤 **jaydec** 2 years, 2 months ago

<https://aws.amazon.com/blogs/machine-learning/transfer-learning-for-custom-labels-using-a-tensorflow-container-and-bring-your-own-algorithm-in-amazon-sagemaker/>

upvoted 3 times

🗨️ 👤 **Antriksh** 2 years, 2 months ago

C and D are correct

upvoted 5 times

🗨️ 👤 **DonaldCMLIN** 2 years, 3 months ago

https://docs.aws.amazon.com/zh_tw/sagemaker/latest/dg/your-algorithms.html

<https://aws.amazon.com/tw/blogs/machine-learning/transfer-learning-for-custom-labels-using-a-tensorflow-container-and-bring-your-own-algorithm-in-amazon-sagemaker/>

https://docs.aws.amazon.com/zh_tw/sagemaker/latest/dg/tf.html

upvoted 1 times

A Machine Learning Specialist built an image classification deep learning model. However, the Specialist ran into an overfitting problem in which the training and testing accuracies were 99% and 75%, respectively.
How should the Specialist address this issue and what is the reason behind it?

- A. The learning rate should be increased because the optimization process was trapped at a local minimum.
- B. The dropout rate at the flatten layer should be increased because the model is not generalized enough.
- C. The dimensionality of dense layer next to the flatten layer should be increased because the model is not complex enough.
- D. The epoch number should be increased because the optimization process was terminated before it reached the global minimum.


Suggested Answer: D

Reference:

https://www.tensorflow.org/tutorials/keras/overfit_and_underfit

Community vote distribution

B (100%)

 **DonaldCMLIN**  3 years, 3 months ago

DROPOUT HELPS PREVENT OVERFITTING

<https://keras.io/layers/core/#dropout>

THE BEAUTIFUL ANSER SHOULD BE B.

upvoted 55 times

 **rsimham** 3 years, 3 months ago

agree. it should be B

upvoted 10 times

 **syu31svc**  3 years, 2 months ago

<https://kharshit.github.io/blog/2018/05/04/dropout-prevent-overfitting>

Answer is B 100%

upvoted 5 times

 **fm99**  8 months, 3 weeks ago

Selected Answer: B

Increasing dropout rate will reduce complexity of the model which inturn reduces overfitting

upvoted 1 times

 **VR10** 10 months, 2 weeks ago

This is clearly B, dont get why the answer is marked as D.

upvoted 1 times

 **endeesa** 1 year, 1 month ago

Selected Answer: B

Regularization will seek to obtain similar accuracies in train and test sets. Anything else will make the overfitting worse

upvoted 1 times

 **elvin_ml_qayiran25091992razor** 1 year, 1 month ago

Selected Answer: B

B is correct, D so stup*d answer

upvoted 1 times

 **loict** 1 year, 3 months ago

Selected Answer: B

A. NO - accuracy on training set is high

B. YES - increased dropout rate => reduce model complexity => less overfitting

C. NO - we want to reduce model complexity

D. NO - the model converged

upvoted 2 times

DavidRou 1 year, 3 months ago

Selected Answer: B

I don't understand why the highlighted "right" answer is D. To increase the number of epochs will make the situation even worse than it is; dropout is the right action to take in this case

upvoted 2 times

kaike_reis 1 year, 5 months ago

Selected Answer: B

B is correct

upvoted 1 times

nilmans 1 year, 6 months ago

agree, B makes more sense here

upvoted 1 times

soonmo 1 year, 6 months ago

Selected Answer: B

Definitely B because overfitting comes from complex model that captures patterns of training data well. But D is getting this model more complex, worsening overfitting.

upvoted 1 times

soonmo 1 year, 6 months ago

Correct my reasoning! D is worsening overfitting because it feeds more data after overfitting arises. D is used for underfitted models.

upvoted 1 times

earthMover 1 year, 7 months ago

Selected Answer: B

Increasing Epoch only makes things worse on a overfitting model. You should perform regularization by introducing drop outs to generalize the model.

upvoted 1 times

user009 1 year, 9 months ago

Option B is the correct answer because increasing the dropout rate at the flatten layer helps prevent overfitting by randomly dropping out units during training, effectively creating a more robust model that can generalize better to new data. Dropout is a regularization technique that helps prevent overfitting by forcing the model to learn redundant representations of the data. By increasing the dropout rate at the flatten layer, the model becomes more generalized, which should help to improve the testing accuracy.

upvoted 1 times

AjoseO 1 year, 10 months ago

Selected Answer: B

Overfitting occurs when a model is too complex and memorizes the training data instead of learning the underlying pattern. As a result, the model performs well on the training data but poorly on new, unseen data.

Increasing the dropout rate, a regularization technique, can help combat overfitting by randomly dropping out some neurons during training, which prevents the model from relying too heavily on any single feature.

upvoted 1 times

sqavi 1 year, 10 months ago

Selected Answer: B

Model is overfitting, I will go with option B, increasing epoch will cause more overfitting

upvoted 2 times

desperatestudent 1 year, 11 months ago

Selected Answer: B

it should answer B.

upvoted 1 times

Shailendraa 2 years, 3 months ago

12-sep exam

upvoted 2 times

A Machine Learning team uses Amazon SageMaker to train an Apache MXNet handwritten digit classifier model using a research dataset. The team wants to receive a notification when the model is overfitting. Auditors want to view the Amazon SageMaker log activity report to ensure there are no unauthorized API calls.

What should the Machine Learning team do to address the requirements with the least amount of code and fewest steps?

- A. Implement an AWS Lambda function to log Amazon SageMaker API calls to Amazon S3. Add code to push a custom metric to Amazon CloudWatch. Create an alarm in CloudWatch with Amazon SNS to receive a notification when the model is overfitting.
- B. Use AWS CloudTrail to log Amazon SageMaker API calls to Amazon S3. Add code to push a custom metric to Amazon CloudWatch. Create an alarm in CloudWatch with Amazon SNS to receive a notification when the model is overfitting.
- C. Implement an AWS Lambda function to log Amazon SageMaker API calls to AWS CloudTrail. Add code to push a custom metric to Amazon CloudWatch. Create an alarm in CloudWatch with Amazon SNS to receive a notification when the model is overfitting.
- D. Use AWS CloudTrail to log Amazon SageMaker API calls to Amazon S3. Set up Amazon SNS to receive a notification when the model is overfitting.

Suggested Answer: B

Community vote distribution

B (80%)

D (20%)

 **DonaldCMLIN** Highly Voted 3 years, 9 months ago

THE ANSWER SHOULD BE B.

YOU DON'T NEED TO THROUGH LAMBDA TO INTERGE CLOUDTRAIL

Log Amazon SageMaker API Calls with AWS CloudTrail

<https://docs.aws.amazon.com/sagemaker/latest/dg/logging-using-cloudtrail.html>

upvoted 41 times

 **rajs** Highly Voted 3 years, 9 months ago

Agreed B for the following reasons

CloudTrail logs captured in S3 without any code/lambda

The custom metrics can be published to Cloudwatch...in this case it would be a test for overfit on MXNET which will set off an alarm which can then be subscribed on SNS

upvoted 11 times

 **JonSno** Most Recent 4 months, 1 week ago

Selected Answer: B

Breakdown of the Chosen Solution (B)

Use AWS CloudTrail to log SageMaker API calls to Amazon S3 ✓

CloudTrail automatically logs all AWS API activity, including SageMaker API calls, for auditing.

S3 stores these logs securely for auditor review.

Push custom metrics to Amazon CloudWatch ✓

Model overfitting can be detected using a custom CloudWatch metric (e.g., validation loss increasing while training loss decreases).

The SageMaker training script can push loss values to CloudWatch during training.

Create a CloudWatch alarm + SNS notification ✓

Set a CloudWatch alarm on the overfitting metric (e.g., validation loss surpassing a threshold).

Use Amazon SNS to send a notification (email, SMS, or Lambda trigger) when the alarm is triggered.

upvoted 1 times

 **MultiCloudIronMan** 8 months ago

Selected Answer: B

Option D involves using AWS CloudTrail to log Amazon SageMaker API calls to Amazon S3 and setting up Amazon SNS to receive a notification when the model is overfitting. While this approach addresses the logging requirement, it does not provide a mechanism for pushing custom metrics to

Amazon CloudWatch, which is necessary for monitoring model performance and detecting overfitting. So 'B' is correct
upvoted 2 times

🗨️ 👤 **Chiquitabandita** 1 year, 2 months ago

Selected Answer: D

https://docs.aws.amazon.com/fr_fr/sagemaker/latest/dg/training-metrics.html#define-train-metrics

It detects hardware resource usage issues (such as CPU, GPU, and I/O bottlenecks) and non-convergent model issues (such as overfitting, disappearing gradients, and tensor explosion).

why couldn't the answer be D, as this covers all of the requirements, and B seems to add an extra step with adding push code, when it already has a builtin metric for overfitting.

upvoted 1 times

🗨️ 👤 **Mobasher** 4 months, 3 weeks ago

This would have been correct had the question not mentioned that the algorithm is "hand-written" which means it's not a built in algorithm. So, for SageMaker AI to understand your custom algorithm's metrics, it needs a regex definition to apply to the logs in order to generate those custom metrics and then alert on them using CW Alarms and SNS to deliver notifications. See <https://docs.aws.amazon.com/sagemaker/latest/dg/define-train-metrics.html>

upvoted 1 times

🗨️ 👤 **Aja1** 1 year, 2 months ago

Custom metric Need to built and pushed.

upvoted 1 times

🗨️ 👤 **loict** 1 year, 9 months ago

Selected Answer: B

- A. NO - CloudTrail has built-in SageMaker API calls tracking, no lambda needed
- B. YES - the chain works
- C. NO - CloudTrail has built-in SageMaker API calls tracking, no lambda needed
- D. NO - CloudTrail has not specific Amazon SageMaker integration to detect overfitting

upvoted 1 times

🗨️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: B

Option B

upvoted 1 times

🗨️ 👤 **ADVIT** 2 years ago

"least amount of code and fewest steps?"

I think it's D.

upvoted 2 times

🗨️ 👤 **kukreti18** 1 year, 11 months ago

Agreed, with less code effort.

upvoted 1 times

🗨️ 👤 **Paolo991** 2 years, 3 months ago

I would consider D as well.

You can just setup a SNS that is triggered by a built-in action like here:

<https://docs.aws.amazon.com/sagemaker/latest/dg/debugger-built-in-actions.html>

You can see that overfitting is a built-in rule for MXNet from here:

<https://docs.aws.amazon.com/sagemaker/latest/dg/debugger-built-in-rules.html>

Not that B is not working. Maybe the question was prior to this new solution.

upvoted 2 times

🗨️ 👤 **khchan123** 1 year, 7 months ago

The `loss_not_decreasing`, `overfit`, `overtraining`, and `stalled_training_rule` monitors if your model is optimizing the loss function without those training issues. If the rules detect training anomalies, the rule evaluation status changes to `IssueFound`. You can set up automated actions, such as notifying training issues and stopping training jobs using Amazon CloudWatch Events and AWS Lambda. For more information, see [Action on](#)

Amazon SageMaker Debugger Rules.

<https://docs.aws.amazon.com/sagemaker/latest/dg/use-debugger-built-in-rules.html>

upvoted 1 times

🗳️ 👤 **Valcilio** 2 years, 3 months ago

Selected Answer: B

It's B.

upvoted 1 times

🗳️ 👤 **Ajose0** 2 years, 3 months ago

Selected Answer: B

AWS CloudTrail provides a history of AWS API calls made on the account. The Machine Learning team can use AWS CloudTrail to log Amazon SageMaker API calls to Amazon S3. They can then use CloudWatch to create alarms and receive notifications when the model is overfitting.

To ensure auditors can view the Amazon SageMaker log activity report, the team can add code to push a custom metric to Amazon CloudWatch. This provides a single place to view and analyze logs across all the services and resources in the environment.

upvoted 1 times

🗳️ 👤 **sonalev419** 3 years, 8 months ago

B. cloudwatch + metrics from sagemaker + sns https://docs.aws.amazon.com/fr_fr/sagemaker/latest/dg/training-metrics.html#define-train-metrics

upvoted 4 times

🗳️ 👤 **ybad** 3 years, 8 months ago

B requires the least amount of code and satisfies all conditions

upvoted 2 times

🗳️ 👤 **tochiebby** 3 years, 8 months ago

What does this line do?

"Add code to push a custom metric to Amazon CloudWatch"

upvoted 1 times

🗳️ 👤 **Omar_Cascudo** 3 years, 8 months ago

It creates a metric for overfitting (accuracy of training data and accuracy of test data).

upvoted 5 times

🗳️ 👤 **jonclem** 3 years, 8 months ago

Its not B. Why would you use CloudTrail?

Having used Lambda for API calls I'm inclined to agree with the original answer, C.

upvoted 1 times

🗳️ 👤 **Pja1** 3 years, 8 months ago

<https://docs.aws.amazon.com/sagemaker/latest/dg/logging-using-cloudtrail.html>

upvoted 3 times

🗳️ 👤 **fhuadeen** 3 years, 8 months ago

Because that is the only job of CloudTrail - to log actions taken on your AWS account. So why need a Lambda function to trigger it?

upvoted 3 times

🗳️ 👤 **Antriksh** 3 years, 8 months ago

B it is

upvoted 2 times

🗳️ 👤 **C10ud9** 3 years, 8 months ago

B it is

upvoted 1 times

A Machine Learning Specialist is building a prediction model for a large number of features using linear models, such as linear regression and logistic regression.

During exploratory data analysis, the Specialist observes that many features are highly correlated with each other. This may make the model unstable.

What should be done to reduce the impact of having such a large number of features?

- A. Perform one-hot encoding on highly correlated features.
- B. Use matrix multiplication on highly correlated features.
- C. Create a new feature space using principal component analysis (PCA)
- D. Apply the Pearson correlation coefficient.

Suggested Answer: C

Community vote distribution

C (100%)

🗳️ **cybe001** Highly Voted 2 years, 9 months ago

C is correct

upvoted 19 times

🗳️ **syu31svc** Highly Voted 2 years, 8 months ago

You want to reduce features/dimension so PCA is the answer

upvoted 5 times

🗳️ **kaike_reis** Most Recent 11 months ago

Selected Answer: C

C is the way

upvoted 3 times

🗳️ **FloKo** 11 months ago

Selected Answer: C

C is correct.

D could be correct if the correlation is used to omit features.

upvoted 1 times

🗳️ **Valcilio** 1 year, 3 months ago

Selected Answer: C

PCA and T-SNE are for solving the curse of dimensionality mentioned here!

upvoted 1 times

🗳️ **DS2021** 1 year, 5 months ago

I assume PCA is for unsupervised learning!...and the scenario in the question looks like supervised learning

upvoted 1 times

🗳️ **GiyeonShin** 1 year, 4 months ago

data (x, y) --> (PCA) --> preprocessed data(x', y) --> learning

why not for supervised learning?

upvoted 1 times

🗳️ **BethChen** 1 year, 6 months ago

Selected Answer: C

Tricky. The sentence 'many features are highly correlated with each other' is no use.

upvoted 1 times

🗳️ **kaike_reis** 11 months ago

It's. PCA removes such correlation.

upvoted 1 times

🗳️ **Shailendraa** 1 year, 9 months ago



Answer C: Read through this carefully "What should be done to reduce the impact of having such a large number of features?" only answer comes in mind PCA

upvoted 1 times

  **Urban_Life** 2 years, 8 months ago

Of course, it's PCA.

upvoted 1 times

  **C10ud9** 2 years, 9 months ago

PCA is the solution. So, answer is C

upvoted 2 times

A Machine Learning Specialist is implementing a full Bayesian network on a dataset that describes public transit in New York City. One of the random variables is discrete, and represents the number of minutes New Yorkers wait for a bus given that the buses cycle every 10 minutes, with a mean of 3 minutes.

Which prior probability distribution should the ML Specialist use for this variable?

- A. Poisson distribution
- B. Uniform distribution
- C. Normal distribution
- D. Binomial distribution

Suggested Answer: D

Community vote distribution

A (100%)

 **ComPah**  3 years, 9 months ago

A

If you have information about the average (mean) number of things that happen in some given time period / interval, Poisson distribution can give you a way to predict the odds of getting some other value on a given future day

upvoted 57 times

 **swagy**  3 years, 9 months ago

Ans: A

<https://brilliant.org/wiki/poisson-distribution/>

upvoted 8 times

 **Mobasher**  4 months, 3 weeks ago

Selected Answer: B

ChatGPT's answer: B

Explanation

The problem describes a random variable representing the waiting time for a bus, where buses arrive every 10 minutes, and the mean waiting time is 3 minutes.

In such a periodic arrival process, the waiting time follows a Uniform Distribution because:

- Any given person's waiting time is equally likely to be any value between 0 and 10 minutes.
- There is no clustering around a particular value—every moment within the cycle is equally probable.

Thus, the waiting time follows a Uniform(0, 10) distribution.

upvoted 1 times

 **Mobasher** 4 months, 3 weeks ago

Why Not the Other Options?

(A) Poisson Distribution

Poisson is used for counting discrete events over a fixed period (e.g., number of buses arriving per hour). Since waiting time is continuous, Poisson is not appropriate.

(C) Normal Distribution

Normal (Gaussian) distribution assumes values cluster around the mean and extend infinitely. Here, waiting time is evenly spread between 0–10 minutes, not forming a bell curve.

(D) Binomial Distribution

Binomial is used for counting successes in a fixed number of trials (e.g., flipping a coin multiple times). Waiting time is continuous, not a count of discrete occurrences.

upvoted 1 times

 **growe** 6 months ago

Selected Answer: B

Buses cycle every 10 minutes, and waiting time can be modeled as a uniform random variable between [0, 10] minutes.

The average waiting time of 3 minutes suggests that waiting is uniformly distributed, not event-based like Poisson.



If buses arrive every 10 minutes and riders arrive randomly, the waiting time follows a Uniform Distribution (B) because:

The arrival process is regular (every 10 minutes).

There's no stochastic randomness in the bus arrival schedule, ruling out Poisson.

Poisson would apply if buses arrived randomly at an average rate rather than at fixed intervals.

upvoted 2 times

  **87ebc7d** 7 months, 1 week ago

B

Poisson is suitable for modeling the number of events (like buses arriving) in a fixed time frame, not the time between events when the events occur at regular intervals. The waiting time variable is not about the count of buses but rather the time to the next bus, which is evenly distributed.



upvoted 1 times

  **elvin_ml_qayiran25091992razor** 1 year, 7 months ago

Selected Answer: A

A is correct

upvoted 1 times

  **Fred93** 1 year, 9 months ago

Selected Answer: A

Poisson distribution is discrete, and gives the number of events that occur in a given time interval

upvoted 2 times

  **loict** 1 year, 9 months ago

Selected Answer: A

A. YES - Poisson distribution is discrete, and gives the number of events that occur in a given time interval

B. NO - Uniform distribution is continuous, we want discrete

C. NO - Normal distribution is continuous we want discrete

D. NO - Binomial distribution give the probability that a random variable is A or B (possibly in with different weight)

upvoted 2 times

  **Mickey321** 1 year, 10 months ago

Selected Answer: A

Option A indeed

upvoted 1 times



  **Nadia0012** 2 years, 3 months ago

Selected Answer: A

ANSWER IS A

<https://www.investopedia.com/terms/d/discrete-distribution.asp>



upvoted 2 times

  **bakarys** 2 years, 4 months ago

Selected Answer: A

The Poisson distribution is commonly used for count data, which is the case here as we are interested in the number of minutes New Yorkers wait for a bus. The Poisson distribution is characterized by a single parameter, lambda, which represents the mean and variance of the distribution. In this case, the mean is 3 minutes, so we would set lambda to 3. The Poisson distribution assumes that events occur independently of each other, which is a reasonable assumption in this case since the waiting time for each individual is likely to be independent of the waiting time for others.

upvoted 4 times

  **Ajose0** 2 years, 4 months ago

Selected Answer: A

The Poisson distribution is a discrete probability distribution that is commonly used to model the number of events that occur in a fixed interval of time, given an average rate of occurrence.



Since the buses cycle every 10 minutes and the mean wait time is 3 minutes, it is reasonable to assume that the number of minutes New Yorkers wait for a bus can be modeled by a Poisson distribution.

upvoted 3 times

  **Tomatoteacher** 2 years, 5 months ago



Selected Answer: A

100% A, as discrete, while binomial has to be binary data (success or failure)
upvoted 1 times

  **Sonoko** 2 years, 6 months ago

Selected Answer: A


A is a discrete distribution
upvoted 1 times

  **Peeking** 2 years, 6 months ago

I do choose Poisson. A.
upvoted 1 times

  **Shailendraa** 2 years, 9 months ago

12-sep exam
upvoted 3 times

  **Shailendraa** 2 years, 9 months ago

Answer is A .. these types on footfalls ,etc ..answer always Poisson-distribution
upvoted 1 times

A Data Science team within a large company uses Amazon SageMaker notebooks to access data stored in Amazon S3 buckets. The IT Security team is concerned that internet-enabled notebook instances create a security vulnerability where malicious code running on the instances could compromise data privacy.

The company mandates that all instances stay within a secured VPC with no internet access, and data communication traffic must stay within the AWS network.

How should the Data Science team configure the notebook instance placement to meet these requirements?

- A. Associate the Amazon SageMaker notebook with a private subnet in a VPC. Place the Amazon SageMaker endpoint and S3 buckets within the same VPC.
- B. Associate the Amazon SageMaker notebook with a private subnet in a VPC. Use IAM policies to grant access to Amazon S3 and Amazon SageMaker.
- C. Associate the Amazon SageMaker notebook with a private subnet in a VPC. Ensure the VPC has S3 VPC endpoints and Amazon SageMaker VPC endpoints attached to it.
- D. Associate the Amazon SageMaker notebook with a private subnet in a VPC. Ensure the VPC has a NAT gateway and an associated security group allowing only outbound connections to Amazon S3 and Amazon SageMaker.

Suggested Answer: C

Community vote distribution

C (100%)

🗳️ **DonaldCMLIN** Highly Voted 2 years, 9 months ago

NAT gateway COULD GO OUT TO THE INTERNET AND DOWNLOAD BACK MALICIOUS
D. IS NOT A GOOD ANSWER.

THE SAFE ONE IS ANSWER C. ASSOCIATE WITH VPC_ENDPOINT AND S3_ENDPOINT
upvoted 35 times

🗳️ **BigEv** Highly Voted 2 years, 9 months ago

C is correct

We must use the VPC endpoint (either Gateway Endpoint or Interface Endpoint) to comply with this requirement "Data communication traffic must stay within the AWS network".

<https://docs.aws.amazon.com/sagemaker/latest/dg/notebook-interface-endpoint.html>
upvoted 23 times

🗳️ **loict** Most Recent 9 months, 2 weeks ago

Selected Answer: C

- A. NO - We don't place a S3 bucket in a VPC, it is always in AWS Service Account
 - B. NO - without an S3 VPC endpoint, traffic will go through the Internet
 - C. YES - we need endpoints for both SageMaker and S3 to avoid Internet traffic
 - D. NO - we need endpoints for both SageMaker and S3 to avoid Internet traffic
- upvoted 2 times

🗳️ **Mickey321** 10 months ago

Selected Answer: C

Option C

upvoted 1 times

🗳️ **kaike_reis** 11 months ago

Selected Answer: C

C is the correct. A is not so correct, because it's possible to communicate two different VPCs inside AWS network (which is not optimized).
upvoted 1 times

🗳️ **Ajose0** 1 year, 4 months ago

Selected Answer: C

This configuration would meet the company's requirements for security, as the notebook instance would be placed within a private subnet in a VPC, and data communication traffic would stay within the AWS network through the use of VPC endpoints for S3 and Amazon SageMaker.

Additionally, the VPC would not have internet access, further reducing the security risk.

upvoted 2 times

🗳️ 👤 **rb39** 1 year, 9 months ago

C - "and data communication traffic must stay within the AWS network." that discards D

upvoted 2 times

🗳️ 👤 **StelSen** 2 years, 8 months ago

Answer should be C. Because, Security team don't want Internet Access, Option-D has NAT and will get to Internet somehow. Also connecting S3 and SageMaker EC2 instance via VPC endpoints is best way to secure the resources.

upvoted 4 times

🗳️ 👤 **cloud_trail** 2 years, 8 months ago

Using a NAT gateway is the old way to do it. Option C is the way to do it now. <https://cloudacademy.com/blog/vpc-endpoint-for-amazon-s3/#:~:text=Accessing%20S3%20the%20old%20way%20%28without%20VPC%20Endpoint%29,has%20no%20access%20to%20any%20outside%20public%20re>

upvoted 2 times

🗳️ 👤 **harmanbirstudy** 2 years, 8 months ago

"and data communication traffic must stay within the AWS network", NAT gateway will always go over the Internet to access S3. with NAT you can put your instances in private subnet and NAT itself in public subnet, but still in order to access S3 it will go over the internet. SO answer cannot be D.

-- C is the only correct option here, as S3 VPC endpoints is a real thing "google it" and its sole purpose is to create route from VPC endpoint to S3, without going over the Internet.

upvoted 3 times

🗳️ 👤 **scuzzy2010** 2 years, 8 months ago

C is correct answer. D is only applicable - "If your model needs access to an AWS service that doesn't support interface VPC endpoints or to a resource outside of AWS, create a NAT gateway and configure your security groups to allow outbound connections."

<https://docs.aws.amazon.com/sagemaker/latest/dg/host-vpc.html>

upvoted 3 times

🗳️ 👤 **v24143** 2 years, 8 months ago

D is correct

upvoted 1 times

🗳️ 👤 **krakow1234** 2 years, 8 months ago

Answer is D, read third paragraph <https://docs.aws.amazon.com/sagemaker/latest/dg/appendix-notebook-and-internet-access.html>

upvoted 1 times

🗳️ 👤 **Potato_Noodle** 2 years, 8 months ago

NAT is the way that a VPC connects to internet and other AWS service when there is NO INTERNET ACCESS FOR VPC. Thus the answer is D.

upvoted 1 times

🗳️ 👤 **Th3Dud3** 2 years, 8 months ago

"concerned that internet-enabled notebook instances create a security vulnerability where malicious code running on the instances could compromise data privacy." NAT Gateway does not mitigate this risk!

upvoted 2 times

🗳️ 👤 **yeetusdeleetus** 2 years, 8 months ago

This is the correct answer.

If this answer is confusing, study some of the associate exams before going for this one. VPC endpoint and NAT gateway are similar, but NAT gateway is for giving resources in the VPC the chance to initiate connections with the internet, whereas a VPC endpoint only allows it to go to other AWS services, which is the best solution for this question.

upvoted 2 times

🗳️ 👤 **Th3Dud3** 2 years, 8 months ago



C:

If you configure your VPC so that it doesn't have internet access, models that use that VPC do not have access to resources outside your VPC. If your model needs access to resources outside your VPC, provide access with one of the following options:

If your model needs access to an AWS service that supports interface VPC endpoints, create an endpoint to connect to that service. For a list of services that support interface endpoints, see VPC Endpoints in the Amazon VPC User Guide. For information about creating an interface VPC endpoint, see Interface VPC Endpoints (AWS PrivateLink) in the Amazon VPC User Guide.



If your model needs access to an AWS service that doesn't support interface VPC endpoints or to a resource outside of AWS, create a NAT gateway and configure your security groups to allow outbound connections. For information about setting up a NAT gateway for your VPC, see Scenario 2: VPC with Public and Private Subnets (NAT) in the Amazon Virtual Private Cloud User Guide.

upvoted 5 times

  **sebtac** 2 years, 8 months ago

what is the difference between A & C? are both answers OK?

upvoted 1 times

  **jrff** 1 year, 8 months ago

It is not enough for sagemaker to communicate to S3 if both of them are inside the same VPC. Sagemaker inside a VPC needs to create a endpoint to connect to other AWS services which has endpoint too.

upvoted 1 times

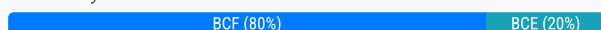
A Machine Learning Specialist has created a deep learning neural network model that performs well on the training data but performs poorly on the test data.

Which of the following methods should the Specialist consider using to correct this? (Choose three.)

- A. Decrease regularization.
- B. Increase regularization.
- C. Increase dropout.
- D. Decrease dropout.
- E. Increase feature combinations.
- F. Decrease feature combinations.

Suggested Answer: BCF

Community vote distribution



cybe001 Highly Voted 3 years, 9 months ago

Yes, answer is BCF

upvoted 24 times

Phong Highly Voted 3 years, 8 months ago

Go for BCF

upvoted 14 times

ninomfr64 Most Recent 1 year ago

Selected Answer: BCF

I think here the point is around the definition of "feature combinations".

If you refer to it as "combine the features to generate a smaller but more effective feature set" this would end up to a smaller feature set thus a good thing for overfitting.

However, if you refer to it as "combine the features to generate additional features" this would end up to a larger feature set thus a bad thing for overfitting.

Also, in some cases you implement feature combinations in your model (see hidden layers in feed-forward network) thus increasing model complexity which is bad for overfitting.

To me this question is poorly worded. I would pick F as my best guess is that you need to implement feature combination in your model, thus decreasing feature combination decrease complexity hence improving with overfitting issue

upvoted 5 times

cloudera3 11 months, 3 weeks ago

Great callout - what exactly the Feature combination is performing has not been elaborated

It can be: Using PCA or t-SNE, it is essentially optimizing features - good to address overfitting, and should be done

Or, it can be: Using Cartesian Product, features are being combined to create additional features - this will aid overfitting and should NOT be done.

Wish questions and answer options are written clearly so that there is no room for ambiguity. Especially, taking into account that in real life, these kind of communication/write-up will trigger follow-up questions until addressed satisfactorily.

upvoted 1 times

Denise123 1 year, 2 months ago

Selected Answer: BCE

About option E:

When increasing feature combinations, the goal is not to simply add more features indiscriminately, which could indeed lead to overfitting. Instead, it

involves selecting and combining features in a way that captures important patterns and relationships in the data.

When done effectively, increasing feature combinations can help the model generalize better to unseen data by providing more informative and discriminative features, thus reducing the risk of overfitting.

upvoted 1 times

🗳️ 👤 **Piyush_N** 1 year, 3 months ago

Selected Answer: BCF

If your model is overfitting the training data, it makes sense to take actions that reduce model flexibility. To reduce model flexibility, try the following:

Feature selection: consider using fewer feature combinations, decrease n-grams size, and decrease the number of numeric attribute bins.

Increase the amount of regularization used.

<https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html>

upvoted 1 times

🗳️ 👤 **Neet1983** 1 year, 6 months ago

Selected Answer: BCF

Best choices are B (Increase regularization), C (Increase dropout), and F (Decrease feature combinations), as these techniques are effective in reducing overfitting and improving the model's ability to generalize to new data.

upvoted 1 times

🗳️ 👤 **akgarg00** 1 year, 7 months ago

Selected Answer: BCE

BCE The model has learnt training data. One approach is to increase complexity by increasing the features or remove some features to increase bias. In deep learning, i thinking increasing feature set is more workable.

upvoted 1 times

🗳️ 👤 **kaike_reis** 1 year, 11 months ago

Selected Answer: BCF

B-C-F. All of those options can be used to reduce model complexity and thus: overfit

upvoted 1 times

🗳️ 👤 **SRB1337** 2 years ago

its BCF

upvoted 1 times

🗳️ 👤 **jackzhao** 2 years, 3 months ago

BCF is correct.

upvoted 2 times

🗳️ 👤 **Ajose0** 2 years, 4 months ago

Selected Answer: BCF

Increasing regularization helps to prevent overfitting by adding a penalty term to the loss function to discourage the model from learning the noise in the data.

Increasing dropout helps to prevent overfitting by randomly dropping out some neurons during training, which forces the model to learn more robust representations that do not depend on the presence of any single neuron.

Decreasing the number of feature combinations helps to simplify the model, making it less likely to overfit.

upvoted 6 times

🗳️ 👤 **Tomatoteacher** 2 years, 5 months ago

Selected Answer: BCE

I see all the comments for BCF, although when you look at F it just says decrease 'feature combinations', not features themselves. In one way to decrease feature combinations results in having more features (less feature engineering), which in turn will cause more overfitting. Unless the question is badly worded, saying less feature combinations just mean those combinations, which components will not be used, then it has to be BCE.

upvoted 1 times

🗳️ 👤 **cpal012** 2 years, 2 months ago

Decrease feature combinations - too many irrelevant features can influence the model by drowning out the signal with noise

upvoted 1 times

🗳️ 👤 **Ajose0** 2 years, 4 months ago

Increasing the number of feature combinations can sometimes improve the performance of a model if the model is underfitting the data.

However, in this context, it is not likely to be a solution to overfitting.

upvoted 1 times

🗨️ 👤 **Shailendraa** 2 years, 9 months ago

BCF - Always remember in case of overfitting - reduce features, Add regularisation and increase dropouts.

upvoted 3 times

🗨️ 👤 **ahquiceno** 3 years, 8 months ago

BCE: The main objective of PCA (technic to feature combination) is to simplify your model features into fewer components to help visualize patterns in your data and to help your model run faster. Using PCA also reduces the chance of overfitting your model by eliminating features with high correlation.

<https://towardsdatascience.com/dealing-with-highly-dimensional-data-using-principal-component-analysis-pca-fea1ca817fe6>

upvoted 2 times

🗨️ 👤 **unitit** 2 years, 5 months ago

AWS Documentation explicitly mentions reducing feature combinations to prevent overfitting - <https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html>

It's B C F

upvoted 3 times

🗨️ 👤 **cloud_trail** 3 years, 8 months ago

B/C/F Easy peasy.

upvoted 1 times

🗨️ 👤 **apnu** 3 years, 8 months ago

BCF 100%

upvoted 1 times

🗨️ 👤 **obaidur** 3 years, 8 months ago

BCF

F

explained in AWS document:

Feature selection: consider using fewer feature combinations, decrease n-grams size, and decrease the number of numeric attribute bins.

Increase the amount of regularization used

<https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html>

upvoted 5 times

A Data Scientist needs to create a serverless ingestion and analytics solution for high-velocity, real-time streaming data.

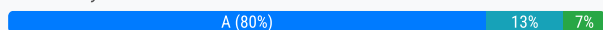
The ingestion process must buffer and convert incoming records from JSON to a query-optimized, columnar format without data loss. The output datastore must be highly available, and Analysts must be able to run SQL queries against the data and connect to existing business intelligence dashboards.

Which solution should the Data Scientist build to satisfy the requirements?

- A. Create a schema in the AWS Glue Data Catalog of the incoming data format. Use an Amazon Kinesis Data Firehose delivery stream to stream the data and transform the data to Apache Parquet or ORC format using the AWS Glue Data Catalog before delivering to Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena, and connect to BI tools using the Athena Java Database Connectivity (JDBC) connector.
- B. Write each JSON record to a staging location in Amazon S3. Use the S3 Put event to trigger an AWS Lambda function that transforms the data into Apache Parquet or ORC format and writes the data to a processed data location in Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena, and connect to BI tools using the Athena Java Database Connectivity (JDBC) connector.
- C. Write each JSON record to a staging location in Amazon S3. Use the S3 Put event to trigger an AWS Lambda function that transforms the data into Apache Parquet or ORC format and inserts it into an Amazon RDS PostgreSQL database. Have the Analysts query and run dashboards from the RDS database.
- D. Use Amazon Kinesis Data Analytics to ingest the streaming data and perform real-time SQL queries to convert the records to Apache Parquet before delivering to Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena and connect to BI tools using the Athena Java Database Connectivity (JDBC) connector.

Suggested Answer: A

Community vote distribution



🗳️ **DonaldCMLIN** Highly Voted 3 years, 3 months ago

Kinesis Data Analytics NO PARQUET FORMAT,
BESIDES THAT JSON NO NEED TO STORE IN S3.
RDS ISN'T serverless ingestion and analytics solution

ANSWER IS A.

upvoted 32 times

🗳️ **georgeZ** Highly Voted 3 years, 3 months ago

I think it should be A please check <https://aws.amazon.com/blogs/big-data/analyzing-apache-parquet-optimized-data-using-amazon-kinesis-data-firehose-amazon-athena-and-amazon-redshift/>

upvoted 14 times

🗳️ **JonSno** Most Recent 4 months, 1 week ago

Selected Answer: A

Amazon Kinesis Data Firehose

Ingests real-time data with automatic buffering.

Supports built-in transformation to Apache Parquet/ORC before writing to Amazon S3.

Requires minimal code and infrastructure.

AWS Glue Data Catalog

Catalogs the schema for structured querying.

Enables Athena to directly query data in S3.

Amazon Athena

Serverless SQL querying on S3-based datasets.

Can connect to BI tools (Tableau, QuickSight) via JDBC.

upvoted 1 times

🗳️ **Alice1234** 10 months, 3 weeks ago

A. Create a schema in the AWS Glue Data Catalog of the incoming data format. Use Amazon Kinesis Data Firehose to buffer and transform the streaming JSON data to a columnar format like Apache Parquet or ORC using the AWS Glue Data Catalog before delivering to Amazon S3. Analysts can then query the data using Amazon Athena and connect to BI dashboards using the Athena JDBC connector. This solution is serverless, manages high-velocity data streams, supports SQL queries, and connects to BI tools—all while being highly available.

upvoted 3 times

🗨️ **loict** 1 year, 3 months ago

Selected Answer: C

A. YES - we need a catalog to create parquet (https://docs.aws.amazon.com/firehose/latest/APIReference/APL_SchemaConfiguration.html)

B. NO - no need for extra staging

C. NO - no need for extra staging

D. NO - we need a catalog

upvoted 1 times

🗨️ **Mickey321** 1 year, 4 months ago

Selected Answer: A

Option A

upvoted 1 times

🗨️ **kaike_reis** 1 year, 5 months ago

Selected Answer: A

A is correct. For those selecting B, answer me: how exactly the json will be stored in the S3? It's not mentioned in the answer. For me it's an incomplete solution.

upvoted 2 times

🗨️ **Ajose0** 1 year, 10 months ago

Selected Answer: A

This solution leverages AWS Glue to create a schema of the incoming data format, which helps to buffer and convert the records to a query-optimized, columnar format without data loss.

The Amazon Kinesis Data Firehose delivery stream is used to stream the data and transform it to Apache Parquet or ORC format using the AWS Glue Data Catalog, and the data is stored in Amazon S3, which is highly available. The Analysts can then query the data directly from Amazon S3 using Amazon Athena, and connect to BI tools using the Athena JDBC connector.

This solution provides a serverless, scalable, and cost-effective solution for real-time streaming data ingestion and analytics.

upvoted 3 times

🗨️ **sqavi** 1 year, 10 months ago

Selected Answer: A

Since you want to buffer and convert data so A is correct answer. No other option is fulfilling this requirement

upvoted 2 times

🗨️ **Peeking** 2 years ago

Selected Answer: A

I go for A. However, I am not sure why AWS Glue is very important here given that Firehose can convert JSON to parquet.

upvoted 2 times

🗨️ **Tony_1406** 1 year, 8 months ago

If I haven't remembered correctly. Athena requires a schema of the S3 object to perform SQL query. That's probably why we need Glue for the schema

upvoted 1 times

🗨️ **ZSun** 1 year, 8 months ago

once you ingest the data using Kinesis Firehose, you can set "generate table" and automatically create Glue schema. I think both Glue and Firehose can do data conversion from JSON to parquet.

upvoted 1 times

🗨️ **itallond** 2 years ago

Why AWS Glue is needed? Firehose could convert to parquet directly...

upvoted 2 times

🗨️ **587df71** 5 months, 3 weeks ago

<https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html>

Amazon Data Firehose requires a schema to determine how to interpret that data. Use AWS Glue to create a schema in the AWS Glue Data Catalog. Amazon Data Firehose then references that schema and uses it to interpret your input data

upvoted 1 times

🗨️ 👤 **Ccindy** 2 years, 1 month ago

Selected Answer: B

Kinesis Data Analytics is near real-time, not real time

upvoted 1 times

🗨️ 👤 **ryuhei** 2 years, 3 months ago

Selected Answer: A

Answer is "A"

upvoted 1 times

🗨️ 👤 **ovokpus** 2 years, 6 months ago

Selected Answer: A

The difference between "real-time" and "near-real-time" is pretty semantic(60s). The fact that the data comes through kinesis data streams (real time) is implied as the only valid input to firehose.

upvoted 1 times

🗨️ 👤 **ovokpus** 2 years, 6 months ago

Mind you, "the ingestion process must buffer and transform incoming records from JSON to a query-optimized, columnar format"

That is exactly what kinesis firehose does.

"Kinesis Data Firehose buffers incoming data before delivering it to Amazon S3. You can configure the values for S3 buffer size (1 MB to 128 MB) or buffer interval (60 to 900 seconds), and the condition satisfied first triggers data delivery to Amazon S3."

See link: [https://aws.amazon.com/kinesis/data-](https://aws.amazon.com/kinesis/data-firehose/faqs/#:~:text=Kinesis%20Data%20Firehose%20buffers%20incoming,data%20delivery%20to%20Amazon%20S3.)

[firehose/faqs/#:~:text=Kinesis%20Data%20Firehose%20buffers%20incoming,data%20delivery%20to%20Amazon%20S3.](https://aws.amazon.com/kinesis/data-firehose/faqs/#:~:text=Kinesis%20Data%20Firehose%20buffers%20incoming,data%20delivery%20to%20Amazon%20S3.)

upvoted 3 times

🗨️ 👤 **TerrancePythonJava** 2 years, 9 months ago

Selected Answer: B

Data Firehose is always Near Real Time not Real Time. The prompt clearly states that process must be done in Real Time.

upvoted 1 times

🗨️ 👤 **anttan** 3 years ago

Why A? Firehose is near real-time, and not real-time which is a requirement

upvoted 1 times

🗨️ 👤 **cpal012** 1 year, 9 months ago

There is no requirement for real time processing. It says the data is in real time but the processing of that data should buffer

upvoted 2 times

🗨️ 👤 **harmanbirstudy** 3 years, 2 months ago

ANSWER is A -- and every statement in it is accurate.

Firehose does integrate with GLue data catalog and it also "Buffers" the data .

"When Kinesis Data Firehose processes incoming events and converts the data to Parquet, it needs to know which schema to apply." This is achieved by glue data catalog and athena and it works on real-time data ingest. See link below.

<https://aws.amazon.com/blogs/big-data/analyzing-apache-parquet-optimized-data-using-amazon-kinesis-data-firehose-amazon-athena-and-amazon-redshift/>

upvoted 5 times

An online reseller has a large, multi-column dataset with one column missing 30% of its data. A Machine Learning Specialist believes that certain columns in the dataset could be used to reconstruct the missing data.

Which reconstruction approach should the Specialist use to preserve the integrity of the dataset?

- A. Listwise deletion
- B. Last observation carried forward
- C. Multiple imputation
- D. Mean substitution

Suggested Answer: C

Reference:


<https://worldwidescience.org/topicpages/i/imputing+missing+values.html>

Community vote distribution

C (100%)

 **rsimham** Highly Voted 2 years, 9 months ago

C looks correct since multiple imputation can be performed based on the related variable as given in the question
upvoted 26 times

 **harmanbirstudy** Highly Voted 2 years, 8 months ago

Multiple Imputation by Chained Equations or MICE, as per udemy this is always the best answer of all
upvoted 8 times

 **sonoluminescence** Most Recent 8 months ago

Why not D:

Doesn't Account for Relationships:

Mean substitution doesn't take into account the potential relationships between variables. In the scenario you provided, it's believed that other columns could help in reconstructing the missing data. Using only the mean of the missing column doesn't leverage this potential inter-column relationship.

Assumption of Missing Completely at Random (MCAR):

Mean substitution often operates under the assumption that the data is Missing Completely at Random (MCAR). In reality, data might be missing for a reason, and that reason might relate to other observed variables. Using mean substitution in such cases can introduce biases.

upvoted 2 times

 **loict** 9 months, 2 weeks ago

Selected Answer: C

- A. NO - Listwise deletion is just dropping rows
 - B. NO - does not reconstruct the data based on other fields
 - C. YES - by definition
 - D. NO - does not reconstruct the data based on other fields
- upvoted 2 times

 **DavidRou** 9 months, 3 weeks ago

Selected Answer: C

MICE is the algorithm to choose here
upvoted 1 times

 **Mickey321** 10 months ago

Selected Answer: C

Option C
upvoted 1 times

 **Ajose0** 1 year, 4 months ago

Selected Answer: C

Multiple imputation is a statistical technique for handling missing data that involves generating multiple versions of the dataset with missing values filled in, and then combining the results to produce a single, complete dataset.

This approach takes into account the relationship between variables in the dataset, and uses statistical models to predict missing values based on the information in other columns. This helps to preserve the integrity of the dataset by avoiding the introduction of bias or systematic error into the results.

upvoted 5 times

🗨️ 👤 **[Removed]** 2 years, 7 months ago

I am trying to understand why Mean Substitution is not the solution. Imputation typically uses the mean if the missing data is random, implying the substitution is not biased.

upvoted 2 times

🗨️ 👤 **cpal012** 1 year, 3 months ago

Mean substitution is limited to the current column. In this case, the requirement is to impute missing data from other columns

upvoted 3 times

🗨️ 👤 **rhuanca** 2 years, 1 month ago

Reason is if you replace 30% of the missing values , likely you will bias the variable.

upvoted 1 times

🗨️ 👤 **syu31svc** 2 years, 8 months ago

If it's handling missing data then imputation comes into play

Answer is C 100%

upvoted 1 times

🗨️ 👤 **Wira** 2 years, 8 months ago

<https://www.countants.com/blogs/heres-how-you-can-configure-automatic-imputation-of-missing-data/> C

upvoted 1 times

🗨️ 👤 **roytruong** 2 years, 8 months ago

it's C

upvoted 1 times

🗨️ 👤 **dhs227** 2 years, 8 months ago

A common strategy used to impute missing values is to replace missing values with the mean or median value. It is important to understand your data before choosing a strategy for replacing missing values. <https://docs.aws.amazon.com/machine-learning/latest/dg/feature-processing.html>

upvoted 2 times

A company is setting up an Amazon SageMaker environment. The corporate data security policy does not allow communication over the internet. How can the company enable the Amazon SageMaker service without enabling direct internet access to Amazon SageMaker notebook instances?

- A. Create a NAT gateway within the corporate VPC.
- B. Route Amazon SageMaker traffic through an on-premises network.
- C. Create Amazon SageMaker VPC interface endpoints within the corporate VPC.
- D. Create VPC peering with Amazon VPC hosting Amazon SageMaker.

Suggested Answer: A

Reference:

<https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf>

(46)

Community vote distribution

C (100%)

 **DonaldCMLIN**  3 years, 3 months ago

NAT CLOUD GO OUT TO THE INTERNET, IT STILL CANNOT PREVENT DOWNLOAD MALICIOUS BY YOURSELF.

THE RIGHT ANSWER IS C.

C.INTERFACE VPC ENDPOINT

<https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf> (516)

https://docs.aws.amazon.com/zh_tw/vpc/latest/userguide/vpc-endpoints.html

upvoted 46 times

 **rsimham** 3 years, 3 months ago

Not sure if C is correct in this particular scenario.

From <https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf>

Page 202 of the SageMaker Guide has:

If you allowed access to resources from your VPC, enable direct internet access. For Direct internet access, choose Enable. Without internet access, you can't train or host models from notebooks on this notebook instance unless your VPC has a NAT gateway and your security group allows outbound connect

upvoted 2 times


 **Selectron** 3 years, 2 months ago

There are two possible solutions, but the safer solution and easier is trough VPC endpoints.

You can connect to your notebook instance from your VPC through an interface endpoint in your Virtual Private Cloud (VPC) instead of connecting over the internet. When you use a VPC interface endpoint, communication between your VPC and the notebook instance is conducted entirely and securely within the AWS network. And there is not problem that the notebooks does not have public internet. Because Amazon SageMaker notebook instances support Amazon Virtual Private Cloud (Amazon VPC) interface endpoints that are powered by AWS PrivateLink. Each VPC endpoint is represented by one or more Elastic Network Interfaces (ENIs) with private IP addresses in your VPC subnets. Each VPC endpoint is represented by one or more Elastic Network Interfaces (ENIs) with private IP addresses in your VPC subnets...

so the Answer is C.

upvoted 5 times

 **rsimham** 3 years, 3 months ago

A may the right answer

upvoted 1 times

 **tap123**  3 years, 3 months ago

C is correct. "The VPC interface endpoint connects your VPC directly to the Amazon SageMaker API or Runtime without an internet gateway, **NAT** device, VPN connection, or AWS Direct Connect connection." <https://docs.aws.amazon.com/sagemaker/latest/dg/interface-vpc-endpoint.html>

upvoted 16 times

  **JonSno**  4 months, 1 week ago

Selected Answer: C

Explanation:

The company's data security policy does not allow internet access, so the solution must allow Amazon SageMaker to function privately within the VPC without internet access.

VPC Interface Endpoints (AWS PrivateLink) for SageMaker allow services to communicate privately over the AWS network, without requiring an Internet Gateway (IGW) or NAT Gateway.

Explanation:

The company's data security policy does not allow internet access, so the solution must allow Amazon SageMaker to function privately within the VPC without internet access.

VPC Interface Endpoints (AWS PrivateLink) for SageMaker allow services to communicate privately over the AWS network, without requiring an Internet Gateway (IGW) or NAT Gateway.

upvoted 1 times

  **Denise123** 10 months, 3 weeks ago

The answer is C.



- If you want to allow internet access, you must use a NAT gateway with access to the internet, for example through an internet gateway.

- If you don't want to allow internet access, create interface VPC endpoints (AWS PrivateLink) to allow Studio Classic to access the following services with the corresponding service names. You must also associate the security groups for your VPC with these endpoints.

This is exactly what's written in the ref. doc given in the answer section of the question. (Check page Security and Permissions 1120- 1121)

<https://docs.aws.amazon.com/pdfs/sagemaker/latest/dg/sagemaker-dg.pdf>

upvoted 1 times

  **phdykd** 11 months, 4 weeks ago

C.

To enable Amazon SageMaker service without enabling direct internet access to Amazon SageMaker notebook instances, while adhering to a corporate data security policy that restricts internet communication, the company can:

C. Create Amazon SageMaker VPC interface endpoints within the corporate VPC.

This option involves setting up VPC (Virtual Private Cloud) interface endpoints for Amazon SageMaker within the corporate VPC (Virtual Private Cloud). This is done using AWS PrivateLink, which allows private connectivity between AWS services using private IP addresses. By creating VPC interface endpoints, the traffic between the corporate VPC and Amazon SageMaker does not traverse the public internet, thereby meeting the corporate data security requirements.

upvoted 1 times

  **sonoluminescence** 1 year, 2 months ago

Selected Answer: C

A would allow instances in a private subnet to initiate outbound internet traffic. This is against the requirement of no direct internet access.

upvoted 2 times

  **Sharath1783** 1 year, 3 months ago

Selected Answer: C

NAT means data will go to internet. C is the right choice.



upvoted 2 times

  **Mickey321** 1 year, 4 months ago

Selected Answer: C



Option c

upvoted 1 times

  **ADVIT** 1 year, 6 months ago

Only C, endpoints.

upvoted 1 times

  **jackzhao** 1 year, 9 months ago

C is correct, NAT allow outband traffic pass through internet.

upvoted 1 times

🗨️ 👤 **Nadia0012** 1 year, 9 months ago

Selected Answer: C

To prevent SageMaker from providing internet access to your Studio notebooks, you can disable internet access by specifying the VPC only network access to Studio or call CreateDomain API. As a result, you won't be able to run a Studio notebook unless your VPC has an interface endpoint to the SageMaker API and with internet access, and your security groups allow outbound connections.

upvoted 2 times

🗨️ 👤 **Nadia0012** 1 year, 9 months ago

To disable direct internet access, under Direct Internet access, simply choose Disable – use VPC only, and select the Create notebook instance button at go.

from: <https://aws.amazon.com/blogs/machine-learning/customize-your-amazon-sagemaker-notebook-instances-with-lifecycle-configurations-and-the-opt-access/#:~:text=To%20disable%20direct%20internet%20access%2C%20under%20Direct%20Internet%20access%2C%20simply,running%2C%20without%20>

upvoted 1 times

🗨️ 👤 **Nadia0012** 1 year, 9 months ago

If you want to allow internet access, you must use a example through an internet gateway. If you don't want to allow internet access, NAT gateway with an interface VPC endpoints (AWS PrivateLink) to allow Studio to access the following services with the corresponding service names. You must also associate your VPC with these endpoints.

upvoted 1 times

🗨️ 👤 **Ajose0** 1 year, 10 months ago

Selected Answer: C

A VPC interface endpoint is a private connection between a VPC and Amazon SageMaker that is powered by AWS PrivateLink. With a VPC interface endpoint, traffic between the VPC and Amazon SageMaker never leaves the Amazon network.

upvoted 3 times

🗨️ 👤 **Ob1KN0B** 2 years, 4 months ago

Selected Answer: C

Page 3438 of <https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf>

upvoted 2 times

🗨️ 👤 **ovokpus** 2 years, 6 months ago

Selected Answer: C

VPC Interface endpoints

upvoted 3 times

🗨️ 👤 **gcpwhiz** 3 years, 1 month ago

If the question just had the last sentence, the answer would be A or C, per this page: <https://docs.aws.amazon.com/sagemaker/latest/dg/appendix-notebook-and-internet-access.html>. "To disable direct internet access, you can specify a VPC for your notebook instance. By doing so, you prevent SageMaker from providing internet access to your notebook instance. As a result, the notebook instance won't be able to train or host models unless your VPC has an interface endpoint (PrivateLink) or a NAT gateway, and your security groups allow outbound connections."

HOWEVER, the question has more context that internet access is not allowed by the corporate policy. ("When you use a VPC interface endpoint, communication between your VPC and the notebook instance is conducted entirely and securely within the AWS network.") Therefore, the answer must be ONLY C.

upvoted 5 times

🗨️ 👤 **scuzzy2010** 3 years, 1 month ago

Answer is C. From <https://docs.aws.amazon.com/sagemaker/latest/dg/interface-vpc-endpoint.html> ->

"The VPC interface endpoint connects your VPC directly to the SageMaker API or Runtime without an internet gateway, NAT device, VPN connection, or AWS Direct Connect connection. The instances in your VPC don't need public IP addresses to communicate with the SageMaker API or Runtime."

upvoted 3 times

🗨️ 👤 **cloud_trail** 3 years, 1 month ago

I see a lot of people employing pretzel logic to try to explain why they should be using NAT. The question states no internet communication. Period. No internet means no NAT. Answer is C.

upvoted 5 times

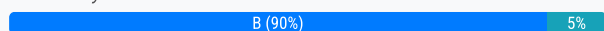
A Machine Learning Specialist is training a model to identify the make and model of vehicles in images. The Specialist wants to use transfer learning and an existing model trained on images of general objects. The Specialist collated a large custom dataset of pictures containing different vehicle makes and models.

What should the Specialist do to initialize the model to re-train it with the custom data?

- A. Initialize the model with random weights in all layers including the last fully connected layer.
- B. Initialize the model with pre-trained weights in all layers and replace the last fully connected layer.
- C. Initialize the model with random weights in all layers and replace the last fully connected layer.
- D. Initialize the model with pre-trained weights in all layers including the last fully connected layer.

Suggested Answer: B

Community vote distribution



rsimham Highly Voted 3 years, 9 months ago

Ans B sounds correct

upvoted 28 times

Ajose0 Highly Voted 2 years, 4 months ago

Selected Answer: B

In transfer learning, a pre-trained model is used as a starting point to train a new model on a different task, typically using a smaller dataset. The pre-trained model contains weights that have been learned from a large amount of data on a related task, and these weights can be leveraged to train the new model more efficiently.

To re-train the model with the custom data, the Specialist should initialize the model with pre-trained weights in all layers, as these weights can provide a good starting point for the new task. The Specialist should then replace the last fully connected layer, which is responsible for making the final predictions, as this layer will likely need to be modified to reflect the new task. By keeping the pre-trained weights in the other layers, the Specialist can take advantage of the knowledge learned from the previous task, and potentially speed up the training process.

upvoted 9 times

JonSno Most Recent 4 months, 1 week ago

Selected Answer: B

Explanation:

The Machine Learning Specialist wants to use transfer learning with an existing model trained on general object images and fine-tune it for vehicle make and model classification. The best approach is:

Use pre-trained weights from the existing model for feature extraction.

Replace the last fully connected (FC) layer to match the number of vehicle classes.

Fine-tune the new model on the vehicle dataset.

Why This Works?

Lower training time: The model has already learned useful features from general objects (e.g., edges, shapes).

Improves accuracy: Instead of training from scratch, transfer learning leverages knowledge from large datasets (e.g., ImageNet).

Avoids catastrophic forgetting: Reusing pre-trained weights preserves learned low- and mid-level features while adapting the last layer for new classes.

upvoted 1 times

itsme1 9 months, 3 weeks ago

Selected Answer: D

Transfer learning helps accelerate the training and at this point, model has yet to learn from the new data. So, all layers including the fully-connected by replaced. Eventually, the training will update the fully-connected layer. The question is about initialization, so we should initialize the fully-connected layers too.

upvoted 1 times

loict 1 year, 9 months ago

Selected Answer: B

- A. NO - random weights does not allow transfer learning
- B. YES - the last layer gives the final classes, we want to have new classes
- C. NO - random weights does not allow transfer learning
- D. NO - the last layer gives the final classes, we want to have new classes

upvoted 2 times

🗨️ **Mickey321** 1 year, 10 months ago

Selected Answer: B

Option B

upvoted 1 times

🗨️ **kaike_reis** 1 year, 11 months ago

Selected Answer: B

For Transfer Learning, A and C are incorrect because we restart the model. The correct is letter B

upvoted 2 times

🗨️ **SRB1337** 2 years ago

B. The reason is, fine-tuning a model means to use the weights/biases trained before. also no matter which strategy you go for in transfer learning (fine-tuning or feature extraction) you always replace the last or last few layers.

upvoted 3 times

🗨️ **mirik** 2 years ago

Selected Answer: C

The task is to "re-train it with the custom data". That means, it is not transfer learning anymore. The "transfer learning" is just a title to make a question tricky.

So, in this case we should randomize the weights and retrain whole model from scratch on custom user's images only.

The correct answer is C.

upvoted 1 times

🗨️ **FloKo** 1 year, 11 months ago

I think retraining refers in this context to the training on the custom data that the expert has already conducted before thinking about transfer learning.

upvoted 1 times

🗨️ **mirik** 2 years ago

The task is to "re-train it with the custom data". That means, it is not transfer learning anymore. The "transfer learning" is just a title to make a question tricky.

So, in this case we should randomize the weights and retrain whole model from scratch on custom user's images only.

The correct answer is C.

upvoted 1 times

🗨️ **Peeking** 2 years, 6 months ago

Selected Answer: B

The fully connected layer will need to be trained from scratch to incorporate the features of his domain problem (Car models)

upvoted 3 times

🗨️ **Shailendraa** 2 years, 9 months ago

12-sep exam

upvoted 2 times

🗨️ **chrisdavid** 2 years, 11 months ago

D is the best - here is why

Question is not to design a final production with deep learning - it is to use it as a dev platform to come up with an edge ML vs. dump load all to S3 - which is very wasteful! AWS did not make deep learning as a toy for devs! it is meant to help companies experiment with edge ML And then copy and reuse the open hardware platform

upvoted 1 times

🗨️ **ckkobe24** 3 years, 1 month ago

Selected Answer: B

one of the methods to implement transfer learning

upvoted 2 times

🗨️ **DzR** 3 years, 8 months ago



I will go with B, we are mainly concerned with the output layer for us to get the desired results, hence we need to replace it.

upvoted 1 times

  **bobdylan1** 3 years, 8 months ago

B is correct

upvoted 2 times

  **sebtac** 3 years, 8 months ago

Actually, it should be NONE of IT!.... it should be like B with exception that 20-40% top layers should be retrained :) -- this is classic transfer learning setup, so B is the answer here.

upvoted 2 times

An office security agency conducted a successful pilot using 100 cameras installed at key locations within the main office. Images from the cameras were uploaded to Amazon S3 and tagged using Amazon Rekognition, and the results were stored in Amazon ES. The agency is now looking to expand the pilot into a full production system using thousands of video cameras in its office locations globally. The goal is to identify activities performed by non-employees in real time

Which solution should the agency consider?

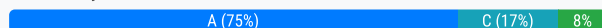
- A. Use a proxy server at each local office and for each camera, and stream the RTSP feed to a unique Amazon Kinesis Video Streams video stream. On each stream, use Amazon Rekognition Video and create a stream processor to detect faces from a collection of known employees, and alert when non-employees are detected.
- B. Use a proxy server at each local office and for each camera, and stream the RTSP feed to a unique Amazon Kinesis Video Streams video stream. On each stream, use Amazon Rekognition Image to detect faces from a collection of known employees and alert when non-employees are detected.
- C. Install AWS DeepLens cameras and use the DeepLens_Kinesis_Video module to stream video to Amazon Kinesis Video Streams for each camera. On each stream, use Amazon Rekognition Video and create a stream processor to detect faces from a collection on each stream, and alert when non-employees are detected.
- D. Install AWS DeepLens cameras and use the DeepLens_Kinesis_Video module to stream video to Amazon Kinesis Video Streams for each camera. On each stream, run an AWS Lambda function to capture image fragments and then call Amazon Rekognition Image to detect faces from a collection of known employees, and alert when non-employees are detected.

Suggested Answer: D

Reference:

<https://aws.amazon.com/blogs/machine-learning/video-analytics-in-the-cloud-and-at-the-edge-with-aws-deeplens-and-kinesis-video-streams/>

Community vote distribution



scuzzy2010 Highly Voted 3 years, 2 months ago

Answer is "A". C and D are out as DeepLens is not offered as a commercial product. It is purely for developers to experiment with.

From <https://aws.amazon.com/deeplens/device-terms-of-use/>

"(i) you may use the AWS DeepLens Device for personal, educational, evaluation, development, and testing purposes, and not to process your production workloads;"

A is correct as it's will analyse live video streams instead of images.

From <https://aws.amazon.com/rekognition/video-features/>

"Amazon Rekognition Video can identify known people in a video by searching against a private repository of face images. "

upvoted 42 times

kaike_reis 1 year, 5 months ago

Agree as well, besides that: (D) uses Rekognition with Image mode, which is wrong for this case.

upvoted 1 times

Mezaji 3 years, 1 month ago

Agreed

upvoted 2 times

WWODIN Highly Voted 3 years, 3 months ago

Why not A?

DeepLens is for development purpose and much more expensive than just a camera.

They are referring to 1000 camera in production scale?

upvoted 12 times

cybe001 3 years, 3 months ago



C is the correct answer. We could use A, since it is for security service, DeepLens allows to notify the security (through aws lamda) immediately when it sees non employee at the office location. So C is more appropriate for the problem than A.

upvoted 6 times

scuzzy2010 3 years, 2 months ago

DeepLens is for developers only, it is not available as a commercial product.

upvoted 5 times

  **sdsfsdsf** 3 years, 3 months ago

A bit off topic but yeah, how could you justify using deep lens for production. Cameras have viewing angles, weather proofing, network connectivity issues (Wifi only), infra red for low lighting conditions, no power over ethernet? Using Deeplens would be laughable for a full production system.

upvoted 8 times

  **JonSno** Most Recent 4 months, 1 week ago

Selected Answer: A

Ans - A -- Proxy Server + Kinesis Video Streams + Rekognition Video

The goal is to scale from 100 cameras to thousands and perform real-time detection of non-employees in office locations globally. The best approach is to use Amazon Kinesis Video Streams + Amazon Rekognition Video for real-time face detection.

upvoted 1 times

  **Denise123** 10 months, 3 weeks ago

The correct answer is D.

Very tricky one but re-read the 2nd sentence in the question;



"Images from the cameras were uploaded to Amazon S3 and tagged using Amazon Rekognition, and the results were stored in Amazon ES."

So, we have 'images' as training data, not videos. This is why it can not be option C - where it says to use Amazon Recognition Video. The only option mentioning Amazon Recognition Image is the option D.

Also check: <https://docs.aws.amazon.com/rekognition/latest/dg/what-is.html>

"...For example, each time a person arrives at your residence, your door camera can upload a photo of the visitor to Amazon S3. This triggers a Lambda function that uses Amazon Rekognition API operations to identify your guest. You can run analysis directly on images that are stored in Amazon S3 without having to load or move the data."

upvoted 3 times

  **phdykd** 11 months, 4 weeks ago

A is answer

upvoted 1 times

  **sukye** 1 year, 1 month ago

Selected Answer: A

A not B: Use Amazon Rekognition Video instead of Amazon Rekognition Image in this case.

upvoted 2 times

  **elvin_ml_qayiran25091992razor** 1 year, 1 month ago

Selected Answer: A

A is correct!

upvoted 1 times

  **sonoluminescence** 1 year, 2 months ago

Selected Answer: A

DeepLens is overkill for mass systems

upvoted 1 times

  **Ioict** 1 year, 3 months ago

Selected Answer: C

A. NO - thousands of cameras would choke network bandwidth

B. NO - thousands of cameras would choke network bandwidth

C. YES - DeepLens is made for edge computing; it might be EOL / Not commercially available, but if they did not want you to use DeepLens the question would not have come in the first place

D. NO - use Amazon Rekognition Video directly instead of Amazon Rekognition Image

upvoted 2 times

  **DavidRou** 1 year, 3 months ago

Why A and not B? Can someone please explain it?

upvoted 1 times

  **Mickey321** 1 year, 4 months ago

Selected Answer: A

Option A

upvoted 1 times

🗨️ 👤 **strike3test** 1 year, 4 months ago

From Chat GPT

The solution that the agency should consider is option A: Use a proxy server at each local office and for each camera, and stream the RTSP feed to a unique Amazon Kinesis Video Streams video stream. On each stream, use Amazon Rekognition Video and create a stream processor to detect faces from a collection of known employees and alert when non-employees are detected.

By using a proxy server at each local office and streaming the RTSP feed to individual Amazon Kinesis Video Streams video streams, the agency can efficiently handle the large number of video cameras in different office locations. Using Amazon Rekognition Video, the agency can create a stream processor to detect faces from a collection of known employees. This allows for real-time identification of non-employees based on facial recognition. Alerts can then be generated when non-employees are detected, ensuring that the agency is able to identify and respond to potential security threats in real-time.

upvoted 2 times

🗨️ 👤 **nilmans** 1 year, 6 months ago

I initially thought it is C but looks like A makes more sense here.

upvoted 1 times

🗨️ 👤 **jyrajan69** 1 year, 7 months ago

The DeepLens Service will reach EOL at the end of Jan 2024, so more than likely that this question will not be asked in the exam

upvoted 2 times

🗨️ 👤 **Valcilio** 1 year, 9 months ago

Selected Answer: D

D is the answer now, DeepLens is used for situations like this!

upvoted 1 times

🗨️ 👤 **cpal012** 1 year, 9 months ago

Maybe, its EOL Jan 2024

upvoted 2 times

🗨️ 👤 **expertguru** 1 year, 11 months ago

Think big picture - you tested something (let say code python) and ready to implement into prod will you move python code or java code! Here in this particular case, they tested with actual video camera and they did not say deeplense so answer is A! For knowledge sake if they say in real exam it is tested with deeplense ---then ideal solution should be model inference happening at deeplense itself with search against existing employees and send back model inference when it detect new faces who are not employees back to cloud may be S3

upvoted 2 times

🗨️ 👤 **Sivadharan** 2 years, 7 months ago

Selected Answer: A

Answer is "A".

As mentioned in below user comment, DeepLens is not offered as a commercial product.

<https://aws.amazon.com/deeplens/device-terms-of-use/>

upvoted 4 times

A Marketing Manager at a pet insurance company plans to launch a targeted marketing campaign on social media to acquire new customers. Currently, the company has the following data in Amazon Aurora:

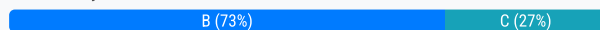
- ⇒ Profiles for all past and existing customers
- ⇒ Profiles for all past and existing insured pets
- ⇒ Policy-level information
- ⇒ Premiums received
- ⇒ Claims paid

What steps should be taken to implement a machine learning model to identify potential new customers on social media?

- A. Use regression on customer profile data to understand key characteristics of consumer segments. Find similar profiles on social media
- B. Use clustering on customer profile data to understand key characteristics of consumer segments. Find similar profiles on social media
- C. Use a recommendation engine on customer profile data to understand key characteristics of consumer segments. Find similar profiles on social media.
- D. Use a decision tree classifier engine on customer profile data to understand key characteristics of consumer segments. Find similar profiles on social media.

Suggested Answer: C

Community vote distribution



🗳️ **DonaldCMLIN** Highly Voted 3 years, 9 months ago

All of the questions in the preceding examples rely on having example data that includes answers. There are times that you don't need, or can't get, example data with answers. This is true for problems whose answers identify groups. For example:

"I want to group current and prospective customers into 10 groups based on their attributes. How should I group them? " You might choose to send the mailing to customers in the group that has the highest percentage of current customers. That is, prospective customers that most resemble current customers based on the same set of attributes. For this type of question, Amazon SageMaker provides the K-Means Algorithm.

<https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html>

Clustering algorithms are unsupervised. In unsupervised learning, labels that might be associated with the objects in the training dataset aren't used. <https://docs.aws.amazon.com/sagemaker/latest/dg/algo-kmeans-tech-notes.html>

THE ANSWER COULD BE B.clustering on customer profile data to understand key characteristic
upvoted 37 times

🗳️ **rsimham** 3 years, 9 months ago

Yes, Clustering seems to be more appropriate in this scenario than recommender system
upvoted 10 times

🗳️ **mirik** 2 years ago

Collaborative filtering recommendation system is also unsupervised
upvoted 1 times

🗳️ **haison8x** 3 years, 8 months ago

<https://towardsdatascience.com/customer-segmentation-with-machine-learning-a0ac8c3d4d84>

B
upvoted 3 times

🗳️ **cloud_trail** Highly Voted 3 years, 8 months ago

Option C. This is not purely unsupervised, as clustering would be, because we have current and past customer profiles to go on. We want to find new customers by finding similar profiles on social media. So it is supervised to some extent. It's not a cluster problem; it is user-user collaborative filtering. The key is to recognize that this is not clustering. You're not blindly trying to group people. You have existing profiles that you are comparing them to.

upvoted 11 times

🗨️ 👤 **MultiCloudIronMan** Most Recent 8 months ago

Selected Answer: B

'B' is correct

upvoted 1 times

🗨️ 👤 **VR10** 1 year, 4 months ago

It is B. Recommendation Engines: Traditionally focus on suggesting products/services to existing customers based on past behavior.

upvoted 2 times

🗨️ 👤 **elvin_ml_qayiran25091992razor** 1 year, 7 months ago

Selected Answer: B

Clustering is right

upvoted 2 times

🗨️ 👤 **DimLam** 1 year, 8 months ago

Selected Answer: B

C would be an answer if wanted to send the promo to the existing customers. But we want to find potential customers. And we can do it only by comparing existing customers with potential customers. It can be done by creating clusters of existing customers and measuring the distance to those clusters for the new potential users.

So my answer is B

upvoted 3 times

🗨️ 👤 **loict** 1 year, 9 months ago

Selected Answer: C

A. NO - Linear Regression not best to understand relationships between data

B. NO - it is supervised (we know premiums received vs. claims paid, so can assign users to GOOD or BAD), so no clustering

C. YES - A recommendation engine in AWS lingua is Amazing Recommender (<https://docs.aws.amazon.com/personalize/latest/dg/what-is-personalize.html> - "Creating a targeted marketing campaign") and can create user segments

D. NO - not as good as C

upvoted 4 times

🗨️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: B

B for me

upvoted 1 times

🗨️ 👤 **teka112233** 1 year, 10 months ago

Selected Answer: B

Recommendation engines is perfect for customers we have, but for implementing a machine learning model to identify potential (new customers on social media) this requires clustering and segmentation.

<https://neptune.ai/blog/customer-segmentation-using-machine-learning>

upvoted 2 times

🗨️ 👤 **jyrajan69** 1 year, 11 months ago

Based on the link below, it must be C

<https://medium.com/voice-tech-podcast/a-simple-way-to-explain-the-recommendation-engine-in-ai-d1a609f59d97>

upvoted 1 times

🗨️ 👤 **kaike_reis** 1 year, 11 months ago

Selected Answer: B

We are divided, but I stick with B.

upvoted 1 times

🗨️ 👤 **Venkatesh_Babu** 1 year, 11 months ago

Selected Answer: C

I think it should be c

upvoted 1 times

🗨️ 👤 **nilmans** 2 years ago

Selected Answer: C

recommender system would help here, as we already have details of all customers

upvoted 1 times

🗨️ 👤 **nilmans** 2 years ago

it should be C - recommender system would be better fit here.

upvoted 2 times

🗨️ 👤 **mirik** 2 years ago

Selected Answer: C

We should use recommendation system to find key characteristics only among company users (past and present). At this step we don't take any users from the web. After we finish processing this CF model we identify key characteristics (important features?) and only after that, we will start looking for similar users on the web.

upvoted 1 times

🗨️ 👤 **earthMover** 2 years, 1 month ago

Selected Answer: B

I would use clustering technique to identify which customers in my database are the target audience and get similar customer profiles from the social media dataset. Its a lot simpler

upvoted 2 times

🗨️ 👤 **vbal** 2 years, 1 month ago

recommendation engines can use either supervised or unsupervised learning. I can't find any reason to NOT use recommendation engine???

upvoted 3 times

A manufacturing company has a large set of labeled historical sales data. The manufacturer would like to predict how many units of a particular part should be produced each quarter.

Which machine learning approach should be used to solve this problem?

- A. Logistic regression
- B. Random Cut Forest (RCF)
- C. Principal component analysis (PCA)
- D. Linear regression

Suggested Answer: B

Community vote distribution

D (90%)

10%

🗳️ **DonaldCMLIN** Highly Voted 3 years, 3 months ago

HOW MANY/MUCH, THOSE ARE REGRESSION TOPIC,
LOGISTIC FOR 0/1, YES/NO

https://docs.aws.amazon.com/zh_tw/machine-learning/latest/dg/regression-model-insights.html

THE ANSWER SHOULD BE D.

upvoted 62 times

🗳️ **rsimham** 3 years, 3 months ago

agree. RCF is mostly used for anomaly detection or separate outliers

upvoted 10 times

🗳️ **syu31svc** Highly Voted 3 years, 2 months ago

Amazon SageMaker Random Cut Forest (RCF) is an unsupervised algorithm for detecting anomalous data points within a data set

Answer is D 100%

upvoted 10 times

🗳️ **JonSno** Most Recent 4 months, 1 week ago

Selected Answer: D

The problem involves predicting the number of units to be produced each quarter based on historical sales data. This is a continuous numerical prediction, making it a regression problem.

Linear regression is ideal for forecasting when there is a linear relationship between input variables (e.g., past sales, seasonal trends) and the target variable (units to be produced).

It helps model the relationship between past sales and future demand.

If there are seasonal effects, a time-series model (like ARIMA or Prophet) could be considered as well.

upvoted 1 times

🗳️ **[Removed]** 7 months, 3 weeks ago

The Answer is D. Random Cut Forest is for Anomaly Detection

upvoted 1 times

🗳️ **t47** 8 months, 3 weeks ago

D should be the answer

upvoted 1 times

🗳️ **endeesa** 1 year, 1 month ago

Selected Answer: D

How many units should give this away as Linear regression

upvoted 1 times

🗳️ **AmeeraM** 1 year, 2 months ago

Selected Answer: D

I do not see any hint of anomalies here, we are looking for a number to be predicted, this seems to be the reason of the correct answer

<https://docs.aws.amazon.com/quicksight/latest/user/how-does-rcf-generate-forecasts.html>

upvoted 1 times

DavidRou 1 year, 3 months ago

Selected Answer: D

How can the right answer be B? That Random Cut Forest is an algorithm written for anomaly detection.

upvoted 3 times

Mickey321 1 year, 4 months ago

Selected Answer: D

option D

upvoted 1 times

kaike_reis 1 year, 5 months ago

Selected Answer: D

D is the correct. B is for outlier detection only.

upvoted 1 times

earthMover 1 year, 7 months ago

Selected Answer: D

It sounds like Linear regression problem and Random Cut is more known for anomaly detection while it can do other types of ML. The answer seems to be strange with no explanation.

upvoted 1 times

jackzhao 1 year, 9 months ago

D is correct!

upvoted 1 times

oso0348 1 year, 9 months ago

Selected Answer: D

D. Linear regression would be the appropriate machine learning approach to solve this problem of predicting the number of units of a particular part to be produced each quarter. Linear regression is a supervised learning algorithm used for predicting continuous variables based on input features. In this case, the historical sales data can be used as input features, and the number of units produced each quarter can be used as the continuous target variable.

upvoted 2 times

Nadia0012 1 year, 9 months ago

Selected Answer: D

definitely D.

upvoted 1 times

Ajose0 1 year, 10 months ago

Selected Answer: D

This is a regression problem where the goal is to predict a continuous outcome, which in this case is the number of units of a particular part that should be produced each quarter. Linear regression is a simple and commonly used approach to solve such problems, where a linear relationship is established between the independent variables (e.g., historical sales data) and the dependent variable (e.g., number of units of a part to be produced).

upvoted 2 times

Tomatoteacher 1 year, 11 months ago

Selected Answer: D

D, RCF answers here just link one article where RCF is implemented to find outliers in time series, or are able to deduce trends, but here they mention already labelled data, RCF is unsupervised, so that data would go to waste.

upvoted 1 times

hamimelon 2 years ago

Honestly, i think these are all bad answers. It should be time series modeling methods.

upvoted 2 times

A financial services company is building a robust serverless data lake on Amazon S3. The data lake should be flexible and meet the following requirements:

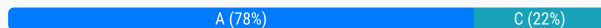
- ⇒ Support querying old and new data on Amazon S3 through Amazon Athena and Amazon Redshift Spectrum.
- ⇒ Support event-driven ETL pipelines
- ⇒ Provide a quick and easy way to understand metadata

Which approach meets these requirements?

- A. Use an AWS Glue crawler to crawl S3 data, an AWS Lambda function to trigger an AWS Glue ETL job, and an AWS Glue Data catalog to search and discover metadata.
- B. Use an AWS Glue crawler to crawl S3 data, an AWS Lambda function to trigger an AWS Batch job, and an external Apache Hive metastore to search and discover metadata.
- C. Use an AWS Glue crawler to crawl S3 data, an Amazon CloudWatch alarm to trigger an AWS Batch job, and an AWS Glue Data Catalog to search and discover metadata.
- D. Use an AWS Glue crawler to crawl S3 data, an Amazon CloudWatch alarm to trigger an AWS Glue ETL job, and an external Apache Hive metastore to search and discover metadata.

Suggested Answer: A

Community vote distribution



🗳️ 👤 **DonaldCMLIN** Highly Voted 2 years, 9 months ago

BOTH A AND B ARE ANSWERS.

BUT external Apache Hive MIGHT BE NOT SERVERLESS SOLUTION.

The AWS Glue Data Catalog is your persistent metadata store. It is a managed service that lets you store, annotate, and share metadata in the AWS Cloud in the same way you would in an Apache Hive metastore.

The Data Catalog is a drop-in replacement for the Apache Hive Metastore

https://docs.aws.amazon.com/zh_tw/glue/latest/dg/components-overview.html

BEAUTIFUL ANSWER IS A.

upvoted 45 times

🗳️ 👤 **rsimham** 2 years, 9 months ago

I am thinking about Answer C, because events can be triggered by cloudwatch w/Glue metastore

upvoted 1 times

🗳️ 👤 **qwerty456** 2 years, 8 months ago

you can't schedule AWS Batch with CloudWatch

upvoted 4 times

🗳️ 👤 **kalyanvarma** 2 years, 7 months ago

We can schedule batch with cloud watch events.

upvoted 1 times

🗳️ 👤 **qwerty456** 2 years, 8 months ago

srr, looks like you can apart from Cron, the argument should be AWS Batch aren't SERVERLESS

upvoted 2 times

🗳️ 👤 **ComPah** 2 years, 9 months ago

if we use Flexible as key word ..Using Lambda might be a constraint

upvoted 4 times

🗳️ 👤 **cybe001** Highly Voted 2 years, 8 months ago

Answer is A. Lambda is the preferred way of implementing event-driven ETL job with S3, when new data arrives in S3, it notifies lambda which can start the ETL job.

upvoted 18 times

  **rb39** 1 year, 9 months ago

agree, event-driven means Lambda, CloudWatch alarms are just to trigger alarms based on log analysis.

upvoted 3 times

  **loict** Most Recent 9 months, 2 weeks ago

Selected Answer: A

A. YES - all integrated components

B. NO - missing a component to invoke the Lambda

C. NO - CloudWatch will not trigger when there is a new file to process

D. NO - CloudWatch will not trigger when there is a new file to process

upvoted 2 times

  **Mickey321** 10 months ago

Selected Answer: A

A for me

upvoted 1 times

  **kaike_reis** 11 months ago

Selected Answer: A



Note that the question asks for a serverless system. In this case, the letters B, C and D are wrong, as they bring options that are managed: AWS Batch (managed) and external Apache Hive (even more managed). For event-driven AWS ETL solutions that are serverless, activation through the Lambda function is recommended, so the correct alternative is Letter A. Note that CloudWatch Alarms only activates from log evaluation, which is not mentioned in the question.

upvoted 1 times

  **jackzhao** 1 year, 3 months ago

I will chose A, I think C & D is wrong, you can use Amazon CloudWatch Event to trigger lambda but not CloudWatch alarm.


upvoted 1 times

  **Valcilio** 1 year, 3 months ago

Selected Answer: A

Batch is more for configurations and other kinds of things by scheduling than event driven and batch data processing with ETL, the answer is A.

upvoted 1 times

  **Jeremy1** 1 year, 7 months ago

Selected Answer: A

Found this supporting A - Lambda used to trigger ETL job after crawler completes. The crawler starts on schedules or events (files arriving).

upvoted 1 times

  **Skychaser** 1 year, 11 months ago

Selected Answer: A

Based on Majority discussion

upvoted 2 times

  **exam887** 2 years, 1 month ago

Selected Answer: C

Quite confused between A&C since they all workable solution. In below AWS Blog, even mix the CloudWatch + Lambda to use the Glue. For key word event trigger, prefer CloudWatch

<https://aws.amazon.com/blogs/big-data/build-and-automate-a-serverless-data-lake-using-an-aws-glue-trigger-for-the-data-catalog-and-etl-jobs/>

<https://docs.aws.amazon.com/glue/latest/dg/automating-awsglue-with-cloudwatch-events.html>

upvoted 2 times

  **ZSun** 1 year, 2 months ago

cloudwatch and lambda function can work together to trigger event. But AWS batch cannot independently conduct ETL and require other service.

when it comes to ETL, glue is much easier choice than Batch

upvoted 1 times

  **VinceCar** 1 year, 7 months ago

Agreed. CloudWatch could trigger event to launch Lambda. Refer to: <https://docs.aws.amazon.com/lambda/latest/dg/services-cloudwatchevents.html>

upvoted 1 times

🗨️ 👤 **syu31svc** 2 years, 8 months ago

Answer is A 100%

upvoted 2 times

🗨️ 👤 **halfway** 2 years, 8 months ago

A is preferred. Lambda can trigger ETL pipelines: <https://aws.amazon.com/glue/>

upvoted 3 times

🗨️ 👤 **PRC** 2 years, 8 months ago

A is correct...Lambda is event driven and Glue is serverless as opposed to Hive

upvoted 4 times

A company's Machine Learning Specialist needs to improve the training speed of a time-series forecasting model using TensorFlow. The training is currently implemented on a single-GPU machine and takes approximately 23 hours to complete. The training needs to be run daily. The model accuracy is acceptable, but the company anticipates a continuous increase in the size of the training data and a need to update the model on an hourly, rather than a daily, basis. The company also wants to minimize coding effort and infrastructure changes. What should the Machine Learning Specialist do to the training solution to allow it to scale for future demand?

- A. Do not change the TensorFlow code. Change the machine to one with a more powerful GPU to speed up the training.
- B. Change the TensorFlow code to implement a Horovod distributed framework supported by Amazon SageMaker. Parallelize the training to as many machines as needed to achieve the business goals.
- C. Switch to using a built-in AWS SageMaker DeepAR model. Parallelize the training to as many machines as needed to achieve the business goals.
- D. Move the training to Amazon EMR and distribute the workload to as many machines as needed to achieve the business goals.

Suggested Answer: B

Community vote distribution

B (83%)

Other

 **JayK** Highly Voted 3 years, 9 months ago

the answer is B. using Hovord distribution results in less coding effort
upvoted 38 times

 **cybe001** Highly Voted 3 years, 9 months ago

Answer is B. "minimize coding effort and infrastructure changes" If we use DeepAR then the code and infra has to be changed to work with DeepAR.
upvoted 15 times

 **ninomfr64** Most Recent 1 year ago


Selected Answer: C

- A. NO, this will not address training dataset continuous increase
- B. NO, this will require code effort and infrastructure change
- C. YES, a built-in model ensure low code effort, so only infrastructure change needed*
- D. This will not work

* they say current model accuracy is acceptable, we doo expect good results with DeepAR as it allows to automatically pick among 5 different models what works best for the customer
upvoted 4 times

 **ninomfr64** 1 year ago

DeepAR doesn't pick among 5 models. However, I still think that switching to DeepAR can assure accuracy and minimize coding effort as the model is built-in
upvoted 2 times

 **VR10** 1 year, 4 months ago

A comes with minimum changes, but it wont scale.
B code changes are minimum but infrastructure still needs to be changed to achieve a distributed solution.
C. Is even more significant infra and code change.
D. wont work.
It is really subjective and tricky.
Could be A or B, depending on what change is considered "SMALL".
For scalability, B seems better. for quick win A could work.
I keep going back and forth.
upvoted 1 times

 **loict** 1 year, 9 months ago

Selected Answer: B

- A. NO - one time shot and not scalable
- B. YES - best practice

C. NO - DeepAR is for forecasting
D. NO - code will not benefit from parallelization without change
upvoted 4 times

🗳️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: B

option B
upvoted 2 times

🗳️ 👤 **kaike_reis** 1 year, 11 months ago

Selected Answer: B

Note that we want to increase training speed, minimize code and infrastructure modification effort on AWS. Letter A would only delay the problem and increase costs too much. The solution that best translates the problem would be Letter B: we would keep the code in tensorflow and use Horovod to make our training faster through parallelization. Letter D is too complex and would change the execution infrastructure a lot and Letter C would be too abrupt a turn as we would throw our model away.
upvoted 2 times

🗳️ 👤 **ZSun** 2 years, 2 months ago

A is better option even though B helps. Firstly, you only have One GPU, in this case distributed training Horovod doesn't help much; Secondly, the question is about minimize "coding effort" not minimize budget. adding distributed framework require much more coding, but increase gpu instance only require single click.
upvoted 1 times

🗳️ 👤 **Valcilio** 2 years, 3 months ago

Selected Answer: B

Horovod distribution is accepted by sagemaker, making easy to implement!
upvoted 1 times

🗳️ 👤 **Ajose0** 2 years, 4 months ago

Selected Answer: B

Hovord distribution will allow the Machine Learning Specialist to take advantage of Amazon SageMaker's built-in support for Horovod, which is a popular, open-source distributed deep learning framework.

Implementing Horovod in TensorFlow will allow the Specialist to parallelize the training across multiple GPUs or instances, which can significantly reduce the time it takes to train the model.

This will allow the company to meet its requirement to update the model on an hourly basis, and minimize coding effort and infrastructure changes as it leverages the existing TensorFlow code and infrastructure, along with the scalability and ease of use of Amazon SageMaker.
upvoted 4 times

🗳️ 👤 **joe3232** 2 years, 5 months ago

Are there a 23X differential between the weakest and strongest GPU in AWS? (and allow for future growth). I don't think so.
upvoted 1 times

🗳️ 👤 **vbal** 2 years, 5 months ago

Answer: C- built-in sagemaker DeepAR model. minimize coding & infra changes.
upvoted 1 times

🗳️ 👤 **cpal012** 2 years, 3 months ago

But they are happy with it - just want it to go faster. Not throw the whole thing out.
upvoted 1 times

🗳️ 👤 **KingGuo** 2 years, 11 months ago

Selected Answer: B

the answer is B. using Hovord distribution results in less coding effort
upvoted 2 times

🗳️ 👤 **John_Pongthorn** 3 years, 4 months ago

Selected Answer: A

Most likely, it is A because it is based on AWS technology, why we have to use open source

we exam AWS ML, the answer should be relevant to AWS technology inevitably
<https://aws.amazon.com/sagemaker/distributed-training/>
upvoted 1 times

🗨️ 👤 **cloud_trail** 3 years, 7 months ago

This one reminds me of an old saying by Yogi Berra: "When you come to a fork in the road, take it." If you see Horovod as an option in a question about scaling TF, take it. Answer is B.

upvoted 9 times

🗨️ 👤 **RaniaSayed** 3 years, 7 months ago

I Think it's B

<https://aws.amazon.com/blogs/machine-learning/launching-tensorflow-distributed-training-easily-with-horovod-or-parameter-servers-in-amazon-sagemaker/>

&

<https://aws.amazon.com/blogs/machine-learning/multi-gpu-and-distributed-training-using-horovod-in-amazon-sagemaker-pipe-mode/>

upvoted 5 times

🗨️ 👤 **harmanbirstudy** 3 years, 8 months ago

Seen similar question on udemy/whizlab , its always Horvord when Tensorflow needs scaling. ANWSER is B

upvoted 5 times

Which of the following metrics should a Machine Learning Specialist generally use to compare/evaluate machine learning classification models against each other?

- A. Recall
- B. Misclassification rate
- C. Mean absolute percentage error (MAPE)
- D. Area Under the ROC Curve (AUC)

Suggested Answer: D

Community vote distribution

D (100%)

🗳️ **DonaldCMLIN** Highly Voted 3 years, 3 months ago

RECALL IS ONE OF FACTOR IN CLASSIFY,

AUC IS MORE FACTORS TO COMPREHENSIVE JUDGEMENT

https://docs.aws.amazon.com/zh_tw/machine-learning/latest/dg/cross-validation.html

ANSWER MIGHT BE D.

upvoted 38 times

🗳️ **devsean** 3 years, 3 months ago

AUC is to determine hyperparams in a single model, not compare different models.

upvoted 6 times

🗳️ **DScore** 3 years, 2 months ago

Not might be, but should be D

upvoted 5 times

🗳️ **Ajose0** Highly Voted 1 year, 10 months ago

Selected Answer: D

Area Under the ROC Curve (AUC) is a commonly used metric to compare and evaluate machine learning classification models against each other. The AUC measures the model's ability to distinguish between positive and negative classes, and its performance across different classification thresholds. The AUC ranges from 0 to 1, with a score of 1 representing a perfect classifier and a score of 0.5 representing a classifier that is no better than random.

While recall is an important evaluation metric for classification models, it alone is not sufficient to compare and evaluate different models against each other. Recall measures the proportion of actual positive cases that are correctly identified as positive, but does not take into account the false positive rate.

upvoted 5 times

🗳️ **ccpmad** 1 year, 5 months ago

chatgpt answers, all your answers are from chatgpt

upvoted 2 times

🗳️ **AsusTuf** Most Recent 1 year, 2 months ago

why not C?

upvoted 1 times

🗳️ **Scrook** 7 months, 2 weeks ago

it's a classification problem, mape is for regression

upvoted 2 times

🗳️ **Mickey321** 1 year, 4 months ago

Selected Answer: D

option D

upvoted 1 times

🗨️ 👤 **Valcilio** 1 year, 9 months ago

Selected Answer: D

AUC is the best metric.

upvoted 1 times

🗨️ 👤 **cloud_trail** 3 years, 1 month ago

D. AUC is always used to compare ML classification models. The others can all be misleading. Consider the cases where classes are highly imbalanced. In those cases accuracy, misclassification rate and the like are useless. Recall is only useful if used in combination with precision or specificity, which what AUC does.

upvoted 4 times

🗨️ 👤 **harmanbirstudy** 3 years, 2 months ago

AUC/ROC work well with special case of Binary Classification not in general

upvoted 5 times

🗨️ 👤 **MohamedSharaf** 3 years, 1 month ago

AUC is to compare different models in terms of their separation power. 0.5 is useless as it's the diagonal line. 1 is perfect. I would go with F1 Score if it was an option. However, taking Recall only as a metric for comparing between models, would be misleading.

upvoted 4 times

🗨️ 👤 **harmanbirstudy** 3 years, 2 months ago

Its Accuracy, Precision, Recall and F1 score, there is no mention of AUC/ROC for comparing models in many articles, so ANSWER is A

upvoted 1 times

🗨️ 👤 **DavidRou** 1 year, 3 months ago

When you draw the ROC graph, you're considering True and False Positive Rate. The first one is also called Recall ;)

upvoted 1 times

🗨️ 👤 **Thai_Xuan** 3 years, 2 months ago

D. AUC is scale- and threshold-invariant, enabling it compare models.

<https://towardsdatascience.com/how-to-evaluate-a-classification-machine-learning-model-d81901d491b1>

upvoted 1 times

🗨️ 👤 **johnny_chick** 3 years, 2 months ago

Actually A, B and D seem to be correct

upvoted 1 times

🗨️ 👤 **deep_n** 3 years, 2 months ago

Probably D

<https://towardsdatascience.com/metrics-for-evaluating-machine-learning-classification-models-python-example-59b905e079a5>

upvoted 2 times

🗨️ 👤 **hughhughhugh** 3 years, 2 months ago

why not B?

upvoted 1 times

🗨️ 👤 **PRC** 3 years, 2 months ago

Answer should be D..ROC is used to determine the diagnostic capability of classification model varying on threshold

upvoted 3 times

🗨️ 👤 **Hypermasterd** 3 years, 2 months ago

Should be A. A is the only one that generally works for classification.

AUC only works with binary classification.

upvoted 4 times

🗨️ 👤 **oMARKOo** 3 years, 2 months ago

Actually AUC could be generalized for multi-class problem.

<https://www.datascienceblog.net/post/machine-learning/performance-measures-multi-class-problems/>

upvoted 1 times

🗨️ 👤 **sebas10** 3 years, 2 months ago

Could be, you mean in a multiclass classification problem. But in that context recall directly can't be compared because first you have to decide recall of what of the classes, in a 3 classes problem we have 3 recalls or you suppose a weighted recall or average recall ?. Do you think in that ?


upvoted 2 times

🗨️ 👤 **mrsimoes** 3 years, 2 months ago

Also in multi-class classification, if you follow an One-vs_Rest strategy you can still use AUC.

https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html#sphx-glr-auto-examples-model-selection-plot-roc-py

upvoted 1 times

  **stamarpadar** 3 years, 3 months ago

Correct Answer is D. Another benefit of using AUC is that it is classification-threshold-invariant like log loss.

<https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226>

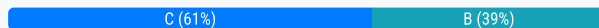
upvoted 3 times

A company is running a machine learning prediction service that generates 100 TB of predictions every day. A Machine Learning Specialist must generate a visualization of the daily precision-recall curve from the predictions, and forward a read-only version to the Business team. Which solution requires the LEAST coding effort?

- A. Run a daily Amazon EMR workflow to generate precision-recall data, and save the results in Amazon S3. Give the Business team read-only access to S3.
- B. Generate daily precision-recall data in Amazon QuickSight, and publish the results in a dashboard shared with the Business team.
- C. Run a daily Amazon EMR workflow to generate precision-recall data, and save the results in Amazon S3. Visualize the arrays in Amazon QuickSight, and publish them in a dashboard shared with the Business team.
- D. Generate daily precision-recall data in Amazon ES, and publish the results in a dashboard shared with the Business team.

Suggested Answer: C

Community vote distribution



rsimham Highly Voted 3 years, 9 months ago

Ans C is reasonable
upvoted 28 times

cloud_trail Highly Voted 3 years, 7 months ago

Agree with C. Quicksight cannot handle 100TB each day.
upvoted 6 times

MultiCloudIronMan Most Recent 8 months, 1 week ago

Selected Answer: C

Amazon QuickSight, particularly when using its SPICE (Super-fast, Parallel, In-memory Calculation Engine) feature, has specific data capacity limits. For the Enterprise Edition, SPICE can handle up to 1 billion rows or 1 TB per dataset¹. This means that while QuickSight is highly capable, handling 100 TB of data per day would exceed its current capacity limits.
upvoted 1 times

AMEJack 8 months, 3 weeks ago

Selected Answer: B

The limit of QuickSight for 1TB is soft limit which can be increased to unlimited number of TBs.
upvoted 1 times

Ali_Redha 1 year, 3 months ago

Ans C Because Quicksight Can't handle

100 TB even in Entiripise

Quotas for SPICE are as follows:

2,047 Unicode characters for each field

127 Unicode characters for each column name

2,000 columns for each file

1,000 files for each manifest

For Standard edition, 25 million (25,000,000) rows or 25 GB for each dataset

For Enterprise edition, 1 billion (1,000,000,000) rows or 1 TB for each dataset

<https://docs.aws.amazon.com/quicksight/latest/user/data-source-limits.html>

upvoted 2 times

🗳️ 👤 **VR10** 1 year, 4 months ago

QuickSight can handle large volumes of data for analytics and visualizations. Some key points:

QuickSight scales seamlessly from hundreds of megabytes to many terabytes of data without needing to manage infrastructure.

It uses an in-memory engine called SPICE to enable high performance analytics on large datasets.

so the choice is B

upvoted 1 times

🗳️ 👤 **kyuhuck** 1 year, 4 months ago

Selected Answer: B

B. Generate daily precision-recall data in Amazon QuickSight, and publish the results in a dashboard shared with the Business team.

This solution leverages QuickSight's managed service capabilities for both data processing and visualization, which should minimize the coding effort required to provide the Business team with the necessary insights. However, it's important to note that QuickSight's ability to calculate the precision-recall data depends on its support for the necessary statistical functions or the availability of such calculations in the dataset. If QuickSight cannot perform these calculations directly, option C might be necessary, despite the increased effort.

upvoted 1 times

🗳️ 👤 **Topg4u** 1 year, 4 months ago

The question does not ask for processing of 1Tb data. it asks for visuals/predications of that data. So B

upvoted 2 times

🗳️ 👤 **phdykd** 1 year, 5 months ago

C.

Considering the large volume of data (100 TB daily), Option C seems to be the most appropriate solution

upvoted 1 times

🗳️ 👤 **iskorini** 1 year, 7 months ago

Selected Answer: C

B it's not correct because of 100tb data size.

C is the answer: <https://docs.aws.amazon.com/quicksight/latest/user/data-source-limits.html>

upvoted 2 times

🗳️ 👤 **Snape** 1 year, 8 months ago

Selected Answer: C

ANs c is correct

upvoted 1 times

🗳️ 👤 **loict** 1 year, 9 months ago

Selected Answer: C

A. NO - we want a dashboard for business

B. NO - 100TB is very large, it will not fit in memory (1TB max for SPICE dataset) or return within the 2min limit if delegated to a DB (<https://docs.aws.amazon.com/quicksight/latest/user/data-source-limits.html>)

C. YES - best combination; EMR can distribute the computation of precision-recall for each slice of data

D. NO - ES cannot help to generate precision-recall

upvoted 1 times

🗳️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: B

although C is tempting but goes with B due to less effort

upvoted 1 times

🗳️ 👤 **teka112233** 1 year, 10 months ago

it is not about the least effort only, since the least effort solution here will not get your job done, look at the quick sight max data it can deal with when it compared to EMR which is built to deal with Big data.

upvoted 1 times

🗳️ 👤 **teka112233** 1 year, 10 months ago

Selected Answer: C

using quick sight for creation of the precision recall with 100 TB every day can't be done since the max size for quick sight to deal with is :
For Standard edition, 25 million (25,000,000) rows or 25 GB for each dataset
For Enterprise edition, 1 billion (1,000,000,000) rows or 1 TB for each dataset
acc to AWS documentation :
<https://docs.aws.amazon.com/quicksight/latest/user/data-source-limits.html>
but we can do it with EMR and latterly use quick sight to visualize the results
upvoted 2 times

🗨️ 👤 **kaike_reis** 1 year, 11 months ago

Selected Answer: C

Looking at the QuickSight documentation: it has a limit of 1 TB per dataset. So it's necessary a previous layer. Letter C is the correct one.
upvoted 1 times

🗨️ 👤 **ADVIT** 2 years ago

It's 100TB daily, need EMR to reduce, option C is correct.
upvoted 1 times

🗨️ 👤 **petervu** 2 years ago

Selected Answer: C

Quicksight can handle maximum 1TB data set only. We have 100TB data set so we need EMR.
<https://docs.aws.amazon.com/quicksight/latest/user/data-source-limits.html>
upvoted 3 times

A Machine Learning Specialist is preparing data for training on Amazon SageMaker. The Specialist is using one of the SageMaker built-in algorithms for the training. The dataset is stored in .CSV format and is transformed into a numpy.array, which appears to be negatively affecting the speed of the training.

What should the Specialist do to optimize the data for training on SageMaker?

- A. Use the SageMaker batch transform feature to transform the training data into a DataFrame.
- B. Use AWS Glue to compress the data into the Apache Parquet format.
- C. Transform the dataset into the RecordIO protobuf format.
- D. Use the SageMaker hyperparameter optimization feature to automatically optimize the data.

Suggested Answer: C


Community vote distribution

C (100%)

 **rsimham** Highly Voted 2 years, 9 months ago

C is okay

upvoted 19 times

 **stamarpadar** Highly Voted 2 years, 8 months ago

Answer is C.

Most Amazon SageMaker algorithms work best when you use the optimized protobuf recordIO format for the training data.

<https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html>


upvoted 16 times

 **Mickey321** Most Recent 10 months ago

Selected Answer: C

option C

upvoted 1 times


 **Ajose0** 1 year, 4 months ago

Selected Answer: C

The Specialist should transform the dataset into the RecordIO protobuf format. This format is optimized for use with SageMaker and has been shown to improve the speed and efficiency of training algorithms.

Using the RecordIO protobuf format is a best practice for preparing data for use with Amazon SageMaker, and it is specifically recommended for use with the built-in algorithms.


upvoted 1 times

 **Jeremy1** 1 year, 7 months ago

Selected Answer: C

I would assume the issue is the transformation. It can be nasty slow between pandas / csv / numpy. Go to protobuf.

upvoted 1 times

 **C10ud9** 2 years, 7 months ago

C is the best

upvoted 5 times

 **PRC** 2 years, 8 months ago

Agree with C

upvoted 6 times

A Machine Learning Specialist is required to build a supervised image-recognition model to identify a cat. The ML Specialist performs some tests and records the following results for a neural network-based image classifier:

Total number of images available = 1,000

Test set images = 100 (constant test set)

The ML Specialist notices that, in over 75% of the misclassified images, the cats were held upside down by their owners.

Which techniques can be used by the ML Specialist to improve this specific test error?

- A. Increase the training data by adding variation in rotation for training images.
- B. Increase the number of epochs for model training
- C. Increase the number of layers for the neural network.
- D. Increase the dropout rate for the second-to-last layer.

Suggested Answer: B

Community vote distribution

A (100%)

 **DonaldCMLIN**  3 years, 3 months ago

NO CORRECT TRAINING DATA, MORE WORKS JUST WASTE TIME.

ONE OF THE REASONS FOR POOR ACCURACY COULD BE INSUFFICIENT DATA. THIS CAN BE OVERCOME BY IMAGE AUGMENTATION. IMAGE AUGMENTATION IS A TECHNIQUE OF INCREASING THE DATASET SIZE BY PROCESSING (MIRRORING, FLIPPING, ROTATING, INCREASING/DECREASING BRIGHTNESS, CONTRAST, COLOR) THE IMAGES.

[HTTPS://MEDIUM.COM/DATADRIVENINVESTOR/AUTO-MODEL-TUNING-FOR-KERAS-ON-AMAZON-SAGEMAKER-PLANT-SEEDLING-DATASET-7B591334501E](https://medium.com/datadriveninvestor/auto-model-tuning-for-keras-on-amazon-sagemaker-plant-seedling-dataset-7b591334501e)

ANSWER A. ADD MORE TRAINING DATA FOR ROTATION IMAGES COULD BE A WAY TO DEAL WITH ISSUE

upvoted 63 times

 **Jeremy1** 2 years, 1 month ago

Donald, your caps lock is on.

upvoted 13 times

 **kaike_reis** 1 year, 4 months ago


Okay, was funny

upvoted 1 times

 **Nadia0012** 1 year, 9 months ago

LOL :D

upvoted 1 times

 **ccpmad** 1 year, 5 months ago

is it possible no using MAYUS? it is annoying

upvoted 1 times

 **tap123** 3 years, 2 months ago

The key phrase might be "constant test set", so you can't increase training set by shrinking the size of test set. Thus the only feasible choice is to increase training time by increasing the number of epochs => answer B.

upvoted 2 times

 **mawsman** 3 years, 2 months ago

The problem is images are upside down and misclassified. If right side up then the model would classify correctly. This can only be fixed by rotating not by trying to recognise upside down cat more times.

upvoted 3 times

 **Urban_Life** 3 years, 2 months ago

What's your answer B?

upvoted 1 times

🗳️ 👤 **VB** 3 years, 2 months ago

A . Increase the training data by adding variation in rotation for training images.

It never says to move the images from Test data set (because it is constant)... only variations are added to the images..so, A is correct.

upvoted 1 times

🗳️ 👤 **rsimham** 3 years, 3 months ago

agree with A

upvoted 9 times

🗳️ 👤 **phdykd** Most Recent 11 months, 3 weeks ago

A is answer

upvoted 1 times

🗳️ 👤 **Kensev** 1 year ago

Selected Answer: A

Data Augmentation would fix the missing conditional data

upvoted 2 times

🗳️ 👤 **cgsoft** 1 year, 1 month ago

Selected Answer: A

ChatGPT says the answer is A. Trust a model to answer an ML question correctly! ;)

upvoted 1 times

🗳️ 👤 **AmeeraM** 1 year, 2 months ago

Selected Answer: A

how come more epochs it better than augmentation?

upvoted 1 times

🗳️ 👤 **Mickey321** 1 year, 4 months ago

Selected Answer: A

option A

upvoted 1 times

🗳️ 👤 **kaike_reis** 1 year, 4 months ago

Selected Answer: A

The question is clear and the answer is clear as well

upvoted 1 times

🗳️ 👤 **nilmans** 1 year, 6 months ago

Selected Answer: A

should be A

upvoted 1 times

🗳️ 👤 **earthMover** 1 year, 7 months ago

Selected Answer: A

More epochs is not a good approach to fundamental data issues

upvoted 2 times

🗳️ 👤 **oso0348** 1 year, 9 months ago

Selected Answer: A

the Specialist can apply data augmentation techniques to increase the training data by adding variation in rotation for training images. This technique will allow the model to learn to recognize cats in various orientations, including upside down.

upvoted 1 times

🗳️ 👤 **Ajose0** 1 year, 10 months ago

Selected Answer: A

Adding more variation in rotation to the training data can help the model to learn how to classify cats in different orientations, including when they are held upside down. This can improve the model's ability to identify cats in this position and reduce the misclassification rate for images in which the cats are upside down.

By adding more rotation to the training data, the model can be trained to generalize better to new images, including those with cats in different orientations. This can help to reduce overfitting and improve the model's overall performance.

upvoted 1 times

🗨️ 👤 **Tomatoteacher** 1 year, 11 months ago

Selected Answer: A

Only logical answer 100% A.

upvoted 1 times

🗨️ 👤 **Jeremy1** 2 years, 1 month ago

Selected Answer: A

More data is a good answer. A

upvoted 1 times

🗨️ 👤 **ryuhei** 2 years, 3 months ago

Selected Answer: A

Answer is "A"

upvoted 1 times

🗨️ 👤 **Morsa** 2 years, 5 months ago

Answer is A

upvoted 1 times

🗨️ 👤 **ovokpus** 2 years, 6 months ago

Selected Answer: A

This is a clear case of Data Augumentation solution.

upvoted 1 times

🗨️ 👤 **yc1005** 2 years, 6 months ago

Selected Answer: A

Common step in CNN, Image augmentation. A.

upvoted 1 times

A Machine Learning Specialist needs to be able to ingest streaming data and store it in Apache Parquet files for exploration and analysis. Which of the following services would both ingest and store this data in the correct format?

- A. AWS DMS
- B. Amazon Kinesis Data Streams
- C. Amazon Kinesis Data Firehose
- D. Amazon Kinesis Data Analytics

Suggested Answer: C

Community vote distribution

C (100%)

🗨️ **JayK** Highly Voted 3 years, 9 months ago

the answer is C. as the main point of the question is data transformation to Parquet format which is done by Kinesis Data Firehose not Data Stream. Coming to the data store the data store in Kinesis Data Stream is only for couple of days so it does not serve the purpose here
upvoted 52 times

🗨️ **shammous** 10 months, 2 weeks ago

The storage part will be taken care of by S3 anyway. Firehose would just transform to Parquet on the fly.
upvoted 1 times

🗨️ **eganilovic** Highly Voted 3 years, 7 months ago

Firehose
upvoted 5 times

🗨️ **earthMover** Most Recent 2 years, 1 month ago

Not sure Firehose can store the data Data Stream can store the data. Someone please explain the answer
upvoted 1 times

🗨️ **kaike_reis** 1 year, 11 months ago

Firehose is to Store the data. Stream requires other service to do that.
upvoted 1 times

🗨️ **GOSD** 2 years, 1 month ago

Kinesis Data Streams can Store for up to 365 days, While Firehouse sends it to S3. Which is correct?
upvoted 1 times

🗨️ **Valcilio** 2 years, 3 months ago

Selected Answer: C

Firehose can do it if the data is in JSON or ORC format initially!
upvoted 2 times

🗨️ **DS2021** 2 years, 4 months ago

It should be KDS
upvoted 1 times

🗨️ **Ajose0** 2 years, 4 months ago

Selected Answer: C

Amazon Kinesis Data Firehose is a fully managed service that can automatically load streaming data into data stores and analytics tools.

It can ingest real-time streaming data such as application logs, website clickstreams, and IoT telemetry data, and then store it in the correct format, such as Apache Parquet files, for exploration and analysis.

This makes it a suitable option for the requirement described in the question.

upvoted 1 times

🗨️ **Thai_Xuan** 3 years, 7 months ago

B



<https://github.com/ravsau/aws-exam-prep/issues/10>

upvoted 2 times

  **weslleylc** 3 years, 8 months ago

B) Only Amazon Kinesis Data Streams can store and ingest data. We don't need to apply any transformation; the question asks to ingest and store data in Apache Parquet format. There is no assumption that the data coming in a different format than parquet.

upvoted 3 times

  **joe3232** 2 years, 5 months ago

KDS cant store to s3

<https://stackoverflow.com/questions/66097886/writing-to-s3-via-kinesis-stream-or-firehose>

upvoted 1 times

  **In** 3 years, 8 months ago

It is C with no doubt

https://aws.amazon.com/about-aws/whats-new/2018/05/stream_real_time_data_in_apache_parquet_or_orc_format_using_firehose/

upvoted 5 times

  **GeeBeeEI** 3 years, 8 months ago

It appears all agree that the answer is between Firehose and Analytics. Kinesis Firehose is used for ingestion. Both firehose and analytics can store, only firehose can ingest. <https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html> shows firehose can store parquet to S3

upvoted 3 times

  **GeeBeeEI** 3 years, 8 months ago



It appears all agree that the answer is between Firehose and Analytics. Data Streams handle stuff like event data, clickstream etc. Its not interested in special format, the focus is speed. The question did not talk of transformation, only ingestion. Kinesis Firehose is used for ingestion. Both firehose and analytics can store, only firehose can ingest. <https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html> shows firehose can store parquet to S3

upvoted 1 times

  **Urban_Life** 3 years, 8 months ago



Think just like this – batch process Glue ETL and Streaming process Firehose ETLcovert to parquet or any other format.

upvoted 1 times

  **CMMC** 3 years, 8 months ago

C for Firehose

upvoted 2 times

  **Erso** 3 years, 9 months ago

Just in case https://acloud.guru/forums/aws-certified-big-data-specialty/discussion/-KhI3MgPEo-FY5rfgl3J/what_is_difference_between_kin

upvoted 2 times

  **BigEv** 3 years, 9 months ago

Amazon Kinesis Data Firehose can convert the format of your input data from JSON to Apache Parquet or Apache ORC before storing the data in Amazon S3.

https://github.com/awsdocs/amazon-kinesis-data-firehose-developer-guide/blob/master/doc_source/record-format-conversion.md

upvoted 3 times

  **rsimham** 3 years, 9 months ago

I would go with B. Kinesis data streams stores data, while Firehose not.

upvoted 3 times

  **cloud_trail** 3 years, 7 months ago

It's the other way around. Firehouses stores data; data streams does not.

upvoted 1 times

A data scientist has explored and sanitized a dataset in preparation for the modeling phase of a supervised learning task. The statistical dispersion can vary widely between features, sometimes by several orders of magnitude. Before moving on to the modeling phase, the data scientist wants to ensure that the prediction performance on the production data is as accurate as possible.

Which sequence of steps should the data scientist take to meet these requirements?

- A. Apply random sampling to the dataset. Then split the dataset into training, validation, and test sets.
- B. Split the dataset into training, validation, and test sets. Then rescale the training set and apply the same scaling to the validation and test sets.
- C. Rescale the dataset. Then split the dataset into training, validation, and test sets.
- D. Split the dataset into training, validation, and test sets. Then rescale the training set, the validation set, and the test set independently.

Suggested Answer: D

Reference:

<https://www.kdnuggets.com/2018/12/six-steps-master-machine-learning-data-preparation.html>

Community vote distribution

B (76%)

C (24%)

  **cron0001** Highly Voted 3 years, 2 months ago

Selected Answer: C

C would be my answer here. Rescaling each set independently could lead to strange skews. Training set, Test set and Evaluation set should be on the same scale

upvoted 17 times

  **GiyeonShin** 2 years, 6 months ago

You're right. test set and val set should be rescaled on the same scale.

But the scale value should be extracted by only statistical value from training data.

I think C means that the rescaling stage is affected by the values from the whole data (with val, test set)

So, I think B is correct

upvoted 9 times

  **masoa3b** Highly Voted 2 years, 8 months ago

Selected Answer: B

<https://stackoverflow.com/questions/49444262/normalize-data-before-or-after-split-of-training-and-testing-data>

C also leads to data leakage. You are using the test data to scale everything. So part of the data in the test set is used to scale for when you build the model on the training and check against the validation set.

upvoted 17 times

  **ML_2** Most Recent 10 months, 2 weeks ago

Selected Answer: B

If you Rescale all the data first you are going to do data leakage by showing all the variance of data with in training. The rescaling needs to be after splitting the data and not before it

upvoted 1 times



  **Denise123** 1 year, 4 months ago

Selected Answer: B

The best practice is --> to split the dataset into training, validation, and test sets first, and then rescale the training set and apply the SAME scaling to the validation and test sets. This ensures that the scaling parameters (e.g., mean and standard deviation for standardization or min and max values for min-max scaling) are calculated only based on the training set to prevent data leakage and maintain the integrity of the evaluation process.

By following this approach, you prevent information from the validation and test sets from influencing the scaling parameters, which could lead to data leakage and overestimation of model performance. Keeping the scaling consistent across all subsets ensures a fair evaluation of the model's generalization performance on new, unseen data.

upvoted 5 times

  **phdykd** 1 year, 5 months ago

Answer is B.

The other options have shortcomings:

A: Random sampling is a good practice, but it doesn't address the issue of feature scaling. Also, rescaling should occur after splitting the data.

C: Rescaling the entire dataset before splitting could lead to data leakage, where information from the validation/test sets inadvertently influences the training process.

D: Rescaling the sets independently would lead to inconsistencies in scale across the training, validation, and test sets, which could negatively impact model performance and evaluation.

upvoted 2 times

🗳️ 👤 **Sukhi4fornet** 1 year, 6 months ago

OPTION C. Rescale the dataset. Then split the dataset into training, validation, and test sets.

Explanation:

Rescaling the dataset:

This is the first step to address the varying statistical dispersion among features. By rescaling, you ensure that all features are on a similar scale, which is important for many machine learning algorithms.

Splitting into training, validation, and test sets:

After rescaling, the dataset is split into training, validation, and test sets. This ensures that the model is trained on one set, validated on another set, and tested on a third set. This separation helps evaluate the model's performance on unseen data.

Option C ensures that the rescaling is applied before splitting the data, ensuring consistency in the scaling across different sets. This approach prevents data leakage and provides a more accurate representation of how the model will perform on new, unseen data.

upvoted 1 times

🗳️ 👤 **akgarg00** 1 year, 6 months ago

Selected Answer: B

Validation and test set should be scaled as per parameters used for scaling of training set. Independent scaling of test set would mean that drift of model in production will be way quicker and is not recommended in data science

upvoted 1 times

🗳️ 👤 **elvin_ml_qayiran25091992razor** 1 year, 7 months ago

Selected Answer: B

B is correct, scale on train and apply the others. prevent to data leakage

upvoted 1 times

🗳️ 👤 **akgarg00** 1 year, 7 months ago

Selected Answer: B

Answer B, C is not a good data science practise.

upvoted 1 times

🗳️ 👤 **DimLam** 1 year, 8 months ago

Selected Answer: B

We need firstly split the data to avoid data leakage from test/eval sets, then rescale data in all sets using statistics from training set

upvoted 1 times

🗳️ 👤 **DavidRou** 1 year, 9 months ago

Selected Answer: B

I think the right answer here is B. We need to split the dataset into Training, Validation and Test set. Then we can only scale (by using some technique) data contained in the Training set. Data that belong to Validation and Test set must be scaled by using the parameters used on the training.

For example, if we want to apply a standardization, we can do that only on the Training set as we should not be allowed to use mean and standard deviation computed on Validation/Test set. We must act as we don't own those data!

upvoted 2 times

🗳️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: B

option B

upvoted 1 times

🗨️ 👤 **kaike_reis** 1 year, 11 months ago

Data Science 101:

(A) Given the question, doesn't solve the magnitude problem.

(B) Correct

(C) Data Leakage

(D) It's not correct, still data leakage.

upvoted 1 times

🗨️ 👤 **gusta_dantas** 1 year, 11 months ago

Tricky question, but, D, definitely!

B: You can't apply the same scaling to the validation and test sets 'cause you may suffer data leakage!

C: You shouldn't rescale the whole dataset then split into training, validation and test, it's not a good practice and may suffer data leakage as well.

D: You're first splitting the whole dataset and applying rescaling individually, preventing any data leakage and each set is rescaled based in your own statistics.

upvoted 1 times

🗨️ 👤 **DavidRou** 1 year, 9 months ago

Theoretically, you should not have Test set data at Training time (when you're doing the scaling), so how do you think to do that?

What if you will not have an entire Test set, but you will receive each new row at a time?

upvoted 1 times

🗨️ 👤 **kaike_reis** 1 year, 11 months ago

but you are leaking information from validation samples between themselves.

upvoted 1 times

🗨️ 👤 **JK1977** 2 years, 1 month ago

Selected Answer: B

From Bing chat (and it makes complete sense)

"Based on the search results, I think the best sequence of steps for the data scientist to take is B. Split the dataset into training, validation, and test sets. Then rescale the training set and apply the same scaling to the validation and test sets.

This sequence of steps ensures that the data scientist can evaluate the model performance on different subsets of data that have not been used for training or tuning. It also ensures that the data scientist can rescale the features to have a common scale without introducing any data leakage from the validation or test sets. Rescaling the features can help improve the accuracy of some machine learning algorithms that are sensitive to the magnitude or distribution of the data, such as distance-based methods or gradient-based methods 1.

upvoted 3 times

🗨️ 👤 **tommct** 2 years, 1 month ago

Selected Answer: B

You want to measure how the model performs on new data. Scaling with the test set is a no-no.

upvoted 1 times

🗨️ 👤 **GOSD** 2 years, 1 month ago

B or D, I dont understand the semantics of "independently" and the effect it would have. It's most def not done before because of data leakage.

<https://www.linkedin.com/pulse/feature-scaling-dataset-splitting-arnab-mukherjee/>

upvoted 1 times

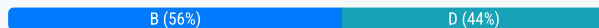
A Machine Learning Specialist is assigned a TensorFlow project using Amazon SageMaker for training, and needs to continue working for an extended period with no Wi-Fi access.

Which approach should the Specialist use to continue working?

- A. Install Python 3 and boto3 on their laptop and continue the code development using that environment.
- B. Download the TensorFlow Docker container used in Amazon SageMaker from GitHub to their local environment, and use the Amazon SageMaker Python SDK to test the code.
- C. Download TensorFlow from tensorflow.org to emulate the TensorFlow kernel in the SageMaker environment.
- D. Download the SageMaker notebook to their local environment, then install Jupyter Notebooks on their laptop and continue the development in a local notebook.

Suggested Answer: B

Community vote distribution



DonaldCMLIN Highly Voted 3 years, 2 months ago

ANSWER B.

YOU COULD INSTALL DOCKER-COMPOSE (AND NVIDIA-DOCKER IF TRAINING WITH A GPU) FOR LOCAL TRAINING

[HTTPS://SAGEMAKER.READTHEDOCS.IO/EN/STABLE/OVERVIEW.HTML#LOCAL-MODE](https://sagemaker.readthedocs.io/en/stable/overview.html#local-mode)

[HTTPS://GITHUB.COM/AWSLABS/AMAZON-SAGEMAKER-EXAMPLES/BLOB/MASTER/SAGEMAKER-PYTHON-](https://github.com/aws-labs/amazon-sagemaker-examples/blob/master/sagemaker-python-sdk/tensorflow_distributed_mnist/tensorflow_local_mode_mnist.ipynb)

[SDK/TENSORFLOW_DISTRIBUTED_MNIST/TENSORFLOW_LOCAL_MODE_MNIST.IPYNB](https://github.com/aws-labs/amazon-sagemaker-examples/blob/master/sagemaker-python-sdk/tensorflow_distributed_mnist/tensorflow_local_mode_mnist.ipynb)

upvoted 42 times

sqavi 1 year, 10 months ago

None of these links are working

upvoted 3 times

VB Highly Voted 3 years, 2 months ago

<https://aws.amazon.com/blogs/machine-learning/use-the-amazon-sagemaker-local-mode-to-train-on-your-notebook-instance/>

B

upvoted 10 times

ef12052 Most Recent 3 months ago

Selected Answer: B

<https://aws.amazon.com/blogs/machine-learning/use-the-amazon-sagemaker-local-mode-to-train-on-your-notebook-instance/>

stop using gpt....

upvoted 1 times

sfwewv 4 months, 2 weeks ago

Selected Answer: D

GPT said SageMaker Python SDK is less suitable for offline

upvoted 1 times

rookiee1111 8 months ago

Selected Answer: B

Correction it will be B, while D is possible, it cannot exactly mimic the sagemaker env, with docker all the configuration and libs will be available to the user which would be an ideal working setup for the DS to work with.

upvoted 1 times

rookiee1111 8 months ago

Selected Answer: D

You can easily download the notebook instance, and work locally using jupyter notebook configured on your laptop which is one the advantages of using sagemaker, and that is what Amazon also promotes imo.

upvoted 2 times

🗳️ 👤 **ArchMelody** 10 months, 1 week ago

Selected Answer: D

Both Amazon Q (AWS Expert) and ChatGPT insist on D. Plus all the links that I see here about Docker/Git and stuff, they either not working or deprecated so far. Not to mention their complexity to my eyes.

Thus, I will go for D.

upvoted 3 times

🗳️ 👤 **rav009** 1 year, 2 months ago

Selected Answer: B

the local mode of sagemaker SDK:

<https://sagemaker.readthedocs.io/en/stable/overview.html#local-mode>

B

upvoted 2 times

🗳️ 👤 **Mickey321** 1 year, 4 months ago

Selected Answer: B

Option B

upvoted 1 times

🗳️ 👤 **ADVIT** 1 year, 6 months ago

B,

<https://github.com/aws/sagemaker-tensorflow-serving-container>

upvoted 1 times

🗳️ 👤 **Valcilio** 1 year, 9 months ago

Selected Answer: B

It's B

upvoted 1 times

🗳️ 👤 **SriAkula** 2 years, 9 months ago

Answer : D

upvoted 2 times

🗳️ 👤 **noblade** 3 years ago

why not D?

upvoted 1 times

🗳️ 👤 **AddiWei** 2 years, 10 months ago

My assumption is that D there is no way to test the code. You need the Sagemaker SDK in order to utilize dockerized container of Tensorflow from Sagemaker is my best guess.

upvoted 2 times

🗳️ 👤 **Tomatoteacher** 1 year, 11 months ago

Cannot be D. If you used Jupyter notebook, you are unable to use it without internet access.

upvoted 1 times

🗳️ 👤 **rookiee1111** 8 months ago

That is incorrect, once jupyter notebook is configured you can use it offline.

upvoted 1 times

🗳️ 👤 **CMMC** 3 years, 2 months ago

Agreed for B

upvoted 4 times

A Machine Learning Specialist is working with a large cybersecurity company that manages security events in real time for companies around the world. The cybersecurity company wants to design a solution that will allow it to use machine learning to score malicious events as anomalies on the data as it is being ingested. The company also wants be able to save the results in its data lake for later processing and analysis.

What is the MOST efficient way to accomplish these tasks?

- A. Ingest the data using Amazon Kinesis Data Firehose, and use Amazon Kinesis Data Analytics Random Cut Forest (RCF) for anomaly detection. Then use Kinesis Data Firehose to stream the results to Amazon S3.
- B. Ingest the data into Apache Spark Streaming using Amazon EMR, and use Spark MLlib with k-means to perform anomaly detection. Then store the results in an Apache Hadoop Distributed File System (HDFS) using Amazon EMR with a replication factor of three as the data lake.
- C. Ingest the data and store it in Amazon S3. Use AWS Batch along with the AWS Deep Learning AMIs to train a k-means model using TensorFlow on the data in Amazon S3.
- D. Ingest the data and store it in Amazon S3. Have an AWS Glue job that is triggered on demand transform the new data. Then use the built-in Random Cut Forest (RCF) model within Amazon SageMaker to detect anomalies in the data.

Suggested Answer: B

Community vote distribution

A (95%)

5%

  **DonaldCMLIN** Highly Voted 3 years, 3 months ago

I WOULD LIKE TO CHOOSE ANSWER A.

<https://aws.amazon.com/tw/blogs/machine-learning/use-the-built-in-amazon-sagemaker-random-cut-forest-algorithm-for-anomaly-detection/>
upvoted 60 times

  **hamimelon** 2 years ago

Donald, do you know your CAPS LOCK has been on the whole time?

upvoted 15 times

  **Nadia0012** 1 year, 9 months ago

I know why his caps lock has been on :D to enter the "I am not robot" code easier :D

upvoted 4 times

  **ccpmad** 1 year, 5 months ago



yes, but it works with minus also...

upvoted 1 times

  **JayK** Highly Voted 3 years, 2 months ago

Answer is A. As the word anamoly talks about Random Cut Forest in the exam and that can be done in a cost effective manner using Kinesis Data Analytics

upvoted 15 times

  **Shakespeare** 6 months, 2 weeks ago

I think it would have been more accurate if the options were kinetic data stream -> kinesis data analytics -> kinesis firehose -> S3

upvoted 2 times

  **sacilm** Most Recent 8 months, 2 weeks ago

The question says REAL TIME events doesn't that eliminate Data Firehose as it is technically NEAR real time but not real time like Data Stream?

Though Random Cut Forest seems like the best option for anomaly detection. I'm torn between A and B


upvoted 1 times

  **vkajoria** 9 months ago

Selected Answer: A

Kinesis Firehose and Data Analytics with random cut forest should do it.

upvoted 1 times

  **phdykd** 11 months, 3 weeks ago

A.

Based on these considerations, Option A is the most efficient way to accomplish the tasks. It provides a seamless, real-time data ingestion and

processing pipeline, leverages machine learning for anomaly detection, and efficiently stores data in a data lake, meeting all the key requirements of the cybersecurity company.

upvoted 1 times

🗳️ 👤 **elvin_ml_qayiran25091992razor** 1 year, 1 month ago

Selected Answer: A

ONLY A

upvoted 1 times

🗳️ 👤 **sonoluminescence** 1 year, 2 months ago

Selected Answer: A

B not as efficient for real-time processing and storing results as using Kinesis services.

upvoted 2 times

🗳️ 👤 **DimLam** 1 year, 2 months ago

Selected Answer: B

At least B is a possible solution, but A will not work as KDF doesn't support KDA as a destination service

<https://docs.aws.amazon.com/firehose/latest/dev/create-name.html> . In my opinion, KDF should always be the latest Kinesis Service in a streaming pipeline

upvoted 1 times

🗳️ 👤 **Dun6** 1 year, 1 month ago

KDF does support KDA as destination

upvoted 1 times

🗳️ 👤 **AmeeraM** 1 year, 2 months ago

Selected Answer: A

A has all the required steps

upvoted 1 times

🗳️ 👤 **loict** 1 year, 3 months ago

Selected Answer: A

A. YES - Firehose can pipe into KDA, and KDA supports RCF

B. NO - RCF best for anomaly detection

C. NO - no need for intermediary S3 storage

D. NO - no need for intermediary S3 storage

upvoted 1 times

🗳️ 👤 **Mickey321** 1 year, 4 months ago

Selected Answer: A

option A

upvoted 1 times

🗳️ 👤 **kaike_reis** 1 year, 4 months ago

Selected Answer: A

A is the correct. One tip for the exam: When you see Data Streaming, possibly the solution should contains a Kinesis Service. B is too much complex!

upvoted 3 times

🗳️ 👤 **nilmans** 1 year, 6 months ago

Selected Answer: A

Makes sense to select A here.

upvoted 1 times

🗳️ 👤 **earthMover** 1 year, 7 months ago

Selected Answer: A

I strongly believe A is the right answer. At a minimum there should be some justification provided for your answer.

upvoted 1 times

🗳️ 👤 **Ajose0** 1 year, 10 months ago

Selected Answer: A

Amazon Kinesis Data Firehose is a fully managed service for streaming real-time data to Amazon S3 and can handle the ingestion of large amounts of data in real time. Kinesis Data Analytics Random Cut Forest (RCF) is a fully managed service that can be used to perform anomaly detection on streaming data, making it well suited for this use case. The results of the anomaly detection can then be streamed to Amazon S3 using Kinesis Data Firehose, providing a scalable and cost-effective data lake for later processing and analysis.

upvoted 2 times

🗨️ 👤 **DimLam** 1 year, 2 months ago

The problem with A, is that there is that KDF doesn't support KDA as a destination service

<https://docs.aws.amazon.com/firehose/latest/dev/create-name.html> . In my opinion, KDF should always be the latest Kinesis Service in a streaming pipeline

upvoted 1 times

🗨️ 👤 **OssamaAbdelatif** 2 years, 1 month ago

I would select A

upvoted 1 times

🗨️ 👤 **ovokpus** 2 years, 6 months ago

Selected Answer: A

B is too resource intensive for that use case. I choose A, but I think the data should be better ingested using Kinesis streams

upvoted 3 times

A Data Scientist wants to gain real-time insights into a data stream of GZIP files.
Which solution would allow the use of SQL to query the stream with the LEAST latency?

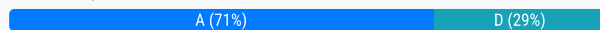
- A. Amazon Kinesis Data Analytics with an AWS Lambda function to transform the data.
- B. AWS Glue with a custom ETL script to transform the data.
- C. An Amazon Kinesis Client Library to transform the data and save it to an Amazon ES cluster.
- D. Amazon Kinesis Data Firehose to transform the data and put it into an Amazon S3 bucket.

Suggested Answer: A

Reference:

<https://aws.amazon.com/big-data/real-time-analytics-featured-partners/>

Community vote distribution



cybe001 Highly Voted 3 years, 2 months ago

A is correct. Kinesis Data Analytics can use lambda to convert GZIP and can run SQL on the converted data.

<https://aws.amazon.com/about-aws/whats-new/2017/10/amazon-kinesis-analytics-can-now-pre-process-data-prior-to-running-sql-queries/>
upvoted 44 times

VB Highly Voted 3 years, 2 months ago

A is correct:

<https://aws.amazon.com/about-aws/whats-new/2017/10/amazon-kinesis-analytics-can-now-pre-process-data-prior-to-running-sql-queries/>

"To get started, simply select an AWS Lambda function from the Kinesis Analytics application source page in the AWS Management console. Your Kinesis Analytics application will automatically process your raw data records using the Lambda function, and send transformed data to your SQL code for further processing.

Kinesis Analytics provides Lambda blueprints for common use cases like converting GZIP

..."

upvoted 17 times

ef12052 Most Recent 3 months ago

Selected Answer: A

Use Amazon Kinesis Data Analytics if you need SQL-based processing and advanced analytics capabilities for streaming data.

Use Amazon Kinesis Data Firehose if your primary requirement is to deliver, transform, and load streaming data into various AWS destinations with simplified configurations, but not for SQL-based processing.

upvoted 1 times

Denise123 10 months, 1 week ago

Selected Answer: D

If gaining real-time insights involves complex analytics or custom processing, Amazon Kinesis Data Analytics with AWS Lambda is likely a more suitable choice. If the requirements can be met with simpler data transformations, Amazon Kinesis Data Firehose might provide a more straightforward and potentially lower-latency solution.

In other words, if this data is in GZIP files and the processing requirements are relatively simple, Amazon Kinesis Data Firehose might be a more straightforward and efficient choice. GZIP files typically contain compressed data, and if our primary objective is to ingest, transform, and load this data into other AWS services for real-time insights, Kinesis Data Firehose provides a managed and streamlined solution that can handle GZIP compression.

upvoted 1 times

Denise123 10 months, 1 week ago

The answer can be A , please comment if you have more clarity. After searching more, I also found out the following:

(I have missed the SQL requirement in the question)

Use Amazon Kinesis Data Analytics if you need SQL-based processing and advanced analytics capabilities for streaming data.

Use Amazon Kinesis Data Firehose if your primary requirement is to deliver, transform, and load streaming data into various AWS destinations with simplified configurations, but not for SQL-based processing.

upvoted 1 times

🗳️ 👤 **elvin_ml_qayiran25091992razor** 1 year, 1 month ago

Selected Answer: A

A is correct, why D xiyarsan sen?

upvoted 1 times

🗳️ 👤 **Mickey321** 1 year, 4 months ago

Selected Answer: A

A is correct

upvoted 1 times

🗳️ 👤 **kaike_reis** 1 year, 4 months ago

Selected Answer: A

"allow the use of <https://www.examttopics.com/exams/amazon/aws-certified-machine-learning-specialty/view/13/#f> SQL to query the stream with the LEAST latency?"

Well, the only solution that presents SQL query is (A). It's a description of KDA.

upvoted 2 times

🗳️ 👤 **Nadia0012** 1 year, 9 months ago

Selected Answer: A

the term "least latency" is the hidden point. with Glue we can have near real-time but Kinesis data analytics will give you real-time transformation with internal lambda

upvoted 3 times

🗳️ 👤 **Valcilio** 1 year, 9 months ago

Selected Answer: A

A is correct, with KDA you can run sql queries in the data during the streaming (real-time SQL queries).

upvoted 2 times

🗳️ 👤 **bakarys** 1 year, 10 months ago

Selected Answer: D

D. Amazon Kinesis Data Firehose to transform the data and put it into an Amazon S3 bucket would be the best solution for allowing the use of SQL to query the stream with the least latency. Amazon Kinesis Data Firehose can be configured to transform the data before writing it to Amazon S3 in real-time. Once the data is in S3, it can be queried using SQL with Amazon Athena, which is a serverless query service that allows running standard SQL queries against data stored in Amazon S3. This approach provides the lowest latency compared to other options and requires minimal setup and maintenance.

upvoted 3 times

🗳️ 👤 **akgarg00** 1 year ago

Query has to be run on stream so firehose not possible.

upvoted 1 times

🗳️ 👤 **OssamaAbdelatif** 2 years, 1 month ago

Selected Answer: A

A is correct.

upvoted 1 times

🗳️ 👤 **AddiWei** 2 years, 10 months ago

And somehow "transformation" is added to the answer as a requirement when it clearly was not part of the requirement from the question.

upvoted 2 times

🗳️ 👤 **apprehensive_scar** 2 years, 11 months ago

AAAAAAA

upvoted 1 times

🗳️ 👤 **HalloSpencer** 3 years, 1 month ago

what about "LEAST latency"?

upvoted 4 times

🗳️ 👤 **Erso** 3 years, 1 month ago

A is correct. you can pre-process data prior to running SQL queries with Kinesis Data Analytics and Lambda (more or less) is always a best practice :)

upvoted 3 times

🗨️ 👤 **JayK** 3 years, 3 months ago

Answer is B. Kinesis Data Analytics does not do any transformation, it is only for querying. Glue ETL can have scripts that can transform the data
upvoted 2 times

🗨️ 👤 **SophieSu** 3 years, 1 month ago

so you need lambda
upvoted 1 times

🗨️ 👤 **am7** 3 years, 3 months ago

But we need to run SQL on real time stream data.
upvoted 1 times

A retail company intends to use machine learning to categorize new products. A labeled dataset of current products was provided to the Data Science team. The dataset includes 1,200 products. The labeled dataset has 15 features for each product such as title dimensions, weight, and price. Each product is labeled as belonging to one of six categories such as books, games, electronics, and movies.

Which model should be used for categorizing new products using the provided dataset for training?

- A. AnXGBoost model where the objective parameter is set to multi:softmax
- B. A deep convolutional neural network (CNN) with a softmax activation function for the last layer
- C. A regression forest where the number of trees is set equal to the number of product categories
- D. A DeepAR forecasting model based on a recurrent neural network (RNN)

Suggested Answer: B

Community vote distribution

A (100%)

 **rsimham** Highly Voted 2 years, 9 months ago

Ans: A XGBoost multi class classification. <https://medium.com/@gabrielziegler3/multiclass-multilabel-classification-with-xgboost-66195e4d9f2d>

CNN is used for image classificaiton problems

upvoted 34 times

 **JayK** Highly Voted 2 years, 9 months ago

Answer is A. This a classification problem thus XGBoost and the fact that there are six categories SOFTMAX is the right activation function

upvoted 14 times

 **sonoluminescence** Most Recent 8 months ago

Selected Answer: A

Deep convolutional neural networks (CNNs) are primarily used for image processing tasks. Given that the dataset provided is structured/tabular in nature (with features like dimensions, weight, and price) and does not mention image data, a CNN is not the most appropriate choice.

upvoted 2 times

 **loict** 9 months, 2 weeks ago

Selected Answer: A

A. YES - perfect fit, multi:softmax the highest probability class is assigned

B. NO - CNN is for imaging

C. NO - regression forest is for continuous variables, we can discrete classification

D. NO - it is classification, not forecasting

upvoted 3 times

 **Mickey321** 10 months ago

Selected Answer: A

Option A XGBoost multi class classification

upvoted 1 times

 **kaike_reis** 11 months ago

Selected Answer: A

A is the answer.

upvoted 1 times

 **oso0348** 1 year, 3 months ago

Selected Answer: A

The XGBoost algorithm is a popular and effective technique for multi-class classification. The objective parameter can be set to multi:softmax, which uses a softmax objective function for multi-class classification. This will train the model to predict the probability of each product belonging to each category, and the most probable category will be chosen as the final prediction.

A deep convolutional neural network (CNN) (B) is a powerful technique commonly used for image recognition tasks. However, it is less appropriate for tabular data like the dataset provided.

upvoted 1 times

🗲️ 👤 **Konga98** 1 year, 5 months ago

Selected Answer: A

A, CNN is used for image classification. It would be suitable if we were classifying products using pictures of them.

upvoted 1 times

🗲️ 👤 **yemauricio** 1 year, 6 months ago

Selected Answer: A

<https://xgboost.readthedocs.io/en/stable/parameter.html>

upvoted 2 times

🗲️ 👤 **GiyeonShin** 1 year, 6 months ago

Selected Answer: A

B - CNN is used for dataset that have "local intermediate features" ex) images, or textCNN, etc

C - We need classification model, not regression model

D - RNN is used for dataset that have sequential features

A is correct

upvoted 3 times

🗲️ 👤 **Peeking** 1 year, 6 months ago

Selected Answer: A

A is the best option here. Only 1200 items and 6 classes are not enough data to involve a deep neural architecture for classification.

upvoted 1 times

🗲️ 👤 **Shailendraa** 1 year, 9 months ago

Ans- A ... For multiclassification - multi: SoftMax

upvoted 2 times

🗲️ 👤 **Morsa** 1 year, 11 months ago

Selected Answer: A

That is a classification problem so A is the answer

upvoted 1 times

🗲️ 👤 **apprehensive_scar** 2 years, 4 months ago

Selected Answer: A

Easy one. A is correct

upvoted 2 times

🗲️ 👤 **Kevinkoo** 2 years, 7 months ago

Selected Answer: A

A is correct

upvoted 3 times

🗲️ 👤 **stardustWu** 2 years, 8 months ago

Definitely A.

upvoted 1 times

🗲️ 👤 **syu31svc** 2 years, 8 months ago

100% is A; the the others are clearly wrong

Convolutional Neural Network (ConvNet or CNN) is a special type of Neural Network used effectively for image recognition and classification

Recurrent neural networks (RNN) are a class of neural networks that is powerful for modeling sequence data such as time series or natural language

upvoted 5 times

A Data Scientist is working on an application that performs sentiment analysis. The validation accuracy is poor, and the Data Scientist thinks that the cause may be a rich vocabulary and a low average frequency of words in the dataset. Which tool should be used to improve the validation accuracy?

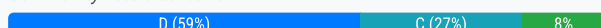
- A. Amazon Comprehend syntax analysis and entity detection
- B. Amazon SageMaker BlazingText cbow mode
- C. Natural Language Toolkit (NLTK) stemming and stop word removal
- D. Scikit-learn term frequency-inverse document frequency (TF-IDF) vectorizer

Suggested Answer: D

Reference:

<https://monkeylearn.com/sentiment-analysis/>

Community vote distribution



tap123 Highly Voted 3 years, 3 months ago

D is correct. Amazon Comprehend syntax analysis \neq Amazon Comprehend sentiment analysis. You need to read choices very carefully.
upvoted 35 times

mawsman 3 years, 3 months ago

We're looking only to improve the validation accuracy and Comprehend syntax analysis would help that because the word set is rich and the sentiment carrying words infrequent. We're not looking to replace the sentiment analysis tool with Comprehend.
upvoted 4 times

DonaldCMLIN Highly Voted 3 years, 3 months ago

AWS COMPREHEND IS A NATURAL LANGUAGE PROCESSING (NLP) SERVICE THAT USES MACHINE LEARNING TO DISCOVER INSIGHTS FROM TEXT. AMAZON COMPREHEND PROVIDES KEYPHRASE EXTRACTION, SENTIMENT ANALYSIS, ENTITY RECOGNITION, TOPIC MODELING, AND LANGUAGE DETECTION APIS SO YOU CAN EASILY INTEGRATE NATURAL LANGUAGE PROCESSING INTO YOUR APPLICATIONS.

[HTTPS://AWS.AMAZON.COM/COMPREHEND/FEATURES/?NC1=H_LS](https://aws.amazon.com/comprehend/features/?nc1=h_ls)

JUST THROUGH AMAZON COMPREHEND IS MUCH EASY THAN OTHER
THE MUCH MORE CONVENIENT ANSWER IS A.
upvoted 23 times

ComPah 3 years, 3 months ago

Agree Also Keyword is TOOL rest are frameworks
upvoted 2 times

VR10 Most Recent 10 months, 2 weeks ago

Selected Answer: A

Both Amazon Comprehend and the TF-IDF with a classifier solution are valid. If ease of use and pre-trained capabilities are high priorities, Comprehend is a solid option. If customization and dataset-specific nuances are crucial, building a custom model with TF-IDF may be needed. Since Comprehend is a tool, I am going with A.
upvoted 1 times

phdykd 11 months, 3 weeks ago

D. Scikit-learn term frequency-inverse document frequency (TF-IDF) vectorizer

Here's why:

TF-IDF Vectorizer: This tool from Scikit-learn is effective in handling issues of rich vocabularies and low frequency words. TF-IDF down-weights words that appear frequently across documents (thus might be less informative) and gives more weight to words that appear less frequently but might be more indicative of the sentiment. This approach can enhance the model's ability to focus on more relevant features, potentially improving validation accuracy.

upvoted 4 times

🗳️ 👤 **geon13** 1 year, 1 month ago

C I think c is correct. stemming involves reducing words to their root or base form, and stop word removal involves removing common words (e.g., "the," "and," "is") that may not contribute much to sentiment analysis. By using NLTK for stemming and stop word removal, you can simplify the vocabulary and potentially improve the model's ability to capture sentiment from the remaining meaningful words.

A - syntax and entity recognition wont solve the scenario

B - blaze text for words.

D - capturing the importance of words in a document collection. frequency of a word in a document.

upvoted 4 times

🗳️ 👤 **elvin_ml_qayiran25091992razor** 1 year, 1 month ago

Selected Answer: D

D is the correct guys

upvoted 1 times

🗳️ 👤 **wendaz** 1 year, 2 months ago

Amazon Comprehend's syntax analysis and entity detection are more about understanding the structure of sentences and identifying entities within the text rather than tackling the problem of a rich vocabulary with low average frequency of words.

TF-IDF vectorization is a technique that can help reduce the impact of common, low-information words in the dataset while emphasizing the importance of more informative, less frequent words. This could potentially improve the validation accuracy by addressing the identified problem.

upvoted 1 times

🗳️ 👤 **loict** 1 year, 3 months ago

Selected Answer: A

A. YES - he works on an application and not a model, Amazon Comprehend is the ready-to-use tool he wants; TF-IDF is built-in

B. NO - word2vec will be challenged with low frequency terms; GloVe and FastText are better for that

C. NO - the vocabulary is rich, so stemming and stop word removal will not address the core issue

D. NO - right approach, but that is not "a tool"

upvoted 1 times

🗳️ 👤 **Mickey321** 1 year, 4 months ago

Selected Answer: D

Option D. This approach can help in reducing the impact of words that occur frequently in the dataset and increasing the impact of words that occur less frequently. This can help in improving the accuracy of the model.

upvoted 2 times

🗳️ 👤 **ashii007** 1 year, 4 months ago

The answer is B.

Blazing text can handle OOV words as explained below. <https://docs.aws.amazon.com/sagemaker/latest/dg/blazingtext.html>

upvoted 2 times

🗳️ 👤 **jyrajan69** 1 year, 4 months ago

This is an AWS exam, so why would you choose anything other than A or B, and based on the link, it looks like B most likely

upvoted 2 times

🗳️ 👤 **kaike_reis** 1 year, 4 months ago

Selected Answer: D

The passage "low average frequency of words" points directly to the use of TF-IDF. Letter A deviates from what the question proposes and is discarded. Letter B proposes a radical change in my POV. Letter C does not solve the passage mentioned at the beginning. Letter D is correct.

upvoted 2 times

🗳️ 👤 **GOSD** 1 year, 7 months ago

The Amazon SageMaker BlazingText algorithm provides highly optimized implementations of the Word2vec and text classification algorithms. The Word2vec algorithm is useful for many downstream natural language processing (NLP) tasks, such as ****sentiment analysis, named entity recognition, machine translation, etc. Text classification is an important task for applications that perform web searches, information retrieval, ranking, and document classification.

upvoted 1 times

🗳️ 👤 **vassof95** 1 year, 8 months ago

Selected Answer: D

I would say since the buzzword "low average frequency" comes up, the safe choice would be the tfidf vectorizer.

I go for D.

upvoted 2 times

🗨️ 👤 **ParkXD** 1 year, 8 months ago

Selected Answer: D

The Scikit-learn term frequency-inverse document frequency (TF-IDF) vectorizer is a widely used tool to mitigate the high dimensionality of text data. Option A, Amazon Comprehend syntax analysis, and entity detection, can help in extracting useful features from the text, but it does not address the issue of high dimensionality.

Option B, Amazon SageMaker BlazingText cbow mode, is a tool for training word embeddings, which can help to represent words in a lower dimensional space. However, it does not directly address the issue of high dimensionality and low frequency of words.

Option C, Natural Language Toolkit (NLTK) stemming and stop word removal, can reduce the dimensionality of the feature space, but it does not address the issue of low-frequency words that are important for sentiment analysis.

upvoted 5 times

🗨️ 👤 **cpal012** 1 year, 9 months ago

Selected Answer: C

Emphasis is on the rich words - so stemming can help reduce these to more common words. Blazing Text in cbow mode doesn't seem relevant as it's about providing words given a context. And TF-IDF I'm not sure would do anything except highlight the problem you are already having?

upvoted 1 times

🗨️ 👤 **bakarys** 1 year, 10 months ago

Selected Answer: D

D. Scikit-learn term frequency-inverse document frequency (TF-IDF) vectorizer would be the best tool to use in this scenario. The TF-IDF vectorizer will give less weight to the less frequent words in the dataset, and allow the more informative and frequent words to have a greater impact on the sentiment analysis. This can help to improve the validation accuracy of the model.

upvoted 5 times

Machine Learning Specialist is building a model to predict future employment rates based on a wide range of economic factors. While exploring the data, the Specialist notices that the magnitude of the input features vary greatly. The Specialist does not want variables with a larger magnitude to dominate the model.

What should the Specialist do to prepare the data for model training?

- A. Apply quantile binning to group the data into categorical bins to keep any relationships in the data by replacing the magnitude with distribution.
- B. Apply the Cartesian product transformation to create new combinations of fields that are independent of the magnitude.
- C. Apply normalization to ensure each field will have a mean of 0 and a variance of 1 to remove any significant magnitude.
- D. Apply the orthogonal sparse bigram (OSB) transformation to apply a fixed-size sliding window to generate new features of a similar magnitude.

Suggested Answer: C

Reference:

<https://docs.aws.amazon.com/machine-learning/latest/dg/data-transformations-reference.html>



Community vote distribution

C (100%)

  **rsimham** Highly Voted 2 years, 9 months ago



Ans: C; Normalization is correct

upvoted 34 times

  **gcpwhiz** 2 years, 7 months ago

Ans is not C. What is listed there is the definition of STANDARDIZATION. Normalization just scales and is not useful for reducing the effect of outliers

upvoted 4 times

  **gcpwhiz** 2 years, 7 months ago

nevermind ignore this

upvoted 4 times

  **Phong** Highly Voted 2 years, 8 months ago

Guys, I passed the exam today. It is a tough one but there are many questions here. Good luck everyone! Thank examtopics

upvoted 14 times

  **haison8x** 2 years, 8 months ago

Hi Phong!

Please add my skype: haison8x

upvoted 2 times

  **Mickey321** Most Recent 10 months ago

Selected Answer: C

Ans: C; Normalization is correct

upvoted 2 times

  **kaike_reis** 11 months ago

C (Yep, STANDARDIZATION is the correct name)

That's an odd question for me

upvoted 1 times

  **OssamaAbdelatif** 1 year, 7 months ago

Selected Answer: C

ans C is correct.

upvoted 1 times

  **Deepsachin** 2 years, 7 months ago

ANS should be C as Normalization work best in case of amplitude diff

upvoted 1 times

🗨️ **grandgale** 2 years, 8 months ago

Hi, guys,

First thanks this website for the information it provided.

However, the ML exam has updated most of the questions. only 20+ questions here are included in today's test. Anyway, it is still helpful.

GOOD LUCK EVERYONE!

upvoted 10 times

🗨️ **joker34** 2 years, 8 months ago

So there are 40+ other questions on the exam that aren't included in Examtopics?

upvoted 2 times

🗨️ **nez15** 2 years, 9 months ago

QUESTION 69

A large consumer goods manufacturer has the following products on sale:

- 34 different toothpaste variants
- 48 different toothbrush variants
- 43 different mouthwash variants

The entire sales history of all these products is available in Amazon S3. Currently, the company is using custom-built autoregressive integrated moving average (ARIMA) models to forecast demand for these products. The company wants to predict the demand for a new product that will soon be launched.

Which solution should a Machine Learning Specialist apply?

- A. Train a custom ARIMA model to forecast demand for the new product.
- B. Train an Amazon SageMaker DeepAR algorithm to forecast demand for the new product.
- C. Train an Amazon SageMaker k-means clustering algorithm to forecast demand for the new product.
- D. Train a custom XGBoost model to forecast demand for the new product.

Correct Answer: B

upvoted 4 times

🗨️ **VB** 2 years, 8 months ago

<https://aws.amazon.com/blogs/machine-learning/forecasting-time-series-with-dynamic-deep-learning-on-aws/>

Answer: B

upvoted 1 times

🗨️ **nez15** 2 years, 9 months ago

QUESTION 68

An agency collects census information within a country to determine healthcare and social program needs by province and city. The census form collects responses for approximately 500 questions from each citizen.

Which combination of algorithms would provide the appropriate insights? (Select TWO.)

- A. The factorization machines (FM) algorithm
- B. The Latent Dirichlet Allocation (LDA) algorithm
- C. The principal component analysis (PCA) algorithm
- D. The k-means algorithm
- E. The Random Cut Forest (RCF) algorithm

Correct Answer: CD

upvoted 5 times

🗨️ **VB** 2 years, 8 months ago

<https://aws.amazon.com/blogs/machine-learning/analyze-us-census-data-for-population-segmentation-using-amazon-sagemaker/>

Answer: C and D

upvoted 4 times

🗨️ **cybe001** 2 years, 9 months ago

I think the answer is A and B.

The census question and answer will be in text. Use LDA (unsupervised algorithm) which takes the census question/answer and groups them into categories. Use the categorization to group the people and identify similar people.

Use the Factorization Machine to group the people. For each person identify if they answer a question or not. Find the total questions they answered and that will be the Target variable. Now the problem is similar to movie recommendation (consider each question a movie and the total

number of questions answered will be the Rating). Based on the questions a Person answered, Factorization Machine groups the people.

Findings from both the algorithms can be used to compare and identify the people for the social programs.

upvoted 2 times

  **kaike_reis** 11 months ago



it's CD

upvoted 1 times

  **jasonsunbao** 2 years, 9 months ago

FM is mainly used in recommendation system to find hidden variables between two known variables to find correlation between two variables.

upvoted 1 times

  **nez15** 2 years, 9 months ago

QUESTION 67

A. Use AWS Lambda to trigger an AWS Step Functions workflow to wait for dataset uploads to complete in Amazon S3. Use AWS Glue to join the datasets. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.

B. Develop the ETL workflow using AWS Lambda to start an Amazon SageMaker notebook instance. Use a lifecycle configuration script to join the datasets and persist the results in Amazon S3. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.

C. Develop the ETL workflow using AWS Batch to trigger the start of ETL jobs when data is uploaded to Amazon S3. Use AWS Glue to join the datasets in Amazon S3. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.

D. Use AWS Lambda to chain other Lambda functions to read and join the datasets in Amazon S3 as soon as the data is uploaded to Amazon S3. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.

Correct Answer: A

upvoted 6 times

  **nez15** 2 years, 9 months ago


QUESTION 67

A Machine Learning Specialist is developing a daily ETL workflow containing multiple ETL jobs. The workflow consists of the following processes:

- Start the workflow as soon as data is uploaded to Amazon S3.
- When all the datasets are available in Amazon S3, start an ETL job to join the uploaded datasets with multiple terabyte-sized datasets already stored in Amazon S3.
- Store the results of joining datasets in Amazon S3.
- If one of the jobs fails, send a notification to the Administrator.

Which configuration will meet these requirements?

upvoted 3 times

  **nez15** 2 years, 9 months ago

QUESTION 66

A Machine Learning Specialist must build out a process to query a dataset on Amazon S3 using Amazon Athena. The dataset contains more than 800,000 records stored as plaintext CSV files. Each record contains 200 columns and is approximately 1.5 MB in size. Most queries will span 5 to 10 columns only.

How should the Machine Learning Specialist transform the dataset to minimize query runtime?

- A. Convert the records to Apache Parquet format.
- B. Convert the records to JSON format.
- C. Convert the records to GZIP CSV format.
- D. Convert the records to XML format.

Correct Answer: A

upvoted 11 times

A Machine Learning Specialist must build out a process to query a dataset on Amazon S3 using Amazon Athena. The dataset contains more than 800,000 records stored as plaintext CSV files. Each record contains 200 columns and is approximately 1.5 MB in size. Most queries will span 5 to 10 columns only.

How should the Machine Learning Specialist transform the dataset to minimize query runtime?

- A. Convert the records to Apache Parquet format.
- B. Convert the records to JSON format.
- C. Convert the records to GZIP CSV format.
- D. Convert the records to XML format.

Suggested Answer: A

Using compressions will reduce the amount of data scanned by Amazon Athena, and also reduce your S3 bucket storage. It's a Win-Win for your AWS bill.

Supported formats: GZIP, LZO, SNAPPY (Parquet) and ZLIB.

Reference:

<https://www.cloudforecast.io/blog/using-parquet-on-athena-to-save-money-on-aws/>



Community vote distribution

A (100%)

  **Erso** Highly Voted 2 years, 8 months ago

Answer A seems correct...

upvoted 12 times

  **Erso** 2 years, 8 months ago

sorry, the link <https://aws.amazon.com/blogs/big-data/prepare-data-for-model-training-and-invoke-machine-learning-models-with-amazon-athena/>

upvoted 1 times

  **sonalev419** Highly Voted 2 years, 8 months ago

A (Most queries will span 5 to 10 columns only)

upvoted 5 times

  **Mickey321** Most Recent 10 months ago

Selected Answer: A

Option A

upvoted 1 times

  **exam_prep** 2 years, 1 month ago

clue is: most queries will span 5 to 10 column while there are 200 columns. Indicating Data Warehouse means columnar storage. Option A is correct.

upvoted 2 times

  **edvardo** 2 years, 7 months ago

Selected Answer: A

A. See <https://aws.amazon.com/blogs/big-data/analyzing-data-in-s3-using-amazon-athena/>

upvoted 4 times

A Machine Learning Specialist is developing a daily ETL workflow containing multiple ETL jobs. The workflow consists of the following processes:

- * Start the workflow as soon as data is uploaded to Amazon S3.
 - * When all the datasets are available in Amazon S3, start an ETL job to join the uploaded datasets with multiple terabyte-sized datasets already stored in Amazon S3.
 - * Store the results of joining datasets in Amazon S3.
 - * If one of the jobs fails, send a notification to the Administrator.
- Which configuration will meet these requirements?

- A. Use AWS Lambda to trigger an AWS Step Functions workflow to wait for dataset uploads to complete in Amazon S3. Use AWS Glue to join the datasets. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.
- B. Develop the ETL workflow using AWS Lambda to start an Amazon SageMaker notebook instance. Use a lifecycle configuration script to join the datasets and persist the results in Amazon S3. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.
- C. Develop the ETL workflow using AWS Batch to trigger the start of ETL jobs when data is uploaded to Amazon S3. Use AWS Glue to join the datasets in Amazon S3. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.
- D. Use AWS Lambda to chain other Lambda functions to read and join the datasets in Amazon S3 as soon as the data is uploaded to Amazon S3. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.

Suggested Answer: A

Reference:

<https://aws.amazon.com/step-functions/use-cases/>

Community vote distribution

A (100%)

  **HaiHN** Highly Voted 2 years, 9 months ago

- A: Correct. S3 events can trigger AWS Lambda function.
- B: Wrong. There's nothing to do with SageMaker in the provided context.
- C: Wrong. AWS Batch cannot receive events from S3 directly.
- D: Wrong. Will not meet the requirement: "When all the datasets are available in Amazon S3..."

<https://docs.aws.amazon.com/step-functions/latest/dg/tutorial-cloudwatch-events-s3.html>

upvoted 35 times

  **scuzzy2010** 2 years, 9 months ago

I agree. Step Functions can be used to implement a workflow. In this case, wait for all the datasets to be loaded before triggering the glue job.

upvoted 3 times

  **cloud_trail** 2 years, 8 months ago

Actually, I think that D does meet the requirement of waiting until all datasets are in S3, BUT you do need Glue to join the datasets. Answer is still A.



upvoted 4 times

  **Mickey321** Most Recent 10 months ago

Selected Answer: A

Option A



upvoted 1 times

  **Valcilio** 1 year, 3 months ago

Selected Answer: A

Batch isn't event driven, answer is A.

upvoted 1 times

  **matteocal** 1 year, 11 months ago

If EMR were present I would have chose that because of the size of dataset, else is Glue

upvoted 1 times

🗨️ 👤 **ZSun** 1 year, 2 months ago

exactly, this is also where I got confused. Since Glue is not good at handling such large dataset, multiple terabyte-sized datasets + multiple ETL jobs + daily

upvoted 1 times

🗨️ 👤 **Huy** 2 years, 7 months ago

A. The answer omits stuffs like Lambda functions and Event Bridge. <https://aws.amazon.com/blogs/big-data/orchestrate-multiple-etl-jobs-using-aws-step-functions-and-aws-lambda/>

upvoted 2 times

🗨️ 👤 **johnvik** 2 years, 8 months ago

<https://d1.awsstatic.com/r2018/a/product-page-diagram-aws-step-functions-use-case-aws-glue.bc69d97a332c2dd29abb724dd747fd82ae110352.png>

upvoted 2 times

An agency collects census information within a country to determine healthcare and social program needs by province and city. The census form collects responses for approximately 500 questions from each citizen.

Which combination of algorithms would provide the appropriate insights? (Choose two.)

- A. The factorization machines (FM) algorithm
- B. The Latent Dirichlet Allocation (LDA) algorithm
- C. The principal component analysis (PCA) algorithm
- D. The k-means algorithm
- E. The Random Cut Forest (RCF) algorithm

Suggested Answer: CD

The PCA and K-means algorithms are useful in collection of data using census form.

Community vote distribution

CD (100%)

  **HaiHN** Highly Voted 3 years, 8 months ago

C: (OK) Use PCA for reducing number of variables. Each citizen's response should have answer for 500 questions, so it should have 500 variables



D: (OK) Use K-means clustering

A: (Not OK) Factorization Machines Algorithm is usually used for tasks dealing with high dimensional sparse datasets

B: (Not OK) The Latent Dirichlet Allocation (LDA) algorithm should be used for task dealing topic modeling in NLP

E: (Not OK) Random Cut Forest should be used for detecting anormal in data

upvoted 33 times

  **hans1234** Highly Voted 3 years, 8 months ago

<https://aws.amazon.com/blogs/machine-learning/analyze-us-census-data-for-population-segmentation-using-amazon-sagemaker/>

Answer: C and D



upvoted 12 times

  **rodrick10** Most Recent 7 months, 2 weeks ago

Selected Answer: BD

If the form contains free-text answers, it would be interesting to apply LDA to identify the most frequent/relevant topics in the answers

upvoted 1 times

  **Mickey321** 1 year, 10 months ago

Selected Answer: CD

Option C and D

upvoted 1 times

  **Mickey321** 1 year, 10 months ago

The answer depends on the type of question is it is open ended then would need LDA hence B and D but if the question is a feature then PCA should work

upvoted 1 times

  **kaike_reis** 1 year, 11 months ago

Selected Answer: CD

C and D are the way

upvoted 1 times



  **ADVIT** 2 years ago

CD,

C - for reduce number of columns.

D - for data clustering

upvoted 1 times

  **Ajose0** 2 years, 4 months ago

Selected Answer: CD



- C. The principal component analysis (PCA) algorithm
- D. The k-means algorithm

PCA is a dimensionality reduction technique that can be used to identify the underlying structure of the census data. This algorithm can help to identify the most important questions and provide an overview of the relationship between the questions and the responses.

K-means is an unsupervised learning algorithm that can be used to segment the population into different groups based on their responses to the census questions. This algorithm can help to determine the healthcare and social program needs by province and city based on the responses collected from each citizen.

These algorithms can help to provide insights into the patterns and relationships within the census data, which can inform decision making for healthcare and social program planning.



upvoted 5 times

  **Peeking** 2 years, 6 months ago

Selected Answer: CD

Reduce dimensionality and cluster subjects.

upvoted 2 times

  **ac427** 3 years, 9 months ago

This is the same question as Topic 2 Q3

upvoted 1 times

  **muralee_xo** 2 years, 5 months ago

how to reach Topic 2 every questions here seem to belong to topic 1

upvoted 1 times

A large consumer goods manufacturer has the following products on sale:

- * 34 different toothpaste variants
- * 48 different toothbrush variants
- * 43 different mouthwash variants

The entire sales history of all these products is available in Amazon S3. Currently, the company is using custom-built autoregressive integrated moving average

(ARIMA) models to forecast demand for these products. The company wants to predict the demand for a new product that will soon be launched. Which solution should a Machine Learning Specialist apply?

- A. Train a custom ARIMA model to forecast demand for the new product.
- B. Train an Amazon SageMaker DeepAR algorithm to forecast demand for the new product.
- C. Train an Amazon SageMaker k-means clustering algorithm to forecast demand for the new product.
- D. Train a custom XGBoost model to forecast demand for the new product.

Suggested Answer: B

The Amazon SageMaker DeepAR forecasting algorithm is a supervised learning algorithm for forecasting scalar (one-dimensional) time series using recurrent neural networks (RNN). Classical forecasting methods, such as autoregressive integrated moving average (ARIMA) or exponential smoothing (ETS), fit a single model to each individual time series. They then use that model to extrapolate the time series into the future.

Reference:

<https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html>

Community vote distribution

B (100%)

 **HaiHN** Highly Voted 3 years, 8 months ago

B

<https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html>

"...When your dataset contains hundreds of related time series, DeepAR outperforms the standard ARIMA and ETS methods. You can also use the trained model to generate forecasts for new time series that are similar to the ones it has been trained on."

upvoted 18 times

 **ninomfr64** Most Recent 11 months, 3 weeks ago

Selected Answer: B

"You can also use the trained model to generate forecasts for new time series that are similar to the ones it has been trained on"

<https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html>

upvoted 1 times

 **james2033** 1 year, 3 months ago

Selected Answer: B

'autoregressive integrated moving average (ARIMA)' <--> DeepAR. <https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html>


upvoted 1 times

 **loict** 1 year, 9 months ago

Selected Answer: B

B - DeepAR is based on GluonTS, and can use multiple time series for learning


upvoted 1 times

 **Mickey321** 1 year, 10 months ago

Selected Answer: B

Option B

upvoted 1 times

 **Valcilio** 2 years, 3 months ago

Selected Answer: B

DeepAr for new products forever!

upvoted 4 times

🗲️ 👤 **Ajose0** 2 years, 4 months ago

Selected Answer: B

The DeepAR algorithm is a powerful time series forecasting algorithm that is designed to handle multiple time series data and can handle irregularly spaced time series data and missing values, making it a good fit for this task.

Additionally, the large amount of sales history data available in Amazon S3 makes the use of a deep learning algorithm like DeepAR more appropriate.
upvoted 2 times

🗲️ 👤 **Shailendraa** 2 years, 10 months ago

B <https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html>

upvoted 1 times

🗲️ 👤 **hans1234** 3 years, 8 months ago

It is B

upvoted 3 times

🗲️ 👤 **ac427** 3 years, 9 months ago

This is the same question as Topic 2 Q4

upvoted 1 times

A Machine Learning Specialist uploads a dataset to an Amazon S3 bucket protected with server-side encryption using AWS KMS. How should the ML Specialist define the Amazon SageMaker notebook instance so it can read the same dataset from Amazon S3?

- A. Define security group(s) to allow all HTTP inbound/outbound traffic and assign those security group(s) to the Amazon SageMaker notebook instance.
- B. Configure the Amazon SageMaker notebook instance to have access to the VPC. Grant permission in the KMS key policy to the notebook's KMS role.
- C. Assign an IAM role to the Amazon SageMaker notebook with S3 read access to the dataset. Grant permission in the KMS key policy to that role.
- D. Assign the same KMS key used to encrypt data in Amazon S3 to the Amazon SageMaker notebook instance.

Suggested Answer: D

Reference:

<https://docs.aws.amazon.com/sagemaker/latest/dg/encryption-at-rest.html>

Community vote distribution

C (93%)

7%

 **seanLu** Highly Voted 3 years, 2 months ago

Should be C.

"You don't need to specify the AWS KMS key ID when you download an SSE-KMS-encrypted object from an S3 bucket. Instead, you need the permission to decrypt the AWS KMS key.

When a user sends a GET request, Amazon S3 checks if the AWS Identity and Access Management (IAM) user or role that sent the request is authorized to decrypt the key associated with the object. If the IAM user or role belongs to the same AWS account as the key, then the permission to decrypt must be granted on the AWS KMS key's policy."

https://aws.amazon.com/premiumsupport/knowledge-center/decrypt-kms-encrypted-objects-s3/?nc1=h_ls

upvoted 29 times

 **askaron** Highly Voted 3 years, 3 months ago

Should be C.

I think it is not possible to assign a key directly to a Sagemaker notebook instance like D suggests.

Normally in AWS in general, IAM roles are used to do so. So C.


upvoted 6 times

 **james2033** Most Recent 9 months, 3 weeks ago

Selected Answer: C

'IAM role' principle of least privilege (PoLP)

upvoted 1 times

 **VR10** 10 months, 1 week ago

Selected Answer: C

IAM roles securely provide temporary AWS credentials that services (like SageMaker notebooks) can assume to access other resources. This avoids using long-lived access keys or directly embedding API keys into code.

KMS Key Policy: This policy controls access to your KMS key. Granting the notebook's role permission within this policy lets SageMaker decrypt the data when reading from S3.

upvoted 1 times

 **endeesa** 1 year, 1 month ago

Selected Answer: C

Seems to follow the best cloud authorization practice

upvoted 1 times

 **sonoluminescence** 1 year, 2 months ago

Selected Answer: C

IAM role associated with the SageMaker notebook instance must be given permissions in the KMS key policy to decrypt the data using the KMS key that was used for encryption.

upvoted 1 times

🗳️ 👤 **AmeeraM** 1 year, 2 months ago

Selected Answer: C

answer is C

upvoted 1 times

🗳️ 👤 **Mickey321** 1 year, 4 months ago

Selected Answer: C

Assign an IAM role to the Amazon SageMaker notebook with S3 read access to the dataset. Grant permission in the KMS key policy to that role. To read data from Amazon S3 that is encrypted with AWS KMS, the Amazon SageMaker notebook instance needs to have both S3 read access and KMS decrypt permissions. This can be achieved by assigning an IAM role to the notebook instance that has the necessary policies attached, and by granting permission in the KMS key policy to that role.

upvoted 1 times

🗳️ 👤 **ADVIT** 1 year, 6 months ago

C only.

upvoted 1 times

🗳️ 👤 **earthMover** 1 year, 7 months ago

Selected Answer: C

Should be C. The reference doc provided did not have any information about assigning keys to the notebook. Doing so become very cumbersome as you can have 100's of notebooks and its not scalable. Someone needs to moderate these answers.

upvoted 1 times

🗳️ 👤 **oso0348** 1 year, 8 months ago

Selected Answer: C

To allow an Amazon SageMaker notebook instance to read a dataset stored in an Amazon S3 bucket that is protected with server-side encryption using AWS KMS, the ML Specialist should assign an IAM role to the Amazon SageMaker notebook with S3 read access to the dataset. The IAM role should have permissions to access the S3 bucket and the KMS key that was used to encrypt the data. This role should be granted permission in the KMS key policy to allow it to decrypt the data.

upvoted 1 times

🗳️ 👤 **Nadia0012** 1 year, 9 months ago

Selected Answer: D

To encrypt the machine learning (ML) storage volume that is attached to notebooks, processing jobs, training jobs, hyperparameter tuning jobs, batch transform jobs, and endpoints, you can pass a AWS KMS key to SageMaker. If you don't specify a KMS key, SageMaker encrypts storage volumes with a transient key and discards it immediately after encrypting the storage volume. For notebook instances, if you don't specify a KMS key, SageMaker encrypts both OS volumes and ML data volumes with a system-managed KMS key.

upvoted 1 times

🗳️ 👤 **Nadia0012** 1 year, 9 months ago

I correct myself- Option C is correct:

Background

AWS Key Management Service (AWS KMS) enables Server-side encryption to protect your data at rest. Amazon SageMaker training works with KMS encrypted data if the IAM role used for S3 access has permissions to encrypt and decrypt data with the KMS key. Further, a KMS key can also be used to encrypt the model artifacts at rest using Amazon S3 server-side encryption. Additionally, a KMS key can also be used to encrypt the storage volume attached to training, endpoint, and transform instances. In this notebook, we demonstrate SageMaker encryption capabilities using KMS-managed keys.

resource: [https://github.com/aws/amazon-sagemaker-](https://github.com/aws/amazon-sagemaker-examples/blob/main/advanced_functionality/handling_kms_encrypted_data/handling_kms_encrypted_data.ipynb)

[examples/blob/main/advanced_functionality/handling_kms_encrypted_data/handling_kms_encrypted_data.ipynb](https://github.com/aws/amazon-sagemaker-examples/blob/main/advanced_functionality/handling_kms_encrypted_data/handling_kms_encrypted_data.ipynb)

Option D is correct if sagemaker does the encryption, if you are dealing with encrypted data then C is 100% correct.

upvoted 3 times

🗳️ 👤 **Ajose0** 1 year, 10 months ago

Selected Answer: C

C. Assign an IAM role to the Amazon SageMaker notebook with S3 read access to the dataset. Grant permission in the KMS key policy to that role.

To access the encrypted dataset in Amazon S3, the Amazon SageMaker notebook instance must have the appropriate permissions. This can be achieved by assigning an IAM role to the notebook with read access to the dataset in Amazon S3 and granting permission in the KMS key policy to that role. This ensures that the notebook has the necessary permissions to access the encrypted data in Amazon S3, while adhering to best practices for securing sensitive data.

upvoted 2 times

🗨️ 👤 **ystotest** 2 years, 1 month ago

Selected Answer: C

agreed with C

upvoted 3 times

🗨️ 👤 **AmakamaxZanny** 2 years, 10 months ago

Answer is C : Open the IAM console. Add a policy to the IAM user that grants the permissions to upload and download from the bucket. You can use a policy that's similar to the following:

<https://aws.amazon.com/premiumsupport/knowledge-center/s3-bucket-access-default-encryption/>

(number 2)

upvoted 1 times

🗨️ 👤 **Deepsachin** 3 years, 1 month ago

Seems to be D

<https://docs.aws.amazon.com/sagemaker/latest/dg/encryption-at-rest-nbi.html>

upvoted 2 times

🗨️ 👤 **Madwyn** 3 years, 2 months ago

Not D as if you assign the key in the notebook, that's not secure, it will make the encryption ineffective. Instead, you assign the access permission by using IAM.

upvoted 1 times

A Data Scientist needs to migrate an existing on-premises ETL process to the cloud. The current process runs at regular time intervals and uses PySpark to combine and format multiple large data sources into a single consolidated output for downstream processing.

The Data Scientist has been given the following requirements to the cloud solution:

- ⇒ Combine multiple data sources.
- ⇒ Reuse existing PySpark logic.
- ⇒ Run the solution on the existing schedule.
- ⇒ Minimize the number of servers that will need to be managed.

Which architecture should the Data Scientist use to build this solution?

A. Write the raw data to Amazon S3. Schedule an AWS Lambda function to submit a Spark step to a persistent Amazon EMR cluster based on the existing schedule. Use the existing PySpark logic to run the ETL job on the EMR cluster. Output the results to a `s3://processed/` location in Amazon S3 that is accessible for downstream use.

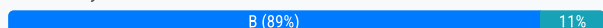
B. Write the raw data to Amazon S3. Create an AWS Glue ETL job to perform the ETL processing against the input data. Write the ETL job in PySpark to leverage the existing logic. Create a new AWS Glue trigger to trigger the ETL job based on the existing schedule. Configure the output target of the ETL job to write to a `s3://processed/` location in Amazon S3 that is accessible for downstream use.

C. Write the raw data to Amazon S3. Schedule an AWS Lambda function to run on the existing schedule and process the input data from Amazon S3. Write the Lambda logic in Python and implement the existing PySpark logic to perform the ETL process. Have the Lambda function output the results to a `s3://processed/` location in Amazon S3 that is accessible for downstream use.

D. Use Amazon Kinesis Data Analytics to stream the input data and perform real-time SQL queries against the stream to carry out the required transformations within the stream. Deliver the output results to a `s3://processed/` location in Amazon S3 that is accessible for downstream use.

Suggested Answer: D

Community vote distribution



🗳️ 👤 **Paul_NoName** Highly Voted 3 years, 9 months ago

B it is .
upvoted 29 times

🗳️ 👤 **[Removed]** 3 years, 8 months ago

I agree, B is serverless and reuses Pyspark. Similar example shown here: <https://docs.aws.amazon.com/glue/latest/dg/aws-glue-programming-python-samples-medicare.html>
upvoted 11 times

🗳️ 👤 **SophieSu** Highly Voted 3 years, 8 months ago

A is not correct because Minimize the number of servers that will need to be managed. EMR is not server-less.
B is correct. AWS Glue supports an extension of the PySpark Python dialect for scripting extract, transform, and load...
C is not correct because using Lambda for ETL you will not be able to Reuse existing PySpark logic
D is not correct because Kinesis is not server-less. And you can not Reuse existing PySpark logic
upvoted 12 times

🗳️ 👤 **xicocaio** Most Recent 9 months ago

Selected Answer: B

Option B (using AWS Glue for the ETL process) is the best solution for the described requirements.

A: This solution requires managing an Amazon EMR cluster, which would involve more server management than AWS Glue, violating the requirement to minimize the number of servers to be managed.

C: AWS Lambda is not ideal for this use case because it has resource limitations, including memory and execution time limits (15 minutes max), which might not be suitable for large-scale ETL operations involving PySpark logic.

D: Amazon Kinesis Data Analytics is focused on real-time stream processing, which doesn't fit the described scheduled batch processing scenario.
upvoted 1 times

🗳️ 👤 **akgarg00** 1 year, 7 months ago

Answer is A, as B clearly mentions that Pyspark code is written with leverage from already existing code. Also, the server architecture used currently is on-premises which will have more servers than solution A.

upvoted 2 times

🗳️ 👤 **sonoluminescence** 1 year, 8 months ago

Selected Answer: B

Amazon Kinesis Data Analytics is more suited for real-time processing and streaming data. The given use case does not indicate a need for real-time processing, so this might not be the best fit. Furthermore, it doesn't support PySpark natively.

upvoted 1 times

🗳️ 👤 **Shenannigan** 1 year, 10 months ago

Selected Answer: B

Voted B based on the serverless (minimum servers) and <https://docs.aws.amazon.com/glue/latest/dg/aws-glue-programming.html>

upvoted 1 times

🗳️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: B

Indeed B using Glue

upvoted 1 times

🗳️ 👤 **kaike_reis** 1 year, 11 months ago

B is the correct.

A you have to manage EMR, so it's wrong.

D you don't use Spark, so it's wrong.

C you will not be using Spark, so it's wrong.

upvoted 1 times

🗳️ 👤 **Maaayaaa** 2 years, 2 months ago

Selected Answer: B

B ticks all boxes. Minimize servers -> AWS managed services -> Glue.

upvoted 2 times

🗳️ 👤 **bakarys** 2 years, 4 months ago

Selected Answer: A

Option A would be the best response for this scenario.

This solution allows the Data Scientist to reuse the existing PySpark logic while migrating the ETL process to the cloud. The raw data is written to Amazon S3, and a Lambda function is scheduled to trigger a Spark step on a persistent EMR cluster based on the existing schedule. The PySpark logic is used to run the ETL job on the EMR cluster, and the results are output to a processed location in Amazon S3 that is accessible for downstream use. This solution minimizes the number of servers that need to be managed, and it allows for a seamless migration of the existing ETL process to the cloud.

upvoted 1 times

🗳️ 👤 **sqavi** 2 years, 4 months ago

Selected Answer: B

Option D is wrong it should be B

upvoted 1 times

🗳️ 👤 **Peeking** 2 years, 6 months ago

D cannot be answer as there is no streaming data or Realtime processing.

upvoted 2 times

🗳️ 👤 **salads** 2 years, 10 months ago

Selected Answer: B

the answer is b

upvoted 2 times

🗳️ 👤 **Nickname_L** 3 years, 8 months ago



Answer should be B. Serverless, on a regular schedule (no real time requirement), reuses PySpark code in Glue ETL script.

upvoted 4 times

🗳️ 👤 **gcpwhiz** 3 years, 8 months ago

Answer is B as they specifically ask about reusing existing PySpark, which can be done with Glue

upvoted 3 times

  **Aashi22** 3 years, 8 months ago

https://docs.aws.amazon.com/glue/latest/dg/creating_running_workflows.html

upvoted 1 times

  **astonm13** 3 years, 8 months ago

It is B. ! "Minimize number of servers to be managed". B is a Serverless solution which fulfils other requirements!

upvoted 2 times

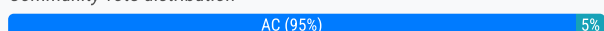
A Data Scientist is building a model to predict customer churn using a dataset of 100 continuous numerical features. The Marketing team has not provided any insight about which features are relevant for churn prediction. The Marketing team wants to interpret the model and see the direct impact of relevant features on the model outcome. While training a logistic regression model, the Data Scientist observes that there is a wide gap between the training and validation set accuracy.

Which methods can the Data Scientist use to improve the model performance and satisfy the Marketing team's needs? (Choose two.)

- A. Add L1 regularization to the classifier
- B. Add features to the dataset
- C. Perform recursive feature elimination
- D. Perform t-distributed stochastic neighbor embedding (t-SNE)
- E. Perform linear discriminant analysis

Suggested Answer: BE

Community vote distribution



bluer1 Highly Voted 2 years, 2 months ago

AC - correct answer
upvoted 13 times

lynn22 Most Recent 6 months, 3 weeks ago

Selected Answer: AE

I think ACE are all correct
upvoted 1 times

loict 9 months, 2 weeks ago

Selected Answer: AC

- A. YES - standard for overfitting
 - B. NO - we have already too much overfitting
 - C. YES - feature elimination can reduce model complexity and thus overfitting
 - D. NO - that does dimensionality reduction to 2D or 3D, for visualization; we want more than a few features
 - E. NO - LDA is an alternative to logistic regression; it may not address overfitting
- upvoted 4 times

Mickey321 10 months ago

Selected Answer: AC

A due to fitting

C Recursive feature elimination (RFE) is a wrapper method that iteratively removes features based on their importance scores from a classifier. RFE starts with all features and then eliminates the least important ones until a desired number of features is reached. This can help to reduce the dimensionality of the dataset and improve the model performance by removing irrelevant or redundant features. The Marketing team can then interpret the model by looking at the remaining features and their importance scores.

upvoted 1 times

kaike_reis 11 months ago

Selected Answer: AC

AC are the correct
upvoted 1 times

earthMover 1 year, 1 month ago

Selected Answer: AC

How can we add features to the dataset provided.... we can't make them up from thin air. Hopefully the moderators can provide some insight on this. I was thinking of paying for this site but the answers are all over the place.

upvoted 1 times

bakarys 1 year, 4 months ago

Selected Answer: AC

A. Add L1 regularization to the classifier and C. Perform recursive feature elimination are the methods that can be used to improve the model performance and satisfy the Marketing team's needs.

Explanation:

A. Adding L1 regularization to the logistic regression classifier can help to improve the model performance and reduce overfitting. This can also help to highlight the relevant features for churn prediction as L1 regularization can shrink the coefficients of irrelevant features to zero.

C. Recursive feature elimination can be used to select the most relevant features for the model. This can help to improve the model performance and highlight the relevant features for churn prediction.

upvoted 3 times

🗨️ 👤 **Ajose0** 1 year, 4 months ago

Selected Answer: AC

A. Adding L1 regularization can help to reduce overfitting by shrinking the coefficients of less important features towards zero, which can improve the model's generalization performance on the validation set.

C. Recursive feature elimination is a feature selection technique that removes the least important feature at each iteration and trains the model on the remaining features until a desired number of features is reached. This method can be used to identify the most relevant features for the prediction task and reduce the dimensionality of the dataset, leading to improved model performance and interpretability for the Marketing team.

upvoted 2 times

🗨️ 👤 **wisoxe8356** 1 year, 6 months ago

AC -

Key: logistic regression model = non linear in terms of Odds and Probability, however it is linear in terms of Log Odds.

Key: Large gap between training & validation = overfitting

=> 5 techniques to prevent overfitting:

1. Simplifying the model | 2. Early stopping
3. Use data augmentation | 4. Use regularization | 5. Use dropouts

A - yes to avoid overfitting (although i am thinking it is talking about regressor)

Not B - add feature will lead to overfitting

C - feature elimination - prevent overfitting

Not D - t-SNE is a nonlinear dimensionality reduction technique

Not E - find feature correlation only - Linear discriminant analysis (LDA), normal discriminant analysis (NDA), or discriminant function analysis is a generalization of Fisher's linear discriminant, a method used in statistics and other fields, to find a linear combination of features that characterizes or separates two or more classes of objects or events.

upvoted 4 times

🗨️ 👤 **itallond** 1 year, 7 months ago

L1 won't do naturally the feature elimination?

I guess AB

upvoted 1 times

🗨️ 👤 **Atreides457** 1 year, 10 months ago

why not A & D? or C & D?

does not t-SNE grant the marketing team's wish for visualization of relationships? or are we to presume that A&C are best as C (recursive feature elimination) grants us some visualization of feature importance.

upvoted 2 times

🗨️ 👤 **tgaos** 2 years ago

Selected Answer: AC

AC is correct

upvoted 3 times

🗨️ 👤 **NeverMinda** 2 years ago

Selected Answer: AC

overfitting: add regularization, remove features

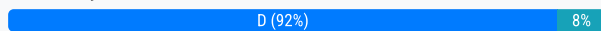
upvoted 4 times

An aircraft engine manufacturing company is measuring 200 performance metrics in a time-series. Engineers want to detect critical manufacturing defects in near- real time during testing. All of the data needs to be stored for offline analysis. What approach would be the MOST effective to perform near-real time defect detection?

- A. Use AWS IoT Analytics for ingestion, storage, and further analysis. Use Jupyter notebooks from within AWS IoT Analytics to carry out analysis for anomalies.
- B. Use Amazon S3 for ingestion, storage, and further analysis. Use an Amazon EMR cluster to carry out Apache Spark ML k-means clustering to determine anomalies.
- C. Use Amazon S3 for ingestion, storage, and further analysis. Use the Amazon SageMaker Random Cut Forest (RCF) algorithm to determine anomalies.
- D. Use Amazon Kinesis Data Firehose for ingestion and Amazon Kinesis Data Analytics Random Cut Forest (RCF) to perform anomaly detection. Use Kinesis Data Firehose to store data in Amazon S3 for further analysis.

Suggested Answer: B

Community vote distribution



Joe_Zhang Highly Voted 3 years, 9 months ago

D near-real time
upvoted 43 times

DimLam 1 year, 8 months ago

The main problem with D is that Amazon Kinesis Data Firehose can not be a source service for Amazon Kinesis Data Analytics.

The answer would be correct if it said

"Using Amazon Kinesis Data Stream to ingest data, using Amazon Kinesis Data Analytics for defect detection and using Amazon Kinesis Data Firehose for storing data for further Analysis"

<https://docs.aws.amazon.com/firehose/latest/dev/create-name.html>

upvoted 2 times

VR10 1 year, 4 months ago

Actually Kinesis Data Firehose can be used for Data Ingestion.

So the correct option is still D

upvoted 2 times

cnethers Highly Voted 3 years, 9 months ago

Glad we are all in agreement D is the correct answer

upvoted 17 times

xicoaio Most Recent 9 months ago

Selected Answer: D

Amazon Kinesis Data Firehose is a fully managed service for real-time data ingestion, which fits the requirement for near-real-time defect detection. It can ingest large volumes of data from various sources and reliably load the data into other AWS services like Amazon S3 for storage.

Amazon Kinesis Data Analytics with Random Cut Forest (RCF) is highly efficient for detecting anomalies in streaming data in near real time, which is what the engineers need to catch manufacturing defects during testing.

After detecting anomalies, the data can be stored in Amazon S3 via Kinesis Data Firehose for offline analysis.

upvoted 1 times

SandyHenshaw 11 months, 1 week ago

Selected Answer: D

D - firehose for near realtime

upvoted 1 times

VR10 1 year, 4 months ago

Selected Answer: D

Kinesis Data Firehose is a fully managed service that can ingest streaming data and load it into destinations like S3, Redshift, Elasticsearch. and with Kinesis Data Analytics and RCF and then Data Firehose again to store on S3.

D is the best choice.

upvoted 1 times

🗳️ 👤 **fa0d8b7** 1 year, 6 months ago

<https://docs.aws.amazon.com/managed-flink/latest/java/get-started-exercise-fh.html>

upvoted 2 times

🗳️ 👤 **endeesa** 1 year, 7 months ago

Selected Answer: D

Kinesis seems like the only viable option

upvoted 1 times

🗳️ 👤 **akgarg00** 1 year, 7 months ago

The answer is D. Since, data is continuously coming in Kinesis datafirehose is our streaming application (also we need near Real time defect detection and storage in S3) and anomaly detection can be done by kinesis data application (RCF algorithm).

upvoted 1 times

🗳️ 👤 **AmeeraM** 1 year, 8 months ago

Selected Answer: D

D, near real-time ingestion is the key

upvoted 1 times

🗳️ 👤 **loict** 1 year, 9 months ago

Selected Answer: D

A. NO - AWS IoT will first store the data, then make it available for Analytics/Jupyter (<https://docs.aws.amazon.com/iotanalytics/latest/userguide/welcome.html>); so not real-time

B. NO - not realtime to store the data before analytics

C. NO - not realtime to store the data before analytics

D. YES - real-time pipe, RCF best for anomalies

upvoted 1 times

🗳️ 👤 **DavidRou** 1 year, 9 months ago

Selected Answer: D

How can someone use S3 for ingestion? Firehose is the right answer

upvoted 1 times

🗳️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: D

This option meets the requirements of performing near-real time defect detection, storing all the data for offline analysis, and handling 200 performance metrics in a time-series. Amazon Kinesis Data Firehose is a fully managed service that can ingest streaming data from various sources and deliver it to destinations such as Amazon S3, Amazon OpenSearch Service, and Amazon Redshift. Amazon Kinesis Data Analytics is a service that can process streaming data using SQL or Apache Flink applications. Amazon Kinesis Data Analytics provides a built-in RANDOM_CUT_FOREST function, a machine learning algorithm that can detect anomalies in streaming data¹. This function can handle high-dimensional data and assign an anomaly score to each record based on how distant it is from other records¹. The anomaly scores can then be delivered to another destination using Kinesis Data Firehose or consumed by other applications using Kinesis Data Streams.

upvoted 1 times

🗳️ 👤 **kaike_reis** 1 year, 11 months ago

D is the correct

If the question says "data streaming", "real time data" or "near real time" you should look for kinesis services.

B and C are totally wrong: It's not possible to use S3 to ingestion, only storage.

upvoted 2 times

🗳️ 👤 **ADVIT** 1 year, 12 months ago

D,

<https://docs.aws.amazon.com/kinesisanalytics/latest/sqlref/sqlrf-random-cut-forest.html>

upvoted 1 times

🗳️ 👤 **earthMover** 2 years, 1 month ago

Selected Answer: D

At a minimum the moderators should put some explanation when the community vote overwhelmingly for a different option.

upvoted 4 times

🗨️ 👤 **oso0348** 2 years, 2 months ago

Selected Answer: C

Option D is not necessarily incorrect, but it may not be the most effective approach to perform near-real time defect detection in this scenario. Here are some potential drawbacks of this approach:

Amazon Kinesis Data Firehose is primarily used for data ingestion and delivery to other services, and may not be the best choice for real-time analysis.

Using Amazon Kinesis Data Analytics for anomaly detection may be less flexible than using Amazon SageMaker, which provides a wide range of algorithms and models for anomaly detection.

Random Cut Forest (RCF) is a popular anomaly detection algorithm used for time-series data, and Amazon SageMaker provides an RCF implementation that can be used for anomaly detection in real-time or offline. While Amazon Kinesis Data Analytics also provides RCF, using Amazon SageMaker may be a better choice for scalability and flexibility.

upvoted 1 times

🗨️ 👤 **oso0348** 2 years, 2 months ago

Selected Answer: C

Yes, option C can provide near real-time defect detection. Amazon SageMaker's Random Cut Forest (RCF) algorithm is designed to work with streaming data and can detect anomalies in near real-time. It can process data in batches as small as a single data point, making it well-suited for real-time anomaly detection.

In this scenario, if the manufacturing process is generating data in real-time, it can be ingested into Amazon S3 and processed by Amazon SageMaker's RCF algorithm, allowing for near real-time detection of critical manufacturing defects during testing.

upvoted 1 times

🗨️ 👤 **ZSun** 2 years, 2 months ago

this is ridiculous. How can you store in s3 and then conduct real-time analysis?

upvoted 2 times

A Machine Learning team runs its own training algorithm on Amazon SageMaker. The training algorithm requires external assets. The team needs to submit both its own algorithm code and algorithm-specific parameters to Amazon SageMaker.

What combination of services should the team use to build a custom algorithm in Amazon SageMaker? (Choose two.)

- A. AWS Secrets Manager
- B. AWS CodeStar
- C. Amazon ECR
- D. Amazon ECS
- E. Amazon S3

Suggested Answer: CE

Community vote distribution

CE (100%)

Paul_NoName Highly Voted 3 years, 9 months ago

CE is the right answer. ECR uses ECS internally while using SGM.
upvoted 11 times

[Removed] 3 years, 8 months ago

CE based on criteria and this documentation: <https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-mkt-create-model-package.html>

"For Location of inference image, type the path to the image that contains your inference code. The image must be stored as a Docker container in Amazon ECR.

For Location of model data artifacts, type the location in S3 where your model artifacts are stored."

upvoted 6 times

ZSun 2 years, 2 months ago

the answer is correct but the explanation is completely wrong.

The question is about how to create your own algorithm using container, not "put the inference in market" (which is your resource link).

the right citation should be "Adapting your own training container": create s3 to store model artifact, and push code to ECR.

upvoted 3 times

SophieSu Highly Voted 3 years, 8 months ago

CE IS THE CORRECT ANSWER 100%

upvoted 5 times

MultiCloudIronMan Most Recent 9 months, 1 week ago

Selected Answer: CE

Amazon ECR (Option C): Amazon Elastic Container Registry (ECR) is used to store, manage, and deploy Docker container images. The team can package their custom algorithm code into a Docker container and store it in Amazon ECR1.

Amazon S3 (Option E): Amazon Simple Storage Service (S3) is used to store external assets and data. The team can store the algorithm-specific parameters and any other required data in Amazon S3

upvoted 1 times

Mickey321 1 year, 10 months ago

Selected Answer: CE

Amazon ECR is a fully managed container registry service that allows users to store, manage, and deploy Docker container images. Amazon SageMaker supports using custom Docker images for training and inference, which can contain the user's own training algorithm and any external assets or dependencies. The user can push their Docker image to Amazon ECR and then reference it in their Amazon SageMaker training job configuration.

upvoted 1 times

jackzhao 2 years, 3 months ago

CE is correct!

upvoted 1 times

Valcilio 2 years, 3 months ago

ECR for the code, S3 for the parameters!

upvoted 1 times

🗨️ 👤 **Valcilio** 2 years, 3 months ago

Selected Answer: CE

C contain the algorithm's image and E contain algorithm's parameters.

upvoted 2 times

🗨️ 👤 **randomnamer** 3 years, 7 months ago

The location of the model artifacts. Model artifacts can either be packaged in the same Docker container as the inference code or stored in Amazon S3. Not so sure.

upvoted 1 times

🗨️ 👤 **cnethers** 3 years, 8 months ago

<https://aws.amazon.com/blogs/machine-learning/bringing-your-own-custom-container-image-to-amazon-sagemaker-studio-notebooks/>

If you wish to use your private VPC to securely bring your custom container, you also need the following:

A VPC with a private subnet

VPC endpoints for the following services:

Amazon Simple Storage Service (Amazon S3)

Amazon SageMaker

Amazon ECR

AWS Security Token Service (AWS STS)

CodeBuild for building Docker containers

Answer C+E

upvoted 3 times

🗨️ 👤 **ahquiceno** 3 years, 9 months ago

For me CD. needs storage and create a custom docker using ECR to store it.

upvoted 1 times

🗨️ 👤 **ahquiceno** 3 years, 8 months ago

Sorry, CE is correct.

upvoted 1 times

🗨️ 👤 **gcpwhiz** 3 years, 7 months ago

Sagemaker will spin up the instances needed with the right image. No need to use ECS. CE is right

upvoted 1 times

A Machine Learning Specialist wants to determine the appropriate SageMakerVariantInvocationsPerInstance setting for an endpoint automatic scaling configuration. The Specialist has performed a load test on a single instance and determined that peak requests per second (RPS) without service degradation is about 20 RPS. As this is the first deployment, the Specialist intends to set the invocation safety factor to 0.5. Based on the stated parameters and given that the invocations per instance setting is measured on a per-minute basis, what should the Specialist set as the SageMakerVariantInvocationsPerInstance setting?

- A. 10
- B. 30
- C. 600
- D. 2,400

Suggested Answer: C

Community vote distribution

C (100%)

 **Paul_NoName** Highly Voted 2 years, 8 months ago

C is correct .

$\text{SageMakerVariantInvocationsPerInstance} = (\text{MAX_RPS} * \text{SAFETY_FACTOR}) * 60$

AWS recommended Saf_fac = 0.5

upvoted 14 times

 **ahquiceno** Highly Voted 2 years, 8 months ago

Answer C: $\text{SageMakerVariantInvocationsPerInstance} = (\text{MAX_RPS} * \text{SAFETY_FACTOR}) * 60$

<https://docs.aws.amazon.com/sagemaker/latest/dg/endpoint-scaling-loadtest.html>

upvoted 6 times

 **Mickey321** Most Recent 10 months ago

Selected Answer: C

To calculate the SageMakerVariantInvocationsPerInstance setting, we can use the following equation from the web search results1:

$\text{SageMakerVariantInvocationsPerInstance} = (\text{MAX_RPS} * \text{SAFETY_FACTOR}) * 60$

Where MAX_RPS is the maximum RPS that the variant can handle, SAFETY_FACTOR is the safety factor that we choose to ensure that we don't exceed the maximum RPS, and 60 is to convert from RPS to invocations-per-minute.


Plugging in the given values, we get:

$\text{SageMakerVariantInvocationsPerInstance} = (20 * 0.5) * 60$ $\text{SageMakerVariantInvocationsPerInstance} = 10 * 60$

$\text{SageMakerVariantInvocationsPerInstance} = 600$

Therefore, the Specialist should set the SageMakerVariantInvocationsPerInstance setting to 600.

upvoted 1 times

 **jackzhao** 1 year, 3 months ago

$\text{SageMakerVariantInvocationsPerInstance} = (\text{MAX_RPS} * \text{SAFETY_FACTOR}) * 60$

upvoted 1 times

 **King_Chess1** 1 year, 5 months ago

$\text{SageMakerVariantInvocationsPerInstance} = (\text{MAX_RPS} * \text{SAFETY_FACTOR}) * 60$

$(20\text{RPS} * 0.5\text{Safety Factor}) * 60$

$(10)*60 = 600$

Answer C

upvoted 1 times

 **Peeking** 1 year, 6 months ago

Selected Answer: C

Maximum request at peak time = 20 RPS = $20 \times 60 = 1200\text{RPM}$

Safety factor of 0.5 = $1200 \times 0.5 = 600$

Basic setting of parameter = 600 (requests per minutes)

upvoted 1 times

A company uses a long short-term memory (LSTM) model to evaluate the risk factors of a particular energy sector. The model reviews multi-page text documents to analyze each sentence of the text and categorize it as either a potential risk or no risk. The model is not performing well, even though the Data Scientist has experimented with many different network structures and tuned the corresponding hyperparameters.

Which approach will provide the MAXIMUM performance boost?

- A. Initialize the words by term frequency-inverse document frequency (TF-IDF) vectors pretrained on a large collection of news articles related to the energy sector.
- B. Use gated recurrent units (GRUs) instead of LSTM and run the training process until the validation loss stops decreasing.
- C. Reduce the learning rate and run the training process until the training loss stops decreasing.
- D. Initialize the words by word2vec embeddings pretrained on a large collection of news articles related to the energy sector.

Suggested Answer: C

Community vote distribution

D (100%)

 **jiadong** Highly Voted 3 years, 9 months ago

I think the right answer is D
upvoted 24 times

 **SophieSu** Highly Voted 3 years, 9 months ago

D is correct.
C is not the best the answer because the question states that tuning parameters doesn't help a lot. Transfer learning would be better solution!
upvoted 11 times

 **xicocaio** Most Recent 9 months ago


Selected Answer: D
Using word2vec embeddings would give the model more accurate representations of words at the start, potentially leading to a significant performance boost for text classification tasks.
upvoted 1 times

 **ninomfr64** 1 year ago

Selected Answer: D
A. NO, transfer learning helps, word2vec > TD-ITF as the first keeps into account part of the word context (there is a hyperparameter for this)
B. LTSM delivers better results wrt GRU which is in turn a compromise architecture to balance accuracy with training time/cost
C. Heperparameters tuning has been already applied, this will not help
D. YES, transfer learning will help and word3vec is better option in this scenario
upvoted 2 times

 **3eb0542** 1 year, 4 months ago

Selected Answer: D
How are the 'correct' answers being provided? I'm seeing so many answers that seem to be wrong and usually, the community vote seems to be correct. This is kind of frustrating.
upvoted 3 times

 **Mickey321** 1 year, 10 months ago

Selected Answer: D
Word2vec is a technique that can learn distributed representations of words, also known as word embeddings, from large amounts of text data. Word embeddings can capture the semantic and syntactic similarities and relationships between words, and can be used as input features for neural network models. Word2vec can be trained on domain-specific corpora to obtain more relevant and accurate word embeddings for a particular task.
upvoted 3 times

 **kaike_reis** 1 year, 11 months ago

Selected Answer: D
From my perspective, B and C are wrong because the DS already tried something close to this. D is correct.
upvoted 1 times

 **vbal** 2 years, 1 month ago

I don't think High Dimensionality is take care by C2V; TF-IDF is required. A.

upvoted 1 times

🗳️ 👤 **Peeking** 2 years, 6 months ago

Selected Answer: D

Transfer learning, in my experience, has been a good way to boost performance when hyperparameter tuning did not work.

upvoted 2 times

🗳️ 👤 **Sidekick** 3 years, 1 month ago

The case ask for predicting labels for sentences, the appropriate algo should be "Text Classification" Which, just as "word2vec,i part of Blazing Text.

upvoted 1 times

🗳️ 👤 **julpeg** 3 years, 2 months ago

Selected Answer: D

The answer should be D. My reasoning is that by using a word embedding which is trained on domain specific material, the embeddings between two words are more domain specific. This means that relations (good or bad) are represented in a better way, which also means that the model should be able to predict the results in a more accurate way.

upvoted 3 times

🗳️ 👤 **bitsplease** 3 years, 5 months ago

both A & D "seem" correct, but word2vec takes ORDER of words into acc (to some extent)--while TF-IDF does not. Thus max boost is from D.

B,C are wrong because the DS has tried several network architectures (aka LSTM) and hyperparameter tuning (aka option C)

upvoted 6 times

🗳️ 👤 **ahmedelbhy** 3 years, 8 months ago

i think answer is A as The model reviews multi-page text documents

upvoted 1 times

🗳️ 👤 **GiyeonShin** 2 years, 6 months ago

I think that the general tf-idf vectors cannot be directly adapted to the deep learning model, because of the large dimension in vector values

upvoted 1 times

🗳️ 👤 **puffpuff** 3 years, 8 months ago

I think it should be B

A/D are false flags because the question doesn't specify what kind of data engineering is currently done on the inputs, as a baseline

Per wikipedia, for GRUs, "GRUs have been shown to exhibit better performance on certain smaller and less frequent datasets", which fits the context of a particular energy sector

upvoted 2 times

🗳️ 👤 **ChanduPatil** 3 years, 8 months ago

why not B??

upvoted 1 times

🗳️ 👤 **GiyeonShin** 2 years, 6 months ago

Generally, LSTM has the better performance then GRU in large datasets such as multi-page documents. GRU has advantages of memory allocation and training time.

upvoted 1 times

🗳️ 👤 **GiyeonShin** 2 years, 6 months ago

Early stopping can give the model better performance, but I think that the model needs more condition like patience value for early stopping. This is because the model doesn't always show the performance at its maximum when the validation loss stops decreasing.

upvoted 1 times

🗳️ 👤 **jkreddy** 3 years, 8 months ago

It cannot be C, because hyper parameter tuning didnt work as given in question. Also, A and D are same, however, word2vec model internally implements tf-idf much more efficiently. So answer got to be D

upvoted 4 times

🗳️ 👤 **YJ4219** 3 years, 8 months ago

but they need to classify the whole sentence i think for such a case we use object2vec not word2vec, but since it's not available in the answers, B is the only answer left.

upvoted 2 times

🗳️ 👤 **tmlid** 3 years, 8 months ago

I go for C

upvoted 2 times

A Machine Learning Specialist needs to move and transform data in preparation for training. Some of the data needs to be processed in near-real time, and other data can be moved hourly. There are existing Amazon EMR MapReduce jobs to clean and feature engineering to perform on the data.

Which of the following services can feed data to the MapReduce jobs? (Choose two.)

- A. AWS DMS
- B. Amazon Kinesis
- C. AWS Data Pipeline
- D. Amazon Athena
- E. Amazon ES

Suggested Answer: AE

Community vote distribution

BC (100%)

🗳️ 👤 **Joe_Zhang** Highly Voted 3 years, 3 months ago

should be BC

upvoted 32 times

🗳️ 👤 **[Removed]** 3 years, 2 months ago

Agreed, AWS Example: <https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/what-is-datapipeline.html>

upvoted 6 times

🗳️ 👤 **Denise123** Most Recent 10 months, 1 week ago

It is obviously B and C, I am frustrated with the number of wrong answers. Why the moderator's answers keep being super weird?

upvoted 2 times

🗳️ 👤 **Mickey321** 1 year, 4 months ago

Selected Answer: BC

B for near real-time

C for hourly

upvoted 2 times

🗳️ 👤 **Khalil11** 1 year, 8 months ago

The right answer is BC

upvoted 2 times

🗳️ 👤 **Ajose0** 1 year, 10 months ago

Selected Answer: BC

AWS Data Pipeline (Option C) can be used to move the hourly data, as it provides a way to move data from various sources to Amazon EMR for processing.

Amazon Kinesis (Option B) can be used to process data in near-real time, as it is a real-time data streaming service that can handle large amounts of incoming data from multiple sources. The data can be fed to Amazon EMR MapReduce jobs for processing.

upvoted 4 times

🗳️ 👤 **DS2021** 2 years ago

Selected Answer: BC

should be BC

upvoted 2 times

🗳️ 👤 **Peeking** 2 years ago

Selected Answer: BC

Kinesis for near realtime data and pipeline for the other data moved hourly.

upvoted 3 times

🗳️ 👤 **John_Pongthorn** 2 years, 10 months ago

Selected Answer: BC

AWS ES is an elastic search , it is nothing to do with this question.

upvoted 3 times

🗨️ **scsas** 2 years, 10 months ago

Kinesis data into EMR: <https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-kinesis.html>

upvoted 1 times

🗨️ **apprehensive_scar** 2 years, 11 months ago

BC. easy

upvoted 2 times

🗨️ **KM226** 2 years, 12 months ago

I believe the answer is BC

upvoted 1 times

🗨️ **Madwyn** 3 years, 2 months ago

BD.

Data Pipeline is to orchestrate the workflow, how can that feed data to the MR jobs?

upvoted 2 times

🗨️ **Vita_Rasta84444** 3 years, 2 months ago

Answer is B and C

upvoted 1 times

🗨️ **astonm13** 3 years, 2 months ago

Answer is for sure BC

upvoted 3 times

🗨️ **takahirokoyama** 3 years, 2 months ago

Ans is BC.

(<https://aws.amazon.com/jp/emr/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc>)

upvoted 2 times

A Machine Learning Specialist previously trained a logistic regression model using scikit-learn on a local machine, and the Specialist now wants to deploy it to production for inference only.

What steps should be taken to ensure Amazon SageMaker can host a model that was trained locally?

- A. Build the Docker image with the inference code. Tag the Docker image with the registry hostname and upload it to Amazon ECR.
- B. Serialize the trained model so the format is compressed for deployment. Tag the Docker image with the registry hostname and upload it to Amazon S3.
- C. Serialize the trained model so the format is compressed for deployment. Build the image and upload it to Docker Hub.
- D. Build the Docker image with the inference code. Configure Docker Hub and upload the image to Amazon ECR.

Suggested Answer: D

Community vote distribution

A (100%)

🗳️ **arulrajjayaraj** Highly Voted 2 years, 8 months ago

Ans : A Refer the below :

<https://sagemaker-workshop.com/custom/containers.html>

upvoted 19 times

🗳️ **Paul_NoName** Highly Voted 2 years, 9 months ago

A

<https://sagemaker-workshop.com/custom/containers.html>

upvoted 7 times

🗳️ **endeesa** Most Recent 7 months ago

Selected Answer: A

You need the container to be hosted on ECR.

upvoted 1 times

🗳️ **loict** 9 months, 2 weeks ago

Selected Answer: A

A. YES - the inference code is built after inspecting the coefficient of the Linear Model (or, alternatively, the model can be serialized via pickle and the inference code is simply to unserialize the model); ECR is only registry supported by SageMaer; tagging the Docker image with the registry hostname (eg. docker tag image1 public.ecr.aws/g6h7x5m5/image1) is required so that the docker push command knows where to push the image

B. NO - no need to compress; image must be on ECR

C. NO - no need to compress; image must be on ECR

D. NO - image must be on ECR

upvoted 5 times

🗳️ **Khalil11** 1 year, 2 months ago

Selected Answer: A

A is the right answer

upvoted 3 times

🗳️ **Nadia0012** 1 year, 3 months ago

Selected Answer: A

For SageMaker to run a container for training or hosting, it needs to be able to find the image hosted in the image repository, Amazon Elastic Container Registry (Amazon ECR). The three main steps to this process are building locally, tagging with the repository location, and pushing the image to the repository.

upvoted 4 times

🗳️ **geekgirl007** 2 years, 5 months ago

Selected Answer: A


A for sure.

upvoted 3 times

🗳️ **astonm13** 2 years, 8 months ago

Answer is A.

upvoted 3 times

  **cnethers** 2 years, 8 months ago

Docker Hub is a repository so ANS D makes no sense. Option A is the way to go.

upvoted 2 times

A trucking company is collecting live image data from its fleet of trucks across the globe. The data is growing rapidly and approximately 100 GB of new data is generated every day. The company wants to explore machine learning uses cases while ensuring the data is only accessible to specific IAM users.

Which storage option provides the most processing flexibility and will allow access control with IAM?

- A. Use a database, such as Amazon DynamoDB, to store the images, and set the IAM policies to restrict access to only the desired IAM users.
- B. Use an Amazon S3-backed data lake to store the raw images, and set up the permissions using bucket policies.
- C. Setup up Amazon EMR with Hadoop Distributed File System (HDFS) to store the files, and restrict access to the EMR instances using IAM policies.
- D. Configure Amazon EFS with IAM policies to make the data available to Amazon EC2 instances owned by the IAM users.

Suggested Answer: C

Community vote distribution

B (81%)

C (19%)

🗳️ 👤 **Paul_NoName** Highly Voted 👍 3 years, 8 months ago

B is the right answer

upvoted 33 times

🗳️ 👤 **clawo** Highly Voted 👍 3 years, 5 months ago

Selected Answer: B

B to use as storage with policies

upvoted 7 times

🗳️ 👤 **xicocaio** Most Recent 🕒 9 months ago

Selected Answer: B

- Amazon S3-backed data lake: S3 is the best storage option for large and rapidly growing datasets like images from trucks. S3 scales easily, handles large volumes of data, and is cost-effective for long-term storage, making it a natural choice for this scenario.

- IAM access control: You can use bucket policies in S3 to set very specific access controls, ensuring that only certain IAM users have permission to access or modify the data. This satisfies the requirement for access control using IAM.

- Processing flexibility: Storing the images in S3 offers flexibility for future machine learning use cases. The data stored in S3 can easily be integrated with other AWS services like SageMaker, Athena, EMR, and more for processing and analysis.

upvoted 1 times

🗳️ 👤 **endeesa** 1 year, 7 months ago

Selected Answer: B

EMR/HDFS is not more 'flexible' than S3

upvoted 1 times

🗳️ 👤 **loict** 1 year, 9 months ago

Selected Answer: B

A. NO - volume too big for a DB

B. YES

C. NO - instance access will not control HDFS access

D. NO - EFS does not use IAM policies (it is unix)

upvoted 1 times

🗳️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: B

S3 indeed

upvoted 1 times

🗳️ 👤 **JK1977** 2 years, 1 month ago

Selected Answer: B

S3 always

upvoted 1 times

🗨️ 👤 **Nadia0012** 2 years, 3 months ago

Selected Answer: B

I would say the answer is B not because of the cost on EMR, that is also a current answer. however: "most processing flexibility" indicates that S3 is a better option. because all ML solutions and work flows integrate with S3. it hasn't spoken what the ML solution and which services so I take the safe side and go with S3

upvoted 2 times

🗨️ 👤 **KlaudYu** 2 years, 11 months ago

Selected Answer: B

C is not affordable because it is ephemeral storage. <https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-file-systems.html>

"HDFS is used by the master and core nodes. One advantage is that it's fast; a disadvantage is that it's ephemeral storage which is reclaimed when the cluster ends. It's best used for caching the results produced by intermediate job-flow steps."

upvoted 4 times

🗨️ 👤 **ZSun** 2 years, 2 months ago

the question does not require long-term storage.

upvoted 1 times

🗨️ 👤 **geekgirl007** 3 years, 5 months ago

Selected Answer: C

C is correct. it says real time data and to be used for ml process so EMR more suitable. also S3 bucket policies not same as IAM users so B is not correct.

upvoted 4 times

🗨️ 👤 **ovokpus** 3 years ago

Why will you need to spin up servers (EMR) just to store visual data for ML?

upvoted 5 times

🗨️ 👤 **Abdo702** 3 years, 7 months ago

I think Amazon EMR is more appropriate, as the data scheme stated is a big data scheme.

<https://aws.amazon.com/emr/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc>

upvoted 3 times

🗨️ 👤 **Sourabh1703** 3 years, 5 months ago

IAM support is required for storage feature , that is not possible as per options described as IAM is supported for HDFS for the instance running on top of it, hence B should be correct

upvoted 2 times

🗨️ 👤 **Vita_Rasta84444** 3 years, 8 months ago

B is the right answer

upvoted 2 times

🗨️ 👤 **srinu3054** 3 years, 8 months ago

S3 is the easy, scalable and secure option to store the image data.

upvoted 1 times

🗨️ 👤 **astonm13** 3 years, 8 months ago

B is the right answer

upvoted 1 times

🗨️ 👤 **zzaibis** 3 years, 8 months ago

B is an appropriate choice

upvoted 2 times

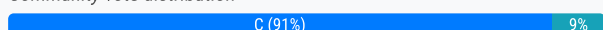
A credit card company wants to build a credit scoring model to help predict whether a new credit card applicant will default on a credit card payment. The company has collected data from a large number of sources with thousands of raw attributes. Early experiments to train a classification model revealed that many attributes are highly correlated, the large number of features slows down the training speed significantly, and that there are some overfitting issues.

The Data Scientist on this project would like to speed up the model training time without losing a lot of information from the original dataset. Which feature engineering technique should the Data Scientist use to meet the objectives?

- A. Run self-correlation on all features and remove highly correlated features
- B. Normalize all numerical values to be between 0 and 1
- C. Use an autoencoder or principal component analysis (PCA) to replace original features with new features
- D. Cluster raw data using k-means and use sample data from each cluster to build a new dataset

Suggested Answer: B

Community vote distribution



ahquiceno Highly Voted 3 years, 9 months ago

Answer C. Need reduce the features preserving the information on it this is achieve using PCA.
upvoted 26 times

Dr_Kiko 3 years, 7 months ago
without losing a lot of information from the original dataset
since when PCA retains information?
upvoted 3 times

VinceCar 2 years, 7 months ago
PCA helps to speed up the training
upvoted 4 times

[Removed] Highly Voted 3 years, 8 months ago

Answer is A, because one must avoid information loss that PCA or autoencoders introduce through new features (<https://www.i2tutorials.com/what-are-the-pros-and-cons-of-the-pca/>). Otherwise, I would perform C.
upvoted 6 times

SophieSu 3 years, 8 months ago
If you REMOVE highly correlated features(that means in pairs), the model lost a lot of information.
upvoted 4 times

rodrigus 2 years, 3 months ago
A doesn't have sense. Self-correlation is for times series data, not for pair correlation
upvoted 2 times

xicocaio Most Recent 9 months ago

Selected Answer: A

This question can be misleading.

I would choose A if self-correlation in the dataset is meaning pair-wise correlation, this is the most typical approach in real life. But if self-correlation means auto-correlation as in the time-series treatment, then it is wrong.

Issues with answer C: Autoencoders are notorious for being hard to interpret. With PCA it is possible, but definitely not easy if you have a large dataset. In real life with this scenario, you would always go with pairwise correlation as the most simple yet effective approach.
upvoted 1 times

Giodefa96 11 months ago

Selected Answer: C

Answer is C
upvoted 1 times

🗳️ 👤 **geoan13** 1 year, 7 months ago

Answer C

PCA (Principal Component Analysis) takes advantage of multicollinearity and combines the highly correlated variables into a set of uncorrelated variables. Therefore, PCA can effectively eliminate multicollinearity between features.

[https://towardsdatascience.com/how-do-you-apply-pca-to-logistic-regression-to-remove-multicollinearity-10b7f8e89f9b#:~:text=PCA%20\(Principal%20Component%20Analysis\)%20takes,effectively%20eliminate%20multicollinearity%20between%20features.](https://towardsdatascience.com/how-do-you-apply-pca-to-logistic-regression-to-remove-multicollinearity-10b7f8e89f9b#:~:text=PCA%20(Principal%20Component%20Analysis)%20takes,effectively%20eliminate%20multicollinearity%20between%20features.)

upvoted 1 times

🗳️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: C

Option C

upvoted 1 times

🗳️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: C

An autoencoder is a type of neural network that can learn a compressed representation of the input data, called the latent space, by encoding and decoding the data through multiple hidden layers¹. PCA is a statistical technique that can reduce the dimensionality of the data by finding a set of orthogonal axes, called the principal components, that capture the most variance in the data². Both methods can transform the original features into new features that are lower-dimensional, uncorrelated, and informative.

upvoted 1 times

🗳️ 👤 **kaike_reis** 1 year, 11 months ago

Selected Answer: C

C is the correct.

Self-correlation is for time series, which is not mention here. Besides that, even if was correlation only, try to do this in thousand features...

upvoted 1 times

🗳️ 👤 **vbal** 2 years, 1 month ago

A . run correlation matrix and remove highly correlated features.

upvoted 1 times

🗳️ 👤 **JK1977** 2 years, 1 month ago

Selected Answer: C

PCA for feature reduction

upvoted 1 times

🗳️ 👤 **GOSD** 2 years, 1 month ago

is it just me or is every 15th answer here PCA?

upvoted 2 times

🗳️ 👤 **oso0348** 2 years, 2 months ago

Selected Answer: C

Using an autoencoder or PCA can help reduce the dimensionality of the dataset by creating new features that capture the most important information in the original dataset while discarding some of the noise and highly correlated features. This can help speed up the training time and reduce overfitting issues without losing a lot of information from the original dataset. Option A may remove too many features and may not capture all the important information in the dataset, while option B only rescales the data and does not address the issue of highly correlated features. Option D is not a feature engineering technique and may not be an effective way to reduce the dimensionality of the dataset.

upvoted 1 times

🗳️ 👤 **Paolo991** 2 years, 3 months ago

Selected Answer: C

PCA builds new features starting from high correlated ones. So it matches the question

upvoted 1 times

🗳️ 👤 **Sneep** 2 years, 5 months ago

It's C.

The Data Scientist should use principal component analysis (PCA) to replace the original features with new features. PCA is a technique that reduces the dimensionality of a dataset by projecting it onto a lower-dimensional space, while preserving as much of the original variation as possible. This can help to speed up the training time of the model and reduce overfitting issues, without losing a significant amount of information from the original dataset.

upvoted 1 times

🗳️ 👤 **Aninina** 2 years, 5 months ago

Selected Answer: C



C: PCA is the solution
upvoted 1 times

  **ovokpus** 3 years ago

Selected Answer: C

Correction to C. Removing correlated features from hundreds of columns will be tedious and time consuming. PCA is the way to go here.

Apologies for the flip
upvoted 2 times

  **ovokpus** 3 years ago

Selected Answer: A

Answer is A. Eliminate features that are highly correlated. This will not compromise the quality of the feature space as much as PCA would.
upvoted 1 times

A Data Scientist is training a multilayer perception (MLP) on a dataset with multiple classes. The target class of interest is unique compared to the other classes within the dataset, but it does not achieve an acceptable recall metric. The Data Scientist has already tried varying the number and size of the MLP's hidden layers, which has not significantly improved the results. A solution to improve recall must be implemented as quickly as possible.

Which techniques should be used to meet these requirements?

- A. Gather more data using Amazon Mechanical Turk and then retrain
- B. Train an anomaly detection model instead of an MLP
- C. Train an XGBoost model instead of an MLP
- D. Add class weights to the MLP's loss function and then retrain

Suggested Answer: C

Community vote distribution

D (100%)

🗳️ 👤 **[Removed]** Highly Voted 👍 2 years, 9 months ago

For me answer is D, adjust to higher weight for class of interest: <https://androidkt.com/set-class-weight-for-imbalance-dataset-in-keras/>. More data may/may not be available and a data labeling job will take time.

upvoted 36 times

🗳️ 👤 **rhuanca** Highly Voted 👍 2 years, 1 month ago

I believe is C, because we already made all changes possible in MLP hidden layers and the results have not improved then we must change model so XGBoost seems the best option

upvoted 5 times

🗳️ 👤 **Mickey321** Most Recent 🕒 10 months ago

Selected Answer: D

In this case, the data scientist is training a multilayer perceptron (MLP), which is a type of neural network, on a dataset with multiple classes. The target class of interest is unique compared to the other classes within the dataset, but it does not achieve an acceptable recall metric. Recall is a measure of how well the model can identify the relevant examples from the minority class. The data scientist has already tried varying the number and size of the MLP's hidden layers, which has not significantly improved the results. A solution to improve recall must be implemented as quickly as possible.

upvoted 2 times

🗳️ 👤 **kaike_reis** 11 months ago

Selected Answer: D

The fastest one is D

upvoted 1 times

🗳️ 👤 **ADVIT** 12 months ago

"quickly as possible" mean do not change to new stuff, so it's D.

upvoted 1 times

🗳️ 👤 **kukreti18** 1 year ago

Not C, as the question ask for a quick solution.

I accept D.

upvoted 1 times

🗳️ 👤 **vbal** 1 year, 1 month ago

Answer C : <https://towardsdatascience.com/boosting-techniques-in-python-predicting-hotel-cancellations-62b7a76ffa6c>

upvoted 1 times

🗳️ 👤 **AjoseO** 1 year, 4 months ago

Selected Answer: D

Adding class weights to the MLP's loss function balances the class frequencies in the cost function during training, so the optimization process focuses more on the underrepresented class, improving recall.

upvoted 3 times

🗨️ 👤 **Tomatoteacher** 1 year, 5 months ago

Selected Answer: D

I have done this before, class weights help with unbalanced data. Only logical solution that would help if not done, XGBoost could be different, but who knows, both NNs and XGBoost have comparable performance. Answer D!

upvoted 4 times

🗨️ 👤 **hamuozi** 1 year, 9 months ago

Selected Answer: D

In this example, it is necessary to improve recall as soon as possible, so instead of creating additional datasets, it is effective to change the weight of each class during learning.

upvoted 4 times

🗨️ 👤 **victorlifan** 1 year, 10 months ago

C: 'distinct' indicates we can simplify this as a binary classification problem; then, NN is just overkill. plus, retraining a NN is much slower than training an XGboost model

upvoted 2 times

🗨️ 👤 **exam_prep** 2 years, 1 month ago

I feel answer is B. Question says Target is different than the input data which is hint for anomaly detection.

upvoted 2 times

🗨️ 👤 **kaike_reis** 11 months ago

stop overthink

upvoted 1 times

🗨️ 👤 **KM226** 2 years, 6 months ago

I believe the answer is C because we need to use hyperparameters to improve model performance.

upvoted 2 times

🗨️ 👤 **ksarda11** 2 years, 8 months ago

In case of the quickest possible way, D seems fine. For XGBoost, it will take a bit of time to code again

upvoted 4 times

🗨️ 👤 **ahquiceno** 2 years, 9 months ago

For me Answer A. Why no other model instead xgBoost, the model need more labeled data to be trained and learn more positive examples.

upvoted 2 times

🗨️ 👤 **SophieSu** 2 years, 8 months ago

A is incorrect. Even if you hire Amazon Mechanical Turk, you won't have more data. This question is NOT asking about "labeling".

upvoted 2 times

A Machine Learning Specialist works for a credit card processing company and needs to predict which transactions may be fraudulent in near-real time.

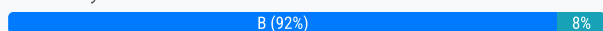
Specifically, the Specialist must train a model that returns the probability that a given transaction may be fraudulent.

How should the Specialist frame this business problem?

- A. Streaming classification
- B. Binary classification
- C. Multi-category classification
- D. Regression classification

Suggested Answer: C

Community vote distribution



ahquiceno Highly Voted 3 years, 9 months ago

Answer B.

upvoted 28 times

SophieSu Highly Voted 3 years, 9 months ago

B IS NOT CORRECT! Return the probability. Not the 1 or 0. D IS THE CORRECT ANSWER.

upvoted 14 times

mdboy93 1 year, 9 months ago

Regression Classification is a made-up term, any binary classifier makes decisions based on probability score.

upvoted 1 times

srinu3054 3 years, 9 months ago

there is nothing like regression classification. (instead it should have said logistic regression). It should be Binary. i.e., either fraud or non fraud. Even with probabilities, we have a threshold to decide the class.

upvoted 11 times

seanLu 3 years, 9 months ago

Logistic regression will give the probability, and logistic regression is a binary classification algorithm.

<https://machinelearningmastery.com/logistic-regression-for-machine-learning/>

upvoted 9 times

Seoyong Most Recent 7 months, 3 weeks ago

Streaming classification: is the process of organizing and categorizing large amounts of data that are continuously flowing. This data can include medical records, banking transactions, and internet records

Binary Classification: Logistic Regression

Multiclass Classification: Softmax regression

Regression Classification is a made-up term

upvoted 1 times

Seoyong 7 months, 3 weeks ago

Random forest is the most suitable model for predicting fraudulent transactions.

Answer is A

upvoted 1 times

ZumbaZim 1 year, 3 months ago

I always see that the community voting is more appropriate and the moderator answer looks

out to be on wrong side. I see this for almost in 1 out of 5 questions. Which answer should we consider here as right one ??

upvoted 1 times

endeesa 1 year, 7 months ago

Selected Answer: B

Its definitely a classification problem, and between Binary and Streaming classification. Binary classification makes more sence
upvoted 1 times

🗨️ **Mickey321** 1 year, 10 months ago

Selected Answer: B

Binary classification
upvoted 1 times

🗨️ **kaike_reis** 1 year, 11 months ago

Selected Answer: B

B, easy.
upvoted 1 times

🗨️ **gusta_dantas** 1 year, 11 months ago

B, obviously!

```
from sklearn.linear_model import LogisticRegression
log_reg = LogisticRegression()
log_reg.fit(X_train, y_train)
log_reg.predict_proba(X_test)
```

=)
upvoted 3 times

🗨️ **rodrigus** 2 years, 3 months ago

The correct solution obviously is binary classification. For the comment above that says that binary classification doesn't returns a probability (for example SVM(classification) only returns a class and logistic, RFClassifier, XGBoostClassifier gives a probability and also a class given a threshold), you should ask yourself if that a regressor model returns always a probability, that is, if there is a restriction in a regressor model to predict values only in $[0,1]$.
upvoted 1 times

🗨️ **Ajose0** 2 years, 4 months ago

Selected Answer: B

The Specialist is trying to determine whether a given transaction is fraudulent or not, which is a binary outcome (yes or no). Therefore, the problem should be framed as binary classification.

The goal is to predict the probability of a transaction being fraudulent or not, and based on that, the Specialist can make a binary decision (fraudulent or not).
upvoted 2 times

🗨️ **Tomatoteacher** 2 years, 5 months ago

Selected Answer: B

This is just binary classification, I don't understand how it could be anything else
upvoted 3 times

🗨️ **Sneep** 2 years, 5 months ago

It's B.

This business problem can be framed as a binary classification problem, where the goal is to predict whether a given transaction is fraudulent (positive class) or not fraudulent (negative class). The model should output a probability for each transaction, indicating the likelihood that it is fraudulent.
upvoted 2 times

🗨️ **DS2021** 2 years, 6 months ago

Selected Answer: D

should be D
upvoted 1 times

🗨️ **RLai** 2 years, 6 months ago

Logistic regression models the probability of the default class (e.g. the first class).

For example, if we are modeling people's sex as male or female from their height, then the first class could be male and the logistic regression model could be written as the probability of male given a person's height.
upvoted 1 times

🗨️ 👤 **theprismdata** 3 years, 1 month ago

I think the answer is B, fraud has various cases which hard to define.

So, Classification result will be fraud or not fraud.

If Multi-category classification, must define case of fraud in detail

upvoted 1 times

🗨️ 👤 **theprismdata** 3 years, 1 month ago

More specifically, anomaly detection model will be needed

upvoted 1 times

🗨️ 👤 **KM226** 3 years, 6 months ago

Selected Answer: B

I believe the answer is B: it a binary classification problem because we are classifying an observation into one of two categories and the target variable in this problem is limited to two options: fraudulent or not fraudulent

upvoted 3 times

🗨️ 👤 **lesh3000** 3 years, 7 months ago

well, regression classification is bullshit, I hope they formulate their questions better on the real exam. binary classification gives probability between 0 and 1

upvoted 3 times

A real estate company wants to create a machine learning model for predicting housing prices based on a historical dataset. The dataset contains 32 features.

Which model will meet the business requirement?

- A. Logistic regression
- B. Linear regression
- C. K-means
- D. Principal component analysis (PCA)

Suggested Answer: B

Community vote distribution

B (100%)

🗲️ 👤 **SophieSu** Highly Voted 👍 2 years, 8 months ago

B is the correct answer.

upvoted 19 times

🗲️ 👤 **Mickey321** Most Recent ⌚ 10 months ago

Selected Answer: B

Linear regression

upvoted 1 times

🗲️ 👤 **kaike_reis** 11 months ago

Selected Answer: B

B, the only model for regression in the options.

upvoted 1 times

🗲️ 👤 **jonsnow777** 2 years, 5 months ago

Selected Answer: B

Answer B

upvoted 2 times

🗲️ 👤 **ahquiceno** 2 years, 9 months ago

Answer B.

upvoted 4 times

A Machine Learning Specialist is applying a linear least squares regression model to a dataset with 1,000 records and 50 features. Prior to training, the ML

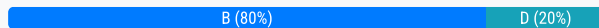
Specialist notices that two features are perfectly linearly dependent.

Why could this be an issue for the linear least squares regression model?

- A. It could cause the backpropagation algorithm to fail during training
- B. It could create a singular matrix during optimization, which fails to define a unique solution
- C. It could modify the loss function during optimization, causing it to fail during training
- D. It could introduce non-linear dependencies within the data, which could invalidate the linear assumptions of the model

Suggested Answer: C

Community vote distribution



Paul_NoName Highly Voted 2 years, 8 months ago

B is correct answer .

upvoted 22 times

hamimelon 1 year, 6 months ago

Agree, B.

upvoted 2 times

pravy 2 years, 8 months ago

why B is the correct answer and not C?

upvoted 1 times

hamimelon 1 year, 6 months ago

For example. If you have two variables, X and Y, and you have two data points. You want to solve the problem: $aX_1 + bY_1 = Z_1$, $aX_2 + bY_2 = Z_2$. However, if $Y=2X \rightarrow Y_1 = 2X_1$, $Y_2 = 2X_2$, then problem becomes: $aX_1 + bY_1 = Z_1$, $a*2X_1 + b*2Y_1 = Z_2 = 2*Z_1$. So you end up with only one function: $aX_1 + bY_1 = Z_1$, meaning there will be more than one answer for (a, b).

If you are familiar with linear algebra, it's easier to express the concept.

upvoted 9 times

SophieSu 2 years, 8 months ago

A square matrix is singular, that is, its determinant is zero, if it contains rows or columns which are proportionally interrelated; in other words, one or more of its rows (columns) is exactly expressible as a linear combination of all or some other its rows (columns), the combination being without a constant term.

upvoted 7 times

Sneep Highly Voted 1 year, 5 months ago

B: If two features in the dataset are perfectly linearly dependent, it means that one feature can be expressed as a linear combination of the other. This can create a singular matrix during optimization, as the linear model would be trying to fit a linear equation to a dataset where one variable is fully determined by the other. This would lead to an ill-defined optimization problem, as there would be no unique solution that minimizes the sum of the squares of the residuals. This could lead to problems during training, as the model would not be able to find appropriate parameter values to fit the data.

upvoted 7 times

Mickey321 Most Recent 10 months ago

Selected Answer: B

Option B

upvoted 1 times

Ajose0 1 year, 4 months ago

Selected Answer: B

The presence of linearly dependent features means that they are redundant, and provide no additional information to the model.

This can result in a matrix that is not invertible, which is a requirement for solving a linear least squares regression problem. The presence of a singular matrix can also cause numerical instability and make it impossible to find an optimal solution to the optimization problem.

upvoted 4 times

🗨️ 👤 **yemauricio** 1 year, 6 months ago

Selected Answer: B

linera dependence creates singular matrix that causes problems at the moment we fit the modle

upvoted 4 times

🗨️ 👤 **wisoxe8356** 1 year, 6 months ago

Selected Answer: B

<https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea>

B - two features are perfectly linearly dependent = singular matrix during optimization

Not D - Not 100% correct (as Multicollinearity happens when independent variables in the regression model are highly correlated to each other) they can still be independent variables

upvoted 3 times

🗨️ 👤 **ovokpus** 2 years ago

Selected Answer: D

Consider one of the 5 assumptions of linear regression. This situation violates the assumption of "No multicollinearity between feature variables"

Hence, D

upvoted 3 times

🗨️ 👤 **jerto97** 2 years, 8 months ago

B. See the multicollinearity problem in wikipedia <https://en.wikipedia.org/wiki/Multicollinearity> (second paragraph)

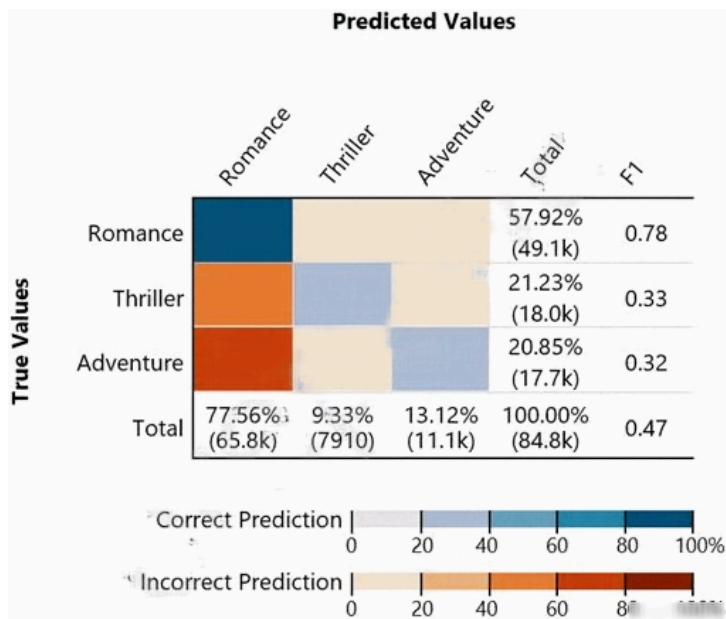
upvoted 4 times

🗨️ 👤 **takahirokoyama** 2 years, 8 months ago

This issue is overfitting.

upvoted 1 times

Given the following confusion matrix for a movie classification model, what is the true class frequency for Romance and the predicted class frequency for Adventure?



- A. The true class frequency for Romance is 77.56% and the predicted class frequency for Adventure is 20.85%
- B. The true class frequency for Romance is 57.92% and the predicted class frequency for Adventure is 13.12%
- C. The true class frequency for Romance is 0.78 and the predicted class frequency for Adventure is (0.47-0.32)
- D. The true class frequency for Romance is 77.56% \div 0.78 and the predicted class frequency for Adventure is 20.85% \div 0.32

Suggested Answer: B

Community vote distribution

B (100%)

SophieSu Highly Voted 2 years, 8 months ago

B is the correct answer. Straightforward!

upvoted 17 times

cnethers Highly Voted 2 years, 8 months ago

<https://docs.aws.amazon.com/machine-learning/latest/dg/multiclass-model-insights.html>

upvoted 10 times

teka112233 Most Recent 10 months, 1 week ago

Selected Answer: B

to be able to understand this Multiclass Model Insights and to be able to answer this question :

True class-frequencies in the evaluation data: The second to last column shows that in the evaluation dataset, 57.92% of the observations in the evaluation data is Romance, 21.23% is Thriller, and 20.85% is Adventure.

Predicted class-frequencies for the evaluation data: The last row shows the frequency of each class in the predictions. 77.56% of the observations is predicted as Romance, 9.33% is predicted as Thriller, and 13.12% is predicted as Adventure.

REF: <https://docs.aws.amazon.com/machine-learning/latest/dg/multiclass-model-insights.html>

upvoted 1 times

Shailendraa 1 year, 9 months ago

12-sep exam

upvoted 2 times

milan_ml 1 year, 11 months ago

Selected Answer: B

The image can be found here:

<https://vceguide.com/what-is-the-true-class-frequency-for-romance-and-the-predicted-class-frequency-for-adventure/>

upvoted 1 times

🗨️ 👤 **obadazx** 1 year, 11 months ago

No image is there!

upvoted 3 times

🗨️ 👤 **SDikeman62** 2 years, 1 month ago

WHy there is no image? Admin. Please fix it.

upvoted 6 times

🗨️ 👤 **Juka3lj** 2 years, 7 months ago

B is correct

upvoted 1 times

🗨️ 👤 **NotAnMLProfessional** 2 years, 8 months ago

A seems to be correct

upvoted 3 times

A Machine Learning Specialist wants to bring a custom algorithm to Amazon SageMaker. The Specialist implements the algorithm in a Docker container supported by Amazon SageMaker.

How should the Specialist package the Docker container so that Amazon SageMaker can launch the training correctly?

- A. Modify the `bash_profile` file in the container and add a bash command to start the training program
- B. Use `CMD` config in the Dockerfile to add the training program as a `CMD` of the image
- C. Configure the training program as an `ENTRYPOINT` named `train`
- D. Copy the training program to directory `/opt/ml/train`

Suggested Answer: B

Community vote distribution

C (82%)

Other

🗳️ **Paul_NoName** Highly Voted 3 years, 3 months ago

C seems correct as per documentations.

upvoted 17 times

🗳️ **[Removed]** Highly Voted 3 years, 3 months ago

I would answer C: <https://docs.aws.amazon.com/sagemaker/latest/dg/your-algorithms-training-algo-dockerfile.html>

"To configure a Docker container to run as an executable, use an `ENTRYPOINT` instruction in a Dockerfile.

SageMaker overrides any default `CMD` statement in a container by specifying the `train` argument after the image name"

upvoted 14 times

🗳️ **awsemort** Most Recent 10 months, 4 weeks ago

Selected Answer: C

I thought it was D, but it is C. It's not D because we copy the `TRAINING` code into `/opt/ml/code/train.py`

upvoted 2 times

🗳️ **MaximusDecimus** 10 months, 4 weeks ago

In Docker, the `ENTRYPOINT` instruction is used to specify the executable that should be run when the container starts. However, Amazon SageMaker expects the training script to be launched by specific commands provided by SageMaker itself, rather than relying solely on the Docker container's `ENTRYPOINT`. The convention for using Docker containers with Amazon SageMaker is to copy the training script and associated resources to specific directories within the container, such as `/opt/ml/code`, and let SageMaker manage the execution of the training process.

I would go with D

upvoted 1 times

🗳️ **cyberfriends** 1 year, 2 months ago

Selected Answer: C

C is correct

upvoted 1 times

🗳️ **Mickey321** 1 year, 4 months ago

Selected Answer: C

C is correct

upvoted 1 times

🗳️ **JK1977** 1 year, 7 months ago

Selected Answer: C

Amazon SageMaker requires that a custom algorithm container has an executable named `train` that runs your training program. This executable can be configured as an `ENTRYPOINT` in the Dockerfile, which specifies the default command to run when the container is launched.

upvoted 3 times

🗳️ **JK1977** 1 year, 7 months ago

Selected Answer: B

Amazon SageMaker requires that a custom algorithm container has an executable named `train` that runs your training program. This executable can be configured as an `ENTRYPOINT` in the Dockerfile, which specifies the default command to run when the container is launched.

upvoted 1 times

🗨️ **Dota_addict** 1 year, 7 months ago

Selected Answer: D

you are all wrong, it is D based on, <https://docs.aws.amazon.com/sagemaker/latest/dg/adapt-training-container.html>

upvoted 1 times

🗨️ **oso0348** 1 year, 8 months ago

Selected Answer: C

To package a Docker container for use with Amazon SageMaker, the training program should be configured as an ENTRYPOINT named train in the Dockerfile. This means that the training program will be automatically executed when the container is launched by Amazon SageMaker, and it can be passed command-line arguments to specify hyperparameters or other training settings.

upvoted 1 times

🗨️ **Ajose0** 1 year, 10 months ago

Selected Answer: C

The recommended option to package the Docker container for Amazon SageMaker is to configure the training program as an ENTRYPOINT named train.

This is because ENTRYPOINT allows you to specify a command that will always be executed when the Docker container is run, ensuring that the training program will always run when the container is launched by Amazon SageMaker.

Additionally, naming the ENTRYPOINT "train" is a convention used by Amazon SageMaker to identify the main training script.

upvoted 1 times

🗨️ **tsangckl** 2 years, 1 month ago

Selected Answer: C

It's C

upvoted 1 times

🗨️ **gnolam** 2 years, 3 months ago

C for sure

as per AWS docs:

> In your Dockerfile, use the exec form of the ENTRYPOINT instruction:

> ENTRYPOINT ["python", "k-means-algorithm.py"]

upvoted 1 times

🗨️ **Juka3lj** 3 years, 2 months ago

C is correct

upvoted 1 times

🗨️ **Aashi22** 3 years, 3 months ago

option C https://github.com/awsdocs/amazon-sagemaker-developer-guide/blob/master/doc_source/your-algorithms-training-algo-dockerfile.md

upvoted 1 times

A Data Scientist needs to analyze employment data. The dataset contains approximately 10 million observations on people across 10 different features. During the preliminary analysis, the Data Scientist notices that income and age distributions are not normal. While income levels shows a right skew as expected, with fewer individuals having a higher income, the age distribution also shows a right skew, with fewer older individuals participating in the workforce.

Which feature transformations can the Data Scientist apply to fix the incorrectly skewed data? (Choose two.)

- A. Cross-validation
- B. Numerical value binning
- C. High-degree polynomial transformation
- D. Logarithmic transformation
- E. One hot encoding

Suggested Answer: AB

Community vote distribution

BD (100%)

  **seanLu** Highly Voted 3 years, 2 months ago

I would go with B,D. Refer to quantile binning and log transform below.

<https://towardsdatascience.com/understanding-feature-engineering-part-1-continuous-numeric-data-da4e47099a7b>

upvoted 31 times

  **OmarSaadEldien** 3 years, 1 month ago

Agree with B &D

B binning for age



D for make income in normal dist

upvoted 7 times

  **omar_bahrain** 3 years, 2 months ago

agree B&D. both are strategies to eliminate the effect of skewing

upvoted 6 times

  **Joe_Zhang** Highly Voted 3 years, 3 months ago

SHOULD BE C,D



upvoted 10 times

  **Togy** Most Recent 2 months, 3 weeks ago

Selected Answer: B

Binning involves grouping numerical values into discrete intervals or bins. While it can simplify the representation of a feature and potentially make the distribution appear less skewed in a histogram, it doesn't fundamentally change the underlying skewness of the continuous data. It discretizes the data rather than transforming its distribution.

upvoted 1 times

  **VR10** 10 months, 1 week ago

Selected Answer: BD

D. Logarithmic Transformation: Addresses the right-skewed income and age distributions. The log function compresses large values, reducing the impact of outliers and making the distributions closer to normal.

B. Numerical Value Binning: Useful for the age distribution. By grouping ages into bins (e.g., 20-29, 30-39, etc.), you reduce the impact of the right skew caused by fewer older individuals. While it doesn't achieve a perfectly normal distribution, it often makes the feature more interpretable and manageable for modeling.

upvoted 1 times

  **AmeeraM** 1 year, 2 months ago

Selected Answer: BD

B and D

upvoted 1 times

🗨️ 👤 **Mickey321** 1 year, 4 months ago

Selected Answer: BD

Agree with B & D B binning for age D for make income in normal dist
upvoted 1 times

🗨️ 👤 **Shailendraa** 2 years, 3 months ago

BD is correct
upvoted 2 times

🗨️ 👤 **[Removed]** 2 years, 6 months ago

A and E, it asks incorrectly
upvoted 1 times

🗨️ 👤 **Sivadharan** 2 years, 7 months ago

Selected Answer: BD

B & D. Reasonable explanation in below discussion.
upvoted 3 times

🗨️ 👤 **angnam** 2 years, 11 months ago

BD
With age, always do quantile binning
With skewed data, always use log.
upvoted 1 times

🗨️ 👤 **Juka3lj** 3 years, 2 months ago

B because we have skewed data with few exeptions
D log transform can change distribution of data
not C - because there is no indicaiton in the text, that data is following any of the HIGH DEGREE polynomial distribution like x^4 10
upvoted 5 times

🗨️ 👤 **Vita_Rasta84444** 3 years, 2 months ago

should be c and d
upvoted 4 times

🗨️ 👤 **achiko** 3 years, 2 months ago

polynomial transformations can also be used for skewed data. <https://machinelearningmastery.com/polynomial-features-transforms-for-machine-learning/>
upvoted 3 times

🗨️ 👤 **jiadong** 3 years, 2 months ago

It seems the ans are C,D
<https://anshikaaxena.medium.com/how-skewed-data-can-skrew-your-linear-regression-model-accuracy-and-transfromation-can-help-62c6d3fe4c53>
upvoted 5 times

A web-based company wants to improve its conversion rate on its landing page. Using a large historical dataset of customer visits, the company has repeatedly trained a multi-class deep learning network algorithm on Amazon SageMaker. However, there is an overfitting problem: training data shows 90% accuracy in predictions, while test data shows 70% accuracy only.

The company needs to boost the generalization of its model before deploying it into production to maximize conversions of visits to purchases. Which action is recommended to provide the HIGHEST accuracy model for the company's test and validation data?

- A. Increase the randomization of training data in the mini-batches used in training
- B. Allocate a higher proportion of the overall data to the training dataset
- C. Apply L1 or L2 regularization and dropouts to the training
- D. Reduce the number of layers and units (or neurons) from the deep learning network

Suggested Answer: D



Community vote distribution

C (81%)

D (19%)

  **knightknt** Highly Voted 2 years, 2 months ago

I think C will be answer, because we even don't know how many layers now, so apply L1,L2 and dropouts layer will be first resort to solve overfitting. If it still does not work, then to reduce layers
upvoted 11 times

  **mamun4105** Most Recent 9 months, 3 weeks ago

D: D is the correct answer. C could be the answer only if it is a regression problem. You cannot apply L1 (Lasso regression) and L2 (Ridge regression) to classification problems. However, you can use dropout here.
upvoted 1 times

  **DimLam** 8 months ago

Why do you think it works only for regression problems? L1/L2 regularizations are just adding penalties to loss functions. I don't see any problems with applying it to DL model
upvoted 1 times

  **Mickey321** 10 months ago

Selected Answer: C

C Regularization
upvoted 2 times

  **kaike_reis** 11 months ago

Selected Answer: C

if you see overfit think regularization.
upvoted 1 times

  **Khalil11** 1 year, 2 months ago


Selected Answer: C

C is the correct answer: The overfitting problem can be addressed by applying regularization techniques such as L1 or L2 regularization and dropouts. Regularization techniques add a penalty term to the cost function of the model, which helps to reduce the complexity of the model and prevent it from overfitting to the training data. Dropouts randomly turn off some of the neurons during training, which also helps to prevent overfitting.
upvoted 2 times

  **Valcilio** 1 year, 3 months ago

Selected Answer: C

D can work, but C is a better answer!
upvoted 2 times

  **drcok87** 1 year, 4 months ago

C and D both seems to be correct but, seems like removing layer is first step in to optimization
<https://www.kaggle.com/general/175912>
d
upvoted 2 times

🗨️ 👤 **Ajose0** 1 year, 4 months ago

Selected Answer: C

C. Apply L1 or L2 regularization and dropouts to the training" because regularization can help reduce overfitting by adding a penalty to the loss function for large weights, preventing the model from memorizing the training data.

Dropout is a regularization technique that randomly drops out neurons during the training process, further reducing the risk of overfitting.
upvoted 1 times

🗨️ 👤 **albu44** 1 year, 5 months ago

Selected Answer: D

"The first step when dealing with overfitting is to decrease the complexity of the model. To decrease the complexity, we can simply remove layers or reduce the number of neurons to make the network smaller."

<https://www.kdnuggets.com/2019/12/5-techniques-prevent-overfitting-neural-networks.html>

upvoted 1 times

🗨️ 👤 **Peeking** 1 year, 6 months ago

Selected Answer: D

Deep learning tuning order:

1. Number of layers
2. Number of neurons (indirectly implements dropout)
3. L1/L2 regularization
4. Dropout

upvoted 4 times

🗨️ 👤 **kaike_reis** 11 months ago

the problem is overfitting, not HP Tuning.

upvoted 1 times

🗨️ 👤 **Shakespeare** 6 months, 2 weeks ago

Can be used for overfitting as well, but the problem does not say it is a deep learning algorithm being used so C would be more appropriate.

upvoted 1 times

🗨️ 👤 **Parth12** 1 year, 11 months ago

Selected Answer: C

Here we are looking to reduce the Overfitting to improve the generalization. In order to do so, L1(or Lasso) regression has always been a good aide.

upvoted 3 times

🗨️ 👤 **mamun4105** 9 months, 3 weeks ago

This is not a regression problem at all.

upvoted 1 times

🗨️ 👤 **mtp1993** 2 years ago

Selected Answer: C

C, Regularization and dropouts should be the first attempt

upvoted 3 times

🗨️ 👤 **ovokpus** 2 years ago

Selected Answer: C

Yes, C is right here. Regularization and Dropouts

upvoted 3 times

🗨️ 👤 **Abdelrahman_Omran** 2 years, 2 months ago

Selected Answer: C

C is the answer

upvoted 4 times

A Machine Learning Specialist is given a structured dataset on the shopping habits of a company's customer base. The dataset contains thousands of columns of data and hundreds of numerical columns for each customer. The Specialist wants to identify whether there are natural groupings for these columns across all customers and visualize the results as quickly as possible.

What approach should the Specialist take to accomplish these tasks?

- A. Embed the numerical features using the t-distributed stochastic neighbor embedding (t-SNE) algorithm and create a scatter plot.
- B. Run k-means using the Euclidean distance measure for different values of k and create an elbow plot.
- C. Embed the numerical features using the t-distributed stochastic neighbor embedding (t-SNE) algorithm and create a line graph.
- D. Run k-means using the Euclidean distance measure for different values of k and create box plots for each numerical column within each cluster.

Suggested Answer: B

Community vote distribution



A (92%)

8%

  **ac71** Highly Voted 2 years, 9 months ago

A is correct. tSNE can do segmentation or grouping as well. Refer: <https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1>

upvoted 21 times

  **SophieSu** Highly Voted 2 years, 8 months ago

A is definitely the correct answer.

Pay attention to what the question is asking:

"whether there are natural groupings for these columns across all customers and visualize the results as quickly as possible"

The key point is to visualize the "groupings"(exactly what t-SNE scatter plot does, it visualize high-dimensional data points on 2D space).

The question does not ask to visualize how many groups you would classify (K-Means Elbow Plot does not visualize the groupings, it is used to determine the optimal # of groups=K).

upvoted 18 times

  **Mickey321** Most Recent 10 months ago

Selected Answer: A

option A

upvoted 1 times

  **kaike_reis** 11 months ago

B doesn't even answer the question: how are you going to see your customer groups in an elbow plot

upvoted 1 times

  **windy9** 9 months ago

Elbow plot helps you identify the correct number of clusters during K-Means clustering. The clustering happens basis of all the features and thus group employees. This is to help your understanding. And the correct answer however is still tSNE because the question focuses on identifying relationships/similarities between the features / columns in the dataset. The correct answer is A



upvoted 1 times

  **kaike_reis** 11 months ago

Selected Answer: A

Euclidean Distance suffers for high dimensional data. tSNE can suffers as well, but from my perspective is the correct one.

upvoted 1 times



  **Sylzys** 1 year, 3 months ago

Selected Answer: A

Elbow plot will not help visualize groups, only try to predict an optimal number of clusters.

I think A is a better choice here

upvoted 2 times

  **Ajose0** 1 year, 4 months ago

Selected Answer: A

A.

The t-SNE algorithm is a popular tool for visualizing high-dimensional datasets, as it can transform high-dimensional data into a 2D scatter plot, which makes it easier to visualize and understand the relationships between data points.

The scatter plot produced by t-SNE can be interpreted as a map that reveals the structure of the data, showing whether there are natural groupings or clusters within the data.

Option A is the quickest and simplest way to visualize the data in a meaningful way, allowing the Specialist to gain insights into the data more efficiently.

upvoted 3 times

🗳️ 👤 **minkhant19** 1 year, 7 months ago

A is correct

upvoted 1 times

🗳️ 👤 **Shailendraa** 1 year, 9 months ago

12-sep exam

upvoted 3 times

🗳️ 👤 **Morsa** 1 year, 11 months ago

Selected Answer: A

A as k-means elbow is erroneous. It does not helping here. Scatter plot and t-sne is the right answer

upvoted 2 times

🗳️ 👤 **ovokpus** 2 years ago

Selected Answer: A

An elbow plot (B) will not give you what the question is asking for. A scatter plot will, and t-SNE is first for visualizing before dimensionality reduction.

upvoted 2 times

🗳️ 👤 **Sadgamaya** 2 years, 3 months ago

A is correct as k means suffer from curse of dimensionality and t-sne will be a better option.

upvoted 1 times

🗳️ 👤 **Mircuz** 2 years, 3 months ago

Selected Answer: A

The B,C,D plots are meaningless wrt the problem → A

upvoted 2 times

🗳️ 👤 **Mircuz** 2 years, 3 months ago

Selected Answer: B

t-SNE suffers curse of dimensionality and is indicated for small datasets

upvoted 1 times

🗳️ 👤 **AddiWei** 2 years, 4 months ago

Additionally the numeric features don't require "embedding". I think they meant to write "standardize"

upvoted 1 times

🗳️ 👤 **apprehensive_scar** 2 years, 4 months ago

Rooting for A

upvoted 1 times

🗳️ 👤 **bitsplease** 2 years, 5 months ago

B & D are wrong—because data contains "thousands of columns" and using k-means with euclidean suffers from "curse of dimensionality"

Thus leaving A & C, you CANNOT viz clusters/groups/segments in a line graph so correct answer is A

upvoted 1 times

A Machine Learning Specialist is planning to create a long-running Amazon EMR cluster. The EMR cluster will have 1 master node, 10 core nodes, and 20 task nodes. To save on costs, the Specialist will use Spot Instances in the EMR cluster. Which nodes should the Specialist launch on Spot Instances?

- A. Master node
- B. Any of the core nodes
- C. Any of the task nodes
- D. Both core and task nodes

Suggested Answer: A

Community vote distribution

C (100%)

🗳️ 👤 **[Removed]** Highly Voted 👍 2 years, 8 months ago

Answer is C. <https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html>
upvoted 23 times

🗳️ 👤 **Sneep** 1 year, 5 months ago

It's definitely C. The fact that this site indicates A is a clear sign that answers are just randomly selected, it would make zero sense to spot-instance the master node for an EMR cluster. Make sure you look at discussions for all of these questions.
upvoted 4 times

🗳️ 👤 **SophieSu** Highly Voted 👍 2 years, 8 months ago

C is the correct answer.

"Long-Running Clusters and Data Warehouses

If you are running a persistent Amazon EMR cluster that has a predictable variation in computational capacity, such as a data warehouse, you can handle peak demand at lower cost with Spot Instances. You can launch your master and core instance groups as On-Demand Instances to handle the normal capacity and launch task instance groups as Spot Instances to handle your peak load requirements."
upvoted 10 times

🗳️ 👤 **teka112233** Most Recent 🕒 10 months, 1 week ago

Selected Answer: C

According to :<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html>

The task nodes process data but do not hold persistent data in HDFS. If they terminate because the Spot price has risen above your maximum Spot price, no data is lost and the effect on your cluster is minimal.

When you launch one or more task instance groups as Spot Instances, Amazon EMR provisions as many task nodes as it can, using your maximum Spot price. This means that if you request a task instance group with six nodes, and only five Spot Instances are available at or below your maximum Spot price, Amazon EMR launches the instance group with five nodes, adding the sixth later if possible.
upvoted 1 times

🗳️ 👤 **Khalil11** 1 year, 2 months ago

Selected Answer: C

The correct answer is C

upvoted 1 times

🗳️ 👤 **Sylzys** 1 year, 3 months ago

Selected Answer: C

I don't get why the wrong answer are still not updated after more than 1 year of everyone showing docs proving answer C..

upvoted 3 times

🗳️ 👤 **gusta_dantas** 11 months, 1 week ago

1 and a half year and still wrong.. Incredible!

upvoted 2 times

🗳️ 👤 **AjoseO** 1 year, 4 months ago

Selected Answer: C

Long-running clusters and data warehouses

If you are running a persistent Amazon EMR cluster that has a predictable variation in computational capacity, such as a data warehouse, you can handle peak demand at lower cost with Spot Instances.

You can launch your primary and core instance groups as On-Demand Instances to handle the normal capacity and launch the task instance group as Spot Instances to handle your peak load requirements.

upvoted 1 times

🗳️ 👤 **SK27** 1 year, 6 months ago

Selected Answer: C

Only task nodes can be deleted without losing data.

upvoted 1 times

🗳️ 👤 **Twist3d** 1 year, 6 months ago

C, If you want to cut cost on an EMR cluster in the most efficient way, use spot instances on the task nodes because it, task nodes do not store data so no risk of data loss

upvoted 1 times

🗳️ 👤 **ovokpus** 2 years ago

Selected Answer: C

For Long running jobs, you do not want to compromise the Master node(sudden termination) or the core nodes (HDFS data loss).

Spot Instances on 20 task nodes are enough cost savings without compromising the job.

Hence, C

upvoted 3 times

🗳️ 👤 **Jump09** 2 years ago

If your primary concern is the cost, then you can run the master node on spot instances.

upvoted 1 times

🗳️ 👤 **Jump09** 2 years ago

Adding the related reference from the AWS documentation:

Master node on a Spot Instance

The master node controls and directs the cluster. When it terminates, the cluster ends, so you should only launch the master node as a Spot Instance if you are running a cluster where sudden termination is acceptable. This might be the case if you are testing a new application, have a cluster that periodically persists data to an external store such as Amazon S3, or are running a cluster where cost is more important than ensuring the cluster's completion.

upvoted 1 times

🗳️ 👤 **Jump09** 2 years ago

In the question , there are no specific conditions mentioned except the concern with the COST, thus I think the answer should be A.

upvoted 1 times

🗳️ 👤 **benson2021** 2 years, 8 months ago

Answer: C. <https://aws.amazon.com/getting-started/hands-on/optimize-amazon-emr-clusters-with-ec2-spot/>

Amazon recommends using On-Demand instances for Master and Core nodes unless you are launching highly ephemeral workloads.

upvoted 5 times

🗳️ 👤 **xpada001** 2 years, 8 months ago

Answer should be C.

upvoted 3 times

🗳️ 👤 **ac71** 2 years, 8 months ago

Only master node is incorrect. Either use all on spot or only task or core on spot. As per:

<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html>

Better to use only task node on spot for long running tasks/jobs

upvoted 3 times

🗳️ 👤 **MahEid** 2 years, 7 months ago

Answer is C

you should only run core nodes on Spot Instances /*when partial HDFS data loss is tolerable*/

Question is what "Should" be launched as spot instance
upvoted 1 times

A manufacturer of car engines collects data from cars as they are being driven. The data collected includes timestamp, engine temperature, rotations per minute (RPM), and other sensor readings. The company wants to predict when an engine is going to have a problem, so it can notify drivers in advance to get engine maintenance. The engine data is loaded into a data lake for training. Which is the MOST suitable predictive model that can be deployed into production?

- A. Add labels over time to indicate which engine faults occur at what time in the future to turn this into a supervised learning problem. Use a recurrent neural network (RNN) to train the model to recognize when an engine might need maintenance for a certain fault.
- B. This data requires an unsupervised learning algorithm. Use Amazon SageMaker k-means to cluster the data.
- C. Add labels over time to indicate which engine faults occur at what time in the future to turn this into a supervised learning problem. Use a convolutional neural network (CNN) to train the model to recognize when an engine might need maintenance for a certain fault.
- D. This data is already formulated as a time series. Use Amazon SageMaker seq2seq to model the time series.

Suggested Answer: B

Community vote distribution

A (100%)

  **ac71** Highly Voted 2 years, 9 months ago

This is a supervised problem and needs labels. Can't use clustering to find when faults can happen. CNN is for images not for timeseries data here. Hence, A seems appropriate.
upvoted 53 times

  **youjun** 2 years, 6 months ago

AGREE WITH YOU
upvoted 2 times

  **astonm13** 2 years, 8 months ago

Agree, the answer is A
upvoted 7 times

  **loict** Most Recent 9 months, 2 weeks ago



Selected Answer: A

- A. YES - RNN good for time series as we want to use previous input
 - B. NO - we know the class (fault) ahead of time, it is supervised
 - C. NO - CNN is for images
 - D. NO - seq2seq is for word generation
- upvoted 1 times

  **Mickey321** 10 months ago

Selected Answer: A



Answer is A
upvoted 1 times

  **Ajose0** 1 year, 4 months ago

Selected Answer: A

A recurrent neural network (RNN) is a more suitable choice than a convolutional neural network (CNN) because the data collected from the engines is a sequence of values over time, and the goal is to predict a future event (an engine fault). RNNs are designed to handle sequential data and can learn patterns and dependencies over time, making them well-suited for time-series data like this.

On the other hand, CNNs are designed for image processing and are not ideal for sequential data.
upvoted 3 times

  **spidy20** 1 year, 10 months ago

Selected Answer: A

Answer should be A
upvoted 1 times

🗨️ 👤 **Morsa** 1 year, 11 months ago

Selected Answer: A

It can only be A. Agree with the comments before
upvoted 1 times

🗨️ 👤 **irimala** 2 years, 1 month ago

Selected Answer: A

Obviously A
upvoted 2 times

🗨️ 👤 **apprehensive_scar** 2 years, 4 months ago

Selected Answer: A

A - obviously.
upvoted 1 times

🗨️ 👤 **bitsplease** 2 years, 5 months ago

Seq2Seq also uses RNN under the hood, BUT option D. did not mention anything about "adding labels"--which is required here--hence --> A
upvoted 3 times

🗨️ 👤 **geekgirl007** 2 years, 5 months ago

Selected Answer: A

A is correct. CNN is for images and RNN is for timeseries.
upvoted 1 times

🗨️ 👤 **loyor94478** 2 years, 7 months ago

AAAAAAAAAAAAA
<https://towardsdatascience.com/how-to-implement-machine-learning-for-predictive-maintenance-4633cdbe4860>
upvoted 2 times

🗨️ 👤 **omar8024** 2 years, 8 months ago

I think A is correct
upvoted 1 times

🗨️ 👤 **Vita_Rasta84444** 2 years, 8 months ago

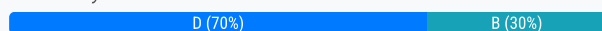
It is A
upvoted 1 times

A company wants to predict the sale prices of houses based on available historical sales data. The target variable in the company's dataset is the sale price. The features include parameters such as the lot size, living area measurements, non-living area measurements, number of bedrooms, number of bathrooms, year built, and postal code. The company wants to use multi-variable linear regression to predict house sale prices. Which step should a machine learning specialist take to remove features that are irrelevant for the analysis and reduce the model's complexity?

- A. Plot a histogram of the features and compute their standard deviation. Remove features with high variance.
- B. Plot a histogram of the features and compute their standard deviation. Remove features with low variance.
- C. Build a heatmap showing the correlation of the dataset against itself. Remove features with low mutual correlation scores.
- D. Run a correlation check of all features against the target variable. Remove features with low target variable correlation scores.

Suggested Answer: D

Community vote distribution



puffpuff Highly Voted 2 years, 8 months ago

D should be the more comprehensive answer. If it's not correlated, you can't make use of it in a linear regression
A lot of others say B, but low variance can also be due to the nature/typical magnitudes of the variable itself
upvoted 29 times

hamimelon 1 year, 6 months ago

I think the problem with B is that what is considered "low variance"? The features are on different scales.
upvoted 1 times

V_B_ 1 year, 10 months ago

Correlation indicates only linear relation, but, there might be non linear as well. To exploit it in the Linear Regression, you can take the variables to some power or run some non linear preprocessing on it, and you don't have to change the algorithm for it.
So, answer B seem much more solid for me.
upvoted 2 times

ahquiceno Highly Voted 2 years, 9 months ago

Answer B. Is not the best solution prior can use other analysis. <https://community.dataquest.io/t/feature-selection-features-with-low-variance/2418>
If the variance is low or close to zero, then a feature is approximately constant and will not improve the performance of the model. In that case, it should be removed. Or if only a handful of observations differ from a constant value, the variance will also be very low.
upvoted 16 times

fshkento 2 years, 7 months ago

Low variance does not mean the feature is not important, right?
If variance of target true value is also small and the correlation between above feature and target, the feature can be important feature.
upvoted 7 times

rb39 1 year, 9 months ago

it does. If feature and target are correlated and you expect the target to change, the feature must have some sort of variance. Otherwise it means feature is almost constant so does target.
upvoted 1 times

akgarg00 Most Recent 7 months, 1 week ago

Selected Answer: D

D is the best answer as it is mentioned multivariable linear regression applied where correlation is strong between dependent and independent variables.
upvoted 1 times

mirik 1 year ago

Selected Answer: D

D: We should remove features that are strongly correlated with each other and weakly correlated with the target:

<https://androidkt.com/find-correlation-between-features-and-target-using-the-correlation-matrix/>

You can evaluate the relationship between each feature and target using a correlation and selecting those features that have the strongest relationship with the target variable.

upvoted 2 times

🗨️ **HunterZ9527** 1 year, 2 months ago

Selected Answer: D

I think D is the correct answer. If I remember correctly, Benjamini-Hochberg Method is essentially answer D if you consider the Hypothesis to be: the feature is powerfully influential to the target.

My problem with B is that the variance can be easily affected by the scale. In the question, the number of bedroom's variance is very low, while the sqrt of the house has a high variance, both of these could be very useful. Furthermore, zip codes are included, and it is safe to assume the variance of zip codes can be high, but the information is very limited, especially if you use them as numerical instead of categorical features.

upvoted 1 times

🗨️ **Valcilio** 1 year, 3 months ago

Selected Answer: D

B is correct but the answer in D is better.

upvoted 2 times

🗨️ **Ajose0** 1 year, 4 months ago

Selected Answer: D

D is preferred over C because the goal is to predict the sale price of houses, which is the target variable. By checking the correlation of each feature against the target variable, the machine learning specialist can identify which features are most relevant to the prediction of the sale price and which are less relevant. Removing features with low correlation to the target variable helps reduce the complexity of the model and potentially improve its accuracy.

On the other hand, a heatmap showing the correlation of the dataset against itself (C) doesn't directly address the relevance of the features to the target variable, and so it's not as effective in reducing the complexity of the model.

upvoted 3 times

🗨️ **expertguru** 1 year, 5 months ago

Answer should be D, THIS is feature elimination /selection during feature Engineering. Choice c is so close just to confuse test takers to pick the wrong choice! See below C and D answers -- C should have been correct if the question asked about how to visualize correlation among independent variables! PROVIDED second sentence in C needs to be removed or to say which feature you will eliminate in such case then the one with low correlation against target out of those two.

C. Build a heatmap showing the correlation of the dataset against itself. Remove features with low mutual correlation scores.

D. Run a correlation check of all features against the target variable. Remove features with low target variable correlation scores.

upvoted 1 times

🗨️ **Ob1KN0B** 1 year, 10 months ago

Selected Answer: D

The multiple regression model is based on the following assumptions:

There is a linear relationship between the dependent variables and the independent variables

The independent variables are not too highly correlated with each other

yi observations are selected independently and randomly from the population

Residuals should be normally distributed with a mean of 0 and variance σ

upvoted 5 times

🗨️ **wakuwaku** 2 years, 4 months ago

I think the answer is D.

If the model is a decision tree or something like that, I don't think it is possible to make a decision based only on the direct correlation with the target variable.

But in multiple linear regression, the only thing that matters is the relationship between the target variable and the feature variable.

B, if the standard deviation is small but not zero, then we have information.

upvoted 3 times

🗨️ **apprehensive_scar** 2 years, 4 months ago

Selected Answer: B

B is correct.

upvoted 2 times

🗨️ **Peasfull** 2 years, 5 months ago

To eliminate extraneous information. So, the answer is D.

upvoted 2 times

🗨️ 👤 **Asrivastava3** 2 years, 6 months ago

Correct answer is D. The reason B is wrong because it is difficult to reason out why would you plot a histogram? Absolutely unnecessary step and distraction choice.

upvoted 4 times

🗨️ 👤 **[Removed]** 2 years, 6 months ago

Selected Answer: B

D is not the proper answer. Here is why:

It says that it is comparing with the target variable (dependent variable), which implies it is comparing the correlation between the dependent and independent variables. This type of comparison is usually done after a model is constructed in order to prevent assessing the predictive strength of the model. To compare the target label, the label you wish to predict, with the other variables before - is premature and will likely result in weakening your model.

Variables with low variance has very less information and the inclusion of which will likely weaken the model performance.

Hence, B.

upvoted 4 times

🗨️ 👤 **Mikky0** 2 years, 7 months ago

Answer is D.

<https://deep-r.medium.com/difference-between-variance-co-variance-and-correlation-ea0b7ddbba1>

upvoted 5 times

🗨️ 👤 **Huy** 2 years, 8 months ago

Answer C. Heatmaps is used to visualize for correlation matrix <https://towardsdatascience.com/better-heatmaps-and-correlation-matrix-plots-in-python-41445d0f2bec>

upvoted 1 times

🗨️ 👤 **mahmoudai** 2 years, 7 months ago

but is mentioned, "Remove features with low mutual correlation scores." which is wrong you should drop features with high correlation scores. so

Answer is D

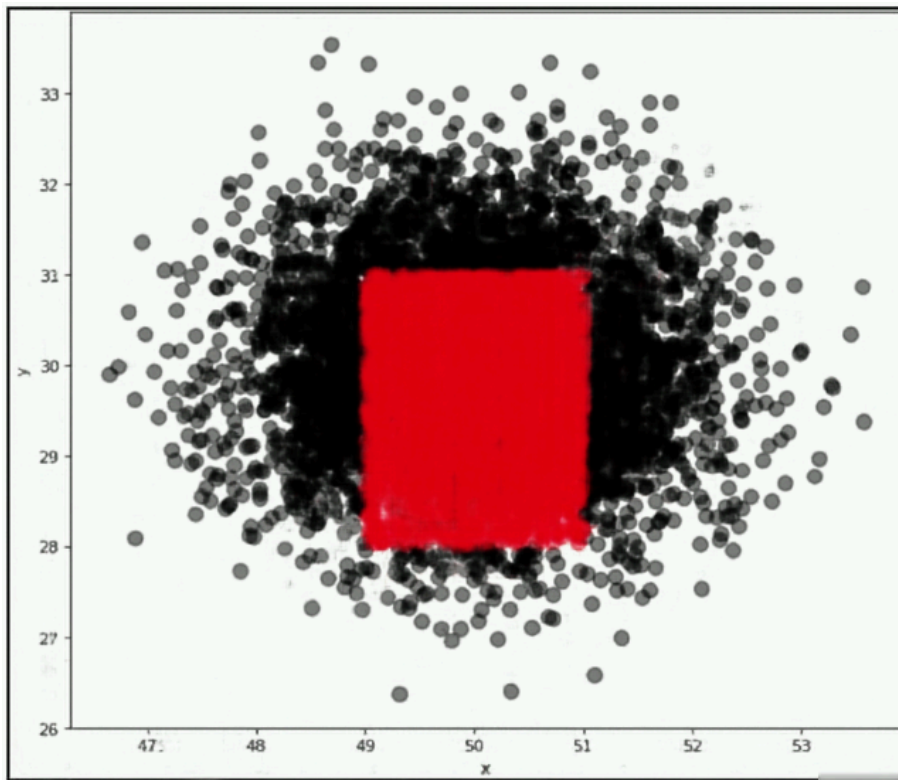
upvoted 5 times

🗨️ 👤 **hero67** 2 years, 8 months ago

The problem with correlation tasks is it capture linear relations only. So, I would go with B

upvoted 1 times

A company wants to classify user behavior as either fraudulent or normal. Based on internal research, a machine learning specialist will build a binary classifier based on two features: age of account, denoted by x , and transaction month, denoted by y . The class distributions are illustrated in the provided figure. The positive class is portrayed in red, while the negative class is portrayed in black.



Which model would have the HIGHEST accuracy?

- A. Linear support vector machine (SVM)
- B. Decision tree
- C. Support vector machine (SVM) with a radial basis function kernel
- D. Single perceptron with a Tanh activation function

Suggested Answer: C

Community vote distribution

B (55%) C (45%)

[Removed] 3 years, 9 months ago

Due to straight angles, I would choose Decision tree. See https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html#sphx-glr-auto-examples-classification-plot-classifier-comparison-py upvoted 24 times

MrCarter 3 years, 8 months ago

From your link it is obvious that the best answer is still SVM with RBF kernel. In your link the SVM-RBF got 88% accuracy on the 'square-like' dataset whereas the Decision tree achieved only 80%. Answer is SVM with RBF kernel upvoted 16 times

ttsun 3 years, 7 months ago

note the data from sklearn link is shaped as a ball of mass not a square. the RBF kernel would be better but the question shows a square. Decision tree should be better fit for this problem. upvoted 7 times

SophieSu 3 years, 9 months ago

B - Decision tree - is not the best answer. If you use decision tree to do clustering, every time you need to partition the space into 2 parts. Hence you will split the space into $2^3=8$. The red points in the center box and the black points will fall into the 7 boxes around it. The black points will be identified as 7 different classes.

C is the correct answer. SVM with non-linear kernel is appropriate for non-linear clustering. Even if the shape is close to rectangular. SVM with non-linear kernel will be able to approximate the rectangular boundary shape.

upvoted 18 times

🗨️ **robotgeek** 1 year, 9 months ago

Your statement "The black points will be identified as 8 different classes" does not make a lot of sense because the leaf node in a tree will be 1 of 2 classes, not 8 different classes just because they are visually in one place or the other

upvoted 1 times

🗨️ **Madwyn** 3 years, 8 months ago

The tree works like this with this branch with 4 nodes:

Age > 49? Y

Age > 51? N

Transaction > 28? Y

Transaction > 31? N

Positive

Correct answer is B.

upvoted 10 times

🗨️ **nick3332** Most Recent 2 months, 3 weeks ago

Selected Answer: B

Tip: When details are missing, assume ideal conditions, so assume no overfitting issues. Therefore B is better than C. If there are overfitting issues or a possibility of overfitting then C is the right answer.

upvoted 1 times

🗨️ **2eb8df0** 4 months ago

Selected Answer: B

Decision tree makes more sense, this decision boundary isn't complex at all and there is no risk of overfitting, all the points are inside the square

upvoted 2 times

🗨️ **MultiCloudIronMan** 8 months, 1 week ago

Selected Answer: C

This is because the RBF kernel can handle non-linear relationships between features, which is often necessary for complex classification tasks.

upvoted 2 times

🗨️ **MJSY** 9 months ago

Selected Answer: C

Decision Tree can treat the training data well but will have a risk of overfitting. the SVM with RBF kernel will be more robust.

upvoted 2 times

🗨️ **rookiee1111** 1 year, 2 months ago

Selected Answer: B

As the positive cases can be interpreted and separated from non positive ones by decision tree easily. SVM would have made sense if the two classes were inseparable or had complex relationship in data.

upvoted 1 times

🗨️ **vkajoria** 1 year, 2 months ago

Selected Answer: C

It is C SVM with RBF Kernel can classify this image. For decision tree, it will be more difficult

upvoted 2 times

🗨️ **kyuhuck** 1 year, 4 months ago

Selected Answer: C

From the visual information provided, an SVM with an RBF kernel (Option C) would likely be the best choice because it can handle the circular class distribution. The RBF kernel is especially good at dealing with such scenarios where the boundary between classes is not linear.

upvoted 2 times

🗨️ **Alice1234** 1 year, 4 months ago

Answer C

B. Decision Tree: Decision trees can capture non-linear patterns and are capable of splitting the feature space in complex ways. They can be very effective if the decision boundary is not linear, but they might also overfit if the decision boundary is too complex.

C. SVM with RBF Kernel: An SVM with a radial basis function (RBF) kernel is designed to handle non-linear boundaries by mapping input features into higher-dimensional spaces where the classes are more likely to be separated by a hyperplane. Given the clustered nature of the classes in the image, an SVM with an RBF kernel would likely be able to separate the classes with a higher degree of accuracy.

upvoted 2 times

🗨️ 👤 **praveenaws** 1 year, 6 months ago

Selected Answer: C

SVM-RBF is the correct solution

upvoted 1 times

🗨️ 👤 **Neet1983** 1 year, 6 months ago

Support vector machine (SVM) with a radial basis function kernel would likely have the highest accuracy for this task because it can handle the non-linear separation required by the data.

upvoted 1 times

🗨️ 👤 **endeesa** 1 year, 7 months ago

Selected Answer: C

I will lean with C

upvoted 1 times

🗨️ 👤 **akgarg00** 1 year, 7 months ago

Answer is B as Decision tree can attain 100% accuracy in this case.

upvoted 1 times

🗨️ 👤 **loict** 1 year, 9 months ago

Selected Answer: C

SVM with RBF and proper C and Gamma value can accomodate this square shape (<https://vitalflux.com/svm-rbf-kernel-parameters-code-sample/>)

upvoted 1 times

🗨️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: C

confusing between SVN or devision tree. learning towards C

upvoted 1 times

🗨️ 👤 **rag1482** 2 years, 1 month ago

Answer C

In general, SVMs are a good choice for tasks where accuracy is critical, such as fraud detection and medical diagnosis. Decision trees are a good choice for tasks where interpretability is important, such as customer segmentation and product recommendation.

upvoted 2 times

A health care company is planning to use neural networks to classify their X-ray images into normal and abnormal classes. The labeled data is divided into a training set of 1,000 images and a test set of 200 images. The initial training of a neural network model with 50 hidden layers yielded 99% accuracy on the training set, but only 55% accuracy on the test set.

What changes should the Specialist consider to solve this issue? (Choose three.)

- A. Choose a higher number of layers
- B. Choose a lower number of layers
- C. Choose a smaller learning rate
- D. Enable dropout
- E. Include all the images from the test set in the training set
- F. Enable early stopping

Suggested Answer: ADE

Community vote distribution

BDF (89%)

11%

 **cnethers** Highly Voted 3 years, 2 months ago

when looking at an overfitting issue :


<https://www.kdnuggets.com/2019/12/5-techniques-prevent-overfitting-neural-networks.html>

1. Simplifying The Model (reduce number of layers)
2. Early Stopping
3. Use Data Augmentation
4. Use Regularization (L1 + L2)
5. Use Dropouts

So looking at the options:

B, D, F

upvoted 48 times

 **SophieSu** Highly Voted 3 years, 2 months ago


BDF !!!

upvoted 8 times

 **asthamishra** Most Recent 11 months, 2 weeks ago

looking at last 100 questions many answers were wrong , thanks to the discussion forum to provide correct answer

upvoted 2 times

 **AmeeraM** 1 year, 2 months ago

Selected Answer: BDF

I would say BCD or BDF


upvoted 1 times

 **Mickey321** 1 year, 4 months ago

Selected Answer: BDF

Agree with BDF

upvoted 1 times

 **JK1977** 1 year, 7 months ago

Selected Answer: BDF

Over fitting problem. All the options B, D, F reduce over fitting.

upvoted 1 times

 **Debayandt91** 1 year, 7 months ago

In what world is ACE the answer ?

upvoted 1 times

🗨️ 👤 **Dota_addict** 1 year, 7 months ago

Selected Answer: BDF

BDF is the answer

upvoted 1 times

🗨️ 👤 **Aninina** 1 year, 12 months ago

Selected Answer: BDF

BDF is the correct

upvoted 1 times

🗨️ 👤 **Peeking** 2 years ago

Selected Answer: BDF

ADE is absolutely wrong. 50 layers is already overfitting the model. We cannot increase the number of layers again.

upvoted 1 times

🗨️ 👤 **GauravLahotiML** 2 years, 1 month ago

Selected Answer: BDF

BDF is the correct answer

upvoted 1 times

🗨️ 👤 **exam887** 2 years, 6 months ago

Selected Answer: BDF

should be BDF

upvoted 1 times

🗨️ 👤 **NILKK** 2 years, 8 months ago

One of the correct answer is showing as A. I wanted to understand how A(Choose Higher Number of Layers) is the correct Answer ?

upvoted 1 times

🗨️ 👤 **KM226** 3 years ago

Selected Answer: BCE

I believe the answer is BCE because the model is overfitting.

upvoted 1 times

🗨️ 👤 **windy9** 1 year, 2 months ago

C might not be, because the model yielded 99% accuracy on the training set

upvoted 2 times

🗨️ 👤 **johnvik** 3 years, 2 months ago

choose smaller learning rate c, d, f,

upvoted 1 times

🗨️ 👤 **johnvik** 3 years, 2 months ago

ignore answer is correct BDF

upvoted 1 times

🗨️ 👤 **Vita_Rasta84444** 3 years, 2 months ago

BDF!!!

upvoted 3 times

🗨️ 👤 **astonm13** 3 years, 2 months ago

It is supposed to be BDF

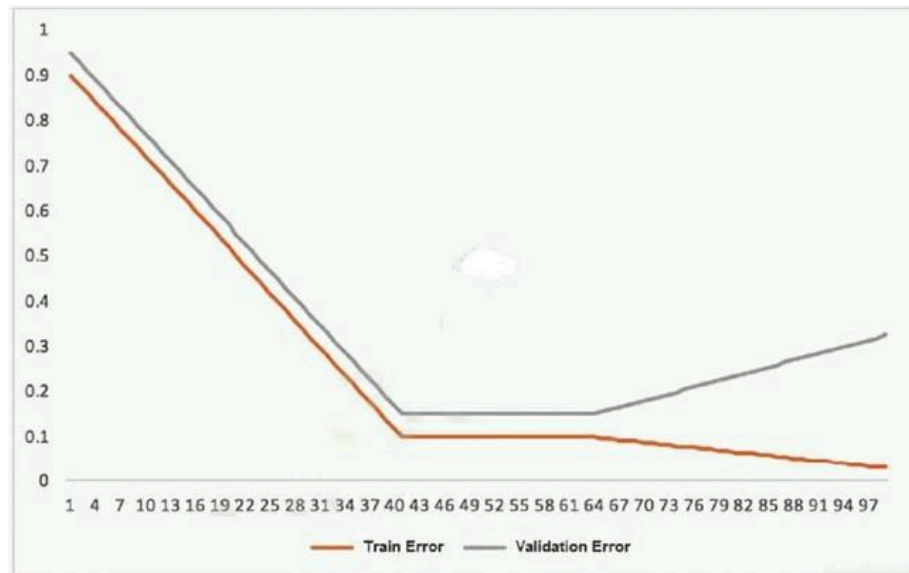
upvoted 2 times

This graph shows the training and validation loss against the epochs for a neural network.

The network being trained is as follows:

- ⇒ Two dense layers, one output neuron
- ⇒ 100 neurons in each layer
- ⇒ 100 epochs

Random initialization of weights



Which technique can be used to improve model performance in terms of accuracy in the validation set?

- A. Early stopping
- B. Random initialization of weights with appropriate seed
- C. Increasing the number of epochs
- D. Adding another layer with the 100 neurons

Suggested Answer: C

Community vote distribution

A (100%)

ahquiceno Highly Voted 3 years, 3 months ago

Answer A.

upvoted 22 times

eganilovic Highly Voted 3 years, 2 months ago

The answer is Early Stopping. Stopp the training before accuracy start do decrease.

upvoted 8 times

StelSen 3 years, 2 months ago

Appreciates your explanation. Cheers

upvoted 2 times

AIWave Most Recent 10 months, 2 weeks ago

I will go with A

Early stopping is a powerful technique to prevent overfitting. It involves monitoring the model's performance on a validation dataset during training. If the validation loss starts increasing or plateaus, early stopping stops further training. This ensures that the model doesn't overfit to the training data.

Based on the graph, if the validation loss begins to stagnate or increase after a certain number of epochs, enabling early stopping could lead to better generalization.

upvoted 1 times

🗲️ 👤 **AmeeraM** 1 year, 2 months ago

Selected Answer: A

early stopping before error increase

upvoted 1 times

🗲️ 👤 **seifski** 1 year, 2 months ago

Selected Answer: A

Early stopping

upvoted 1 times

🗲️ 👤 **Mickey321** 1 year, 4 months ago

Selected Answer: A

Early stopping

upvoted 1 times

🗲️ 👤 **vbal** 1 year, 7 months ago

A: stop the training process of a neural network before it reaches the maximum number of epochs or iterations; in this case stop close to 64 Epochs.

upvoted 1 times

🗲️ 👤 **Peeking** 2 years ago

Selected Answer: A

Early stopping and not increasing epochs.

upvoted 1 times

🗲️ 👤 **ryuhei** 2 years, 3 months ago

Selected Answer: A

Answer is "A"

upvoted 1 times

🗲️ 👤 **chrisabc** 3 years, 2 months ago

Early Stopping can improve the model?

upvoted 3 times

🗲️ 👤 **Vita_Rasta84444** 3 years, 3 months ago

A is the answer

upvoted 2 times

🗲️ 👤 **astonm13** 3 years, 3 months ago

I would go for A

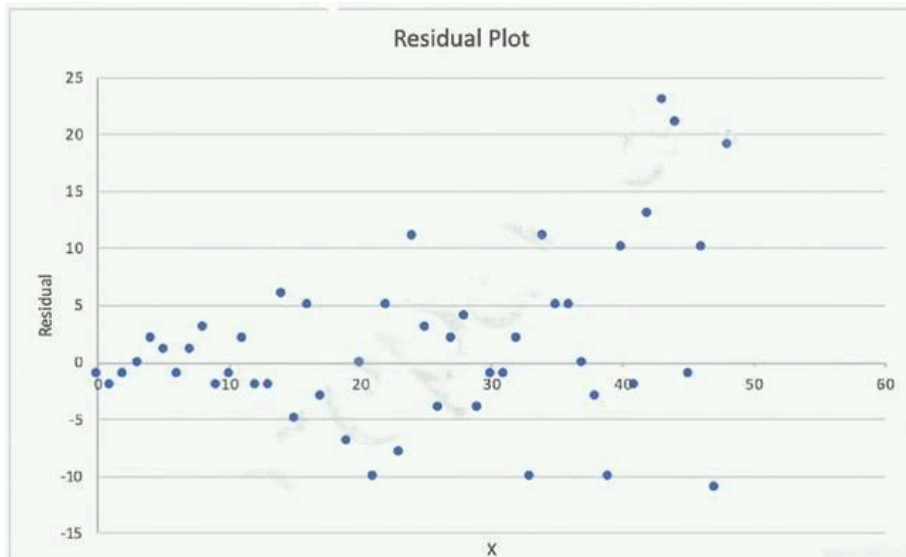
upvoted 2 times

🗲️ 👤 **[Removed]** 3 years, 3 months ago

I would choose A.

upvoted 5 times

A Machine Learning Specialist is attempting to build a linear regression model.



Given the displayed residual plot only, what is the MOST likely problem with the model?

- A. Linear regression is inappropriate. The residuals do not have constant variance.
- B. Linear regression is inappropriate. The underlying data has outliers.
- C. Linear regression is appropriate. The residuals have a zero mean.
- D. Linear regression is appropriate. The residuals have constant variance.

Suggested Answer: D

Community vote distribution

A (75%)

D (25%)

[Removed] Highly Voted 2 years, 9 months ago

I would choose A. See: <https://www.itl.nist.gov/div898/handbook/pmd/section4/pmd442.htm> and <https://blog.minitab.com/blog/the-statistics-game/checking-the-assumption-of-constant-variance-in-regression-analyses>
upvoted 25 times

takahirokoyama Highly Voted 2 years, 9 months ago

Ans. is A.
High-degree polynomial transformation.
upvoted 12 times

TrekkingMachine 2 years, 8 months ago

I think so too.
upvoted 2 times

geon13 Most Recent 7 months, 2 weeks ago

Answer A.
One of the key assumptions of linear regression is that the residuals have constant variance at every level of the predictor variable(s). If this assumption is not met, the residuals are said to suffer from heteroscedasticity. When this occurs, the estimates for the model coefficients become unreliable
<https://www.statology.org/constant-variance-assumption/>
upvoted 1 times

Mickey321 10 months ago

Selected Answer: A

Agree with A
upvoted 1 times

Nadia0012 1 year, 3 months ago

Selected Answer: A

<https://blog.minitab.com/en/the-statistics-game/checking-the-assumption-of-constant-variance-in-regression-analyses>

upvoted 1 times

🗳️ 👤 **Tomatoteacher** 1 year, 5 months ago

Selected Answer: A

Kind of like heteroskedasticity, anyways it is A.

upvoted 1 times

🗳️ 👤 **jrff** 1 year, 8 months ago

Selected Answer: A

Answer is A. It does not has constant variance !

upvoted 3 times

🗳️ 👤 **Shailendraa** 1 year, 10 months ago

A is correct answer

upvoted 1 times

🗳️ 👤 **ovokpus** 2 years ago

These images are broken. I cannot review the question properly!

upvoted 2 times

🗳️ 👤 **John_Pongthorn** 2 years, 4 months ago

Selected Answer: D

D is best answer

all x values are scattering as a whole , no matter what x is

<https://www.statisticshowto.com/residual-plot/>

if you take all x values to plot histogram , it will be bel-curve.

upvoted 2 times

🗳️ 👤 **AddiWei** 2 years, 4 months ago

And it does NOT mean linear regression is not appropriate. It means your linear regression model is biased due to several reasons.

upvoted 1 times

🗳️ 👤 **eeah** 2 years, 2 months ago

yes, it does. One of the main assumptions is homoscedasticity.

upvoted 2 times

🗳️ 👤 **AddiWei** 2 years, 4 months ago

100% A

upvoted 1 times

🗳️ 👤 **u404** 2 years, 5 months ago

I will choose A , because the data is heteroscedastic. It violates a key assumption of linear regression

upvoted 2 times

🗳️ 👤 **Huy** 2 years, 8 months ago

A. <https://www.originlab.com/doc/origin-help/residual-plot-analysis>

upvoted 2 times

🗳️ 👤 **yummytaco** 2 years, 8 months ago

Do not have constant variance

<https://stats.stackexchange.com/questions/52089/what-does-having-constant-variance-in-a-linear-regression-model-mean>

upvoted 2 times

🗳️ 👤 **Vita_Rasta84444** 2 years, 8 months ago

Answer is A. As x raises, the residuals become higher and higher...

upvoted 1 times

🗳️ 👤 **cnethers** 2 years, 8 months ago

Some Good Reading https://www.andrew.cmu.edu/user/achoulde/94842/homework/regression_diagnostics.html

Ans is A

upvoted 7 times

🗳️ 👤 **Sadgamaya** 2 years, 2 months ago

Thank you for sharing.

upvoted 1 times

A large company has developed a BI application that generates reports and dashboards using data collected from various operational metrics. The company wants to provide executives with an enhanced experience so they can use natural language to get data from the reports. The company wants the executives to be able ask questions using written and spoken interfaces. Which combination of services can be used to build this conversational interface? (Choose three.)

- A. Alexa for Business
- B. Amazon Connect
- C. Amazon Lex
- D. Amazon Polly
- E. Amazon Comprehend
- F. Amazon Transcribe

Suggested Answer: BEF

Community vote distribution



astomn13 Highly Voted 3 years, 9 months ago

C - voice and text interface
 E - understanding
 F - Speech to text
 upvoted 35 times

hero67 3 years, 8 months ago

Why would I need to transcribe while I have Lex that do the NLU part? It would be more reasonable to select Either Connect (B) or Polly (D) if the specs to generate output speech.
 upvoted 4 times

Hariru 3 years, 7 months ago

E - is more to express the "feeling" or "mood". We would rather need something, that can speak to the customer. So my suggestion is c,d,f
 upvoted 5 times

F1Fan 1 year, 2 months ago

The question states that the company wants to "provide executives with an enhanced experience so they can use natural language to get data from the reports." The key phrase here is "use natural language," which implies that the executives will be interacting with the system using human-like language, either written or spoken. To understand and interpret natural language inputs from users, whether written or spoken, the system needs to have natural language understanding (NLU) or natural language processing (NLP) capabilities. Without NLU/NLP capabilities, the system would not be able to make sense of the executives' natural language queries and extract the relevant information to retrieve data from the reports and dashboards. Services like Amazon Lex and Amazon Comprehend are specifically designed to provide NLU and NLP functionalities, respectively. Amazon Lex uses NLU models to understand the intent and extract relevant information from user inputs, while Amazon Comprehend provides NLP capabilities to analyze and extract insights from text data.
 upvoted 1 times

eganilovic Highly Voted 3 years, 8 months ago



If we need to build written and spoken interfaces we need :
 F - Transcribe (speech to text)
 D- Polly (text ot speech)
 And for chatbot:
 E - Lex
 upvoted 23 times

eganilovic 3 years, 8 months ago

*C - Lex

So C,D,F

upvoted 17 times

  **weelz** 3 years, 8 months ago

I second that, the keyword here is "conversational interface". so, no conversation without Amazon Lex
upvoted 1 times

  **sheetalconnect** Most Recent 11 months, 4 weeks ago

Selected Answer: ACD

Alexa for Business: Handles the voice interaction, converting spoken queries into text and providing the voice interface that executives use to interact with the BI application.

Amazon Lex: Processes the text input (converted by Alexa) and understands the intent behind the queries, enabling the conversational interface.

Amazon Polly: Optional but useful if you want to convert the textual responses from the BI application back into spoken responses, providing a complete voice-based interaction.


upvoted 1 times

  **ArchMelody** 1 year, 3 months ago

Selected Answer: CDF

Lex for bot service, Polly for text-to-speech (answer) and Transcribe for speech-to-text (question).

upvoted 2 times

  **vkajoria** 1 year, 4 months ago

I believe Answer should be CDF

C: Lex

D: Polly

F: Transcribe

upvoted 1 times

  **kyuhuck** 1 year, 4 months ago

Selected Answer: CDF

For a BI application where executives can ask questions using written and spoken interfaces, the following combination of services would be suitable:

Amazon Lex (Option C): To build the core conversational interface that understands and processes natural language queries.

Amazon Polly (Option D): To provide spoken responses to written queries, giving a more interactive experience for users who are not using the voice interface.

Amazon Transcribe (Option F): To convert spoken queries into text that can be understood by Amazon Lex.

These three services would work together to provide a comprehensive conversational interface that allows for both text and voice interactions, meeting the requirements of the scenario provided.

upvoted 2 times

  **Alice1234** 1 year, 4 months ago

C. Amazon Lex: It provides advanced deep learning functionalities of automatic speech recognition (ASR) for converting speech to text, and natural language understanding (NLU) to recognize the intent of the text, enabling you to build applications with highly engaging user experiences and lifelike conversational interactions.

D. Amazon Polly: This service turns text into lifelike speech using deep learning. It would enable the BI application to deliver the answers to the executives' questions in a spoken format.

F. Amazon Transcribe: This is an automatic speech recognition (ASR) service that makes it easy for developers to add speech-to-text capability to their applications. This would be necessary for the BI application to interpret spoken questions from the executives.

upvoted 1 times

  **CloudHandsOn** 1 year, 5 months ago

Selected Answer: CDF

CDF -> CEF. you dont need comprehend in this scenario.

upvoted 3 times

  **CloudHandsOn** 1 year, 5 months ago

Selected Answer: CDF

Amazon Lex (C): This service is crucial for building conversational interfaces. It provides the capabilities to understand and interpret user input in natural language, which is essential for understanding the questions asked by executives.

Amazon Transcribe (F): For a spoken interface, you need a service that can convert speech into text. Amazon Transcribe does exactly this, allowing the system to process spoken questions by converting them into text that can then be interpreted by Amazon Lex.

Amazon Polly (D): To enhance the user experience by responding to inquiries not only in text but also in spoken form, Amazon Polly is ideal. It converts text responses into lifelike speech, allowing the system to verbally communicate with the executives.

Together, these three services (Amazon Lex, Amazon Transcribe, and Amazon Polly) will enable a comprehensive conversational interface for the BI application, catering to both written and spoken queries and responses

upvoted 2 times

🗳️ 👤 **endeesa** 1 year, 7 months ago

Selected Answer: CDF

why does aws use mulipt service for tts and stt?

upvoted 3 times

🗳️ 👤 **sukye** 1 year, 7 months ago

Selected Answer: CDF

No, don't need E Comprehend because the report has already been generated.

upvoted 2 times

🗳️ 👤 **akgarg00** 1 year, 7 months ago

Answer is CEF --> Input can be speech but the output to the user will be text (as nothing specific is mentioned) using Lex for conversational interface, Transcribe to convert speech to text (if input is speech) and Comprehend for insights from text

upvoted 1 times

🗳️ 👤 **elvin_ml_qayiran25091992razor** 1 year, 7 months ago

Selected Answer: CEF

CEF is correct

upvoted 1 times

🗳️ 👤 **DimLam** 1 year, 8 months ago

Selected Answer: CDE

I will go with:

lex for the chat interface

comprehend for getting insights from reports

Polly for text-to-speech transformation

<https://aws.amazon.com/blogs/machine-learning/deriving-conversational-insights-from-invoices-with-amazon-textract-amazon-comprehend-and-amazon-lex/>

upvoted 1 times

🗳️ 👤 **jopaca1216** 1 year, 9 months ago

Amazon Polly is essential for providing spoken responses in a conversational interface, it doesn't directly handle the natural language understanding and processing aspect, which is why it wasn't included as one of the top three services for building the conversational interface in this scenario.

Correct is C, E, F

upvoted 1 times

🗳️ 👤 **loict** 1 year, 9 months ago

Selected Answer: CDF

A. NO - Alexa for Business

B. NO - Amazon Connect for call centers

C. YES - Amazon Lex for chatbots

D. YES - Lex Text-to-Speech

E. NO - Amazon Comprehend is for topic extraction and sentiment analysis, Transcribe already does it

F. YES - Transcribe Speech-to-Text

upvoted 4 times

🗳️ 👤 **AmeeraM** 1 year, 8 months ago

Transcribe does not do sentiment analysis and topic extraction it just generates transcript from speech so we need Amazon Comprehend

upvoted 1 times

🗳️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: CDF

Agree with CDF

upvoted 2 times

A machine learning specialist works for a fruit processing company and needs to build a system that categorizes apples into three types. The specialist has collected a dataset that contains 150 images for each type of apple and applied transfer learning on a neural network that was pretrained on ImageNet with this dataset.

The company requires at least 85% accuracy to make use of the model.

After an exhaustive grid search, the optimal hyperparameters produced the following:

⇒ 68% accuracy on the training set

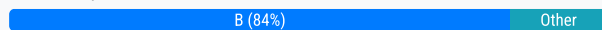
⇒ 67% accuracy on the validation set

What can the machine learning specialist do to improve the system's accuracy?

- A. Upload the model to an Amazon SageMaker notebook instance and use the Amazon SageMaker HPO feature to optimize the model's hyperparameters.
- B. Add more data to the training set and retrain the model using transfer learning to reduce the bias.
- C. Use a neural network model with more layers that are pretrained on ImageNet and apply transfer learning to increase the variance.
- D. Train a new model using the current neural network architecture.

Suggested Answer: B

Community vote distribution



dolorez Highly Voted 2 years, 1 month ago

Selected Answer: B

the answer is B - the model is underfitting = high bias, so we want to reduce it

C is wrong because the intention is not to increase variance which equals overfitting (using a more complex model would be good, but to reduce bias not increase variance)

upvoted 11 times

CloudGyan Most Recent 5 months, 2 weeks ago

Selected Answer: B

The 68% accuracy on the training set and 67% accuracy on the validation set suggest that the model is biased - underfitting and does not have enough capacity or relevant information to learn the underlying patterns in the data.

upvoted 1 times

endeesa 7 months ago

Selected Answer: B

I would think ImageNet network is good enough already, so more data

upvoted 1 times

loict 9 months, 2 weeks ago

Selected Answer: B

A. NO - HPO has already been done though grid search

B. YES - 150 images is very small; need x10 that

C. NO - need bigger training set

D. NO - what would the new model be ?

upvoted 2 times

Mickey321 10 months ago

Selected Answer: B

More data to training set

upvoted 1 times

kaike_reis 11 months ago

Selected Answer: B

Letter B is the correct one. We can add more data with data augmentation. Letter A would be a repetition of what has already been done. Letter C is impractical. Letter D is starting from scratch without need.

upvoted 1 times

🗨️ 👤 **mirik** 1 year ago

Selected Answer: D

I think it should be D: "Train a new model using the current neural network architecture".

Because apples data is very specific and ImageNet weights will be too generic there. We still can leave ImageNet weights for an initial configuration but the model should be retrained from scratch.

upvoted 1 times

🗨️ 👤 **cox1960** 1 year, 2 months ago

Selected Answer: A

450 images should be fine. HPO for me.

upvoted 1 times

🗨️ 👤 **expertguru** 1 year, 5 months ago

BOTH VALIDation set and train set performing equally but performance not good. So the basic problem here is high bias (train error) and high variance (test error). Ideally we want both low, but there is trade-off need to be cautious to avoid overfitting. So this problem needs solution for Low bias first (so training performance improves with decent) for later to figure out whether that leads to overfit or not when you test it! Answer choice B

upvoted 1 times

🗨️ 👤 **deng113jie** 1 year, 10 months ago

why not A?

<https://aws.amazon.com/about-aws/whats-new/2022/07/amazon-sagemaker-automatic-model-tuning-supports-increased-limits-improve-accuracy-models/>

upvoted 2 times

🗨️ 👤 **[Removed]** 2 years ago

not B, c is correct

upvoted 1 times

🗨️ 👤 **edvardo** 2 years, 1 month ago

Given that the model can't even fit the training set properly, it would be convenient to amplify the layers that are trained. If I understood the phrasing correctly, I would go with C.

upvoted 1 times

🗨️ 👤 **Istdanagan** 2 years, 2 months ago

Selected Answer: C

C, accuracy on training set is low, model not complex enough

upvoted 1 times

🗨️ 👤 **spaceexplorer** 2 years, 2 months ago

B is more accurate, while adding more complexity for model is viable but you don't want to increase variance

upvoted 12 times

🗨️ 👤 **NeverMinda** 2 years ago

It only has 150 photos for training, more complex neural network won't help

upvoted 4 times

A company uses camera images of the tops of items displayed on store shelves to determine which items were removed and which ones still remain. After several hours of data labeling, the company has a total of 1,000 hand-labeled images covering 10 distinct items. The training results were poor.

Which machine learning approach fulfills the company's long-term needs?

- A. Convert the images to grayscale and retrain the model
- B. Reduce the number of distinct items from 10 to 2, build the model, and iterate
- C. Attach different colored labels to each item, take the images again, and build the model
- D. Augment training data for each item using image variants like inversions and translations, build the model, and iterate.

Suggested Answer: A

Community vote distribution

D (100%)

🗳️ 👤 **ovokpus** Highly Voted 2 years ago

Selected Answer: D

Data Augmentation is the way to go here.

How does converting to grayscale help? What if the colors of the items are relevant in object identification???

upvoted 11 times

🗳️ 👤 **AmeeraM** Most Recent 8 months, 2 weeks ago

Selected Answer: D

data augemntation

upvoted 1 times

🗳️ 👤 **jopaca1216** 9 months, 2 weeks ago

D is correct

How can I make the decision to use gray images if the question doesn't even indicate whether the images are colored or not? and even so, colored images are important to ensure more accuracy in training than compared to gray imagens.

Due that the model is underfitting, more data like indicated the option D is the correct action.

upvoted 1 times

🗳️ 👤 **mirik** 1 year ago

C: "Attach different colored labels to each item, take the images again, and build the model"

It is also kind of augmentation. It is even better than just inverting and translating existing samples.

upvoted 1 times

🗳️ 👤 **kaike_reis** 11 months ago

But it's done in real life and your manual work would be lost.

upvoted 1 times

🗳️ 👤 **Debayandt91** 1 year, 1 month ago

shouldnt it be reduced to 2 variables , taking image of empty shelf and non empty and that should do it ?

upvoted 1 times

🗳️ 👤 **Sylzys** 1 year, 4 months ago

D is of course the right answer, grayscale only won't help anything

upvoted 3 times

🗳️ 👤 **PHTR** 1 year, 5 months ago

D is the CORRECT ANSWER

<https://research.aimultiple.com/data-augmentation/>

upvoted 1 times

🗨️ 👤 **aScientist** 1 year, 7 months ago

Selected Answer: D

Data augmentation is correct. we need more samples

upvoted 1 times

🗨️ 👤 **tgaos** 2 years ago

D is correct

upvoted 3 times

🗨️ 👤 **cron0001** 2 years, 2 months ago

Selected Answer: D

D is my answer for this. A can help but it'll need more than that.

upvoted 4 times

🗨️ 👤 **Istdanagan** 2 years, 2 months ago

Selected Answer: D

D, i guess

upvoted 4 times

A Data Scientist is developing a binary classifier to predict whether a patient has a particular disease on a series of test results. The Data Scientist has data on 400 patients randomly selected from the population. The disease is seen in 3% of the population. Which cross-validation strategy should the Data Scientist adopt?

- A. A k-fold cross-validation strategy with $k=5$
- B. A stratified k-fold cross-validation strategy with $k=5$
- C. A k-fold cross-validation strategy with $k=5$ and 3 repeats
- D. An 80/20 stratified split between training and validation

Suggested Answer: B

Community vote distribution

B (100%)

🗳️ **scuzzy2010** Highly Voted 3 years, 9 months ago

B - stratified k-fold cross-validation will enforce the class distribution in each split of the data to match the distribution in the complete training dataset.

upvoted 16 times

🗳️ **SophieSu** Highly Voted 3 years, 8 months ago

B is the correct answer. Use Stratified k-Fold Cross-Validation for Imbalanced Classification. Stratified train/test splits is an option too. But the question is specifically asking "cross-validation" strategy.

upvoted 9 times

🗳️ **MultiCloudIronMan** Most Recent 8 months, 1 week ago

Selected Answer: B

In summary, Option B is the most appropriate strategy for handling the imbalanced dataset and ensuring reliable performance metrics for the binary classifier.

upvoted 1 times

🗳️ **Mickey321** 1 year, 10 months ago

Selected Answer: B

for imbalanced data. Stratified k-fold cross-validation ensures that the distribution of the target variable is the same in each fold. This is important for binary classification problems, where the target variable is imbalanced. In this case, the disease is seen in only 3% of the population. This means that if we do not use stratified k-fold cross-validation, then there is a risk that the training and validation sets will not be representative of the actual population.

upvoted 1 times

🗳️ **ADVIT** 1 year, 12 months ago

B

<https://towardsdatascience.com/understanding-8-types-of-cross-validation-80c935a4976d>

upvoted 1 times

🗳️ **Valcilio** 2 years, 3 months ago

Selected Answer: B

Stratified cross validation is for unbalanced data like this!

upvoted 1 times

🗳️ **AWS__Newbie** 3 years, 7 months ago

Why K=5?

upvoted 2 times

🗳️ **eeah** 3 years, 2 months ago

K=5 is just standard

upvoted 1 times

🗳️ **Vita_Rasta84444** 3 years, 8 months ago

Yes, B...

upvoted 1 times

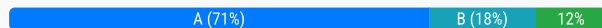
A technology startup is using complex deep neural networks and GPU compute to recommend the company's products to its existing customers based upon each customer's habits and interactions. The solution currently pulls each dataset from an Amazon S3 bucket before loading the data into a TensorFlow model pulled from the company's Git repository that runs locally. This job then runs for several hours while continually outputting its progress to the same S3 bucket. The job can be paused, restarted, and continued at any time in the event of a failure, and is run from a central queue.

Senior managers are concerned about the complexity of the solution's resource management and the costs involved in repeating the process regularly. They ask for the workload to be automated so it runs once a week, starting Monday and completing by the close of business Friday. Which architecture should be used to scale the solution at the lowest cost?

- A. Implement the solution using AWS Deep Learning Containers and run the container as a job using AWS Batch on a GPU-compatible Spot Instance
- B. Implement the solution using a low-cost GPU-compatible Amazon EC2 instance and use the AWS Instance Scheduler to schedule the task
- C. Implement the solution using AWS Deep Learning Containers, run the workload using AWS Fargate running on Spot Instances, and then schedule the task using the built-in task scheduler
- D. Implement the solution using Amazon ECS running on Spot Instances and schedule the task using the ECS service scheduler

Suggested Answer: C

Community vote distribution



jdstone Highly Voted 3 years, 8 months ago

Answer is A

<https://aws.amazon.com/blogs/compute/gpu-workloads-on-aws-batch/>

upvoted 27 times

Juka3lj 3 years, 8 months ago

Makes most sense

upvoted 2 times

astonm13 Highly Voted 3 years, 8 months ago

I would go for D. As far as I know Fargate does not support GPU computing.

upvoted 6 times

Bhadu 1 year, 11 months ago

It does support GPU

<https://docs.aws.amazon.com/batch/latest/userguide/fargate.html>

upvoted 1 times

fa0d8b7 1 year, 6 months ago

this is wrong information. It does not support GPU

upvoted 1 times

teka112233 1 year, 10 months ago

the problem that fargate is serverless which mean you can't control its compute capabilities

upvoted 1 times

MultiCloudIronMan Most Recent 9 months, 1 week ago

Selected Answer: A

To scale the solution at the lowest cost, the best architecture would be Option A: Implement the solution using AWS Deep Learning Containers and run the container as a job using AWS Batch on a GPU-compatible Spot Instance. This approach leverages AWS Batch to manage the job scheduling and execution, while using Spot Instances to significantly reduce costs¹².

Would you like more details on how to set this up or any other aspect of optimizing your architecture?

upvoted 1 times

chewasa 1 year, 3 months ago

Selected Answer: A

fargate doesnt support GPU.

<https://github.com/aws/containers-roadmap/issues/88>

upvoted 1 times

🗨️ **endeesa** 1 year, 7 months ago

Selected Answer: A

AWS batch will easily satisfy the requiremntns

upvoted 1 times

🗨️ **windy9** 1 year, 9 months ago

Fargate doesn't support GPU. So go with AWS Batch and DLC (Deep Learning Container)

upvoted 1 times

🗨️ **loict** 1 year, 9 months ago

Selected Answer: C

A. NO - Fargate provides batch fonctionnalities already fully integrated with ECS

B. NO - too low level

C. YES - AWS Deep Learning Containers are optimized; AWS Fargate is serverless (so less ops complexity); Spot best for cost

D. NO - ECS service scheduler is not serverless

upvoted 1 times

🗨️ **khchan123** 1 year, 7 months ago

Answer is A.

C is not correct. GPU resources aren't supported for jobs that run on Fargate resources.

upvoted 1 times

🗨️ **Shenannigan** 1 year, 10 months ago

A and C are both great answers but when it comes to cost I believe A is the more cost effective solution. So A is my answer

upvoted 1 times

🗨️ **Mickey321** 1 year, 10 months ago

Selected Answer: A

Automate the workload by scheduling the job to run once a week using AWS Batch's built-in scheduler or a cron expression.

Optimize the performance by using AWS Deep Learning Containers that are tailored for GPU acceleration and deep learning frameworks.

Reduce the cost by using Spot Instances that offer significant savings compared to On-Demand Instances.

Handle failures by using AWS Batch's retry strategies that can automatically restart the job on a different instance if the Spot Instance is interrupted.

upvoted 1 times

🗨️ **injoho** 2 years, 2 months ago

Answer is A. (But the question is tricky)

A and D are both correct solutions but pay attention to the words - "Senior managers are concerned about the complexity of the solution's resource management and the costs". With Cost is everything simple - use Spot instances, with resource management - use higher abstraction servicr

AWS Batch is a management/abstraction layer on top of ECS and EC2 (and some other AWS resources). It does some things for you, like cost optimization, that can be difficult to do yourself. Think of it like Elastic Beanstalk for batch operations. It provides a management layer on top of lower-level AWS resources, but if you are comfortable managing those lower level resources yourself and want more control over them it is certainly an option to use those lower-level resources directly.

upvoted 4 times

🗨️ **MIlb** 2 years, 2 months ago

Selected Answer: C

Why not C? AWS Fargate is oriented to manage resources as you need.

upvoted 1 times

🗨️ **austinoy** 2 years, 3 months ago

A looks good to me

<https://aws.amazon.com/blogs/compute/deep-learning-on-aws-batch/>

upvoted 1 times

🗨️ **drcok87** 2 years, 4 months ago

b: for those who think its B because of spot instance interruption, ready question phrase "The job can be paused, restarted, and continued at any time in the event of a failure, and is run from a central queue."

between a and c: at the time of this question i doubt if fargate supported GPU, even if it did I choose aws batch for job and fargate for services/apps that need to run all the time.

a is answer

upvoted 2 times

🗨️ 👤 **Ajose0** 2 years, 4 months ago

Selected Answer: A

Option A is the most cost-effective architecture as it uses GPU-compatible Spot Instance which is the lowest cost compute option for GPU instances in the AWS cloud.

AWS Batch is a fully managed service that schedules, runs, and manages the processing and analysis of batch workloads.

The use of AWS Deep Learning Containers enables the technology startup to use pre-built, optimized Docker containers for deep learning, which reduces the complexity of the solution's resource management and eliminates the need for repeated processing.

upvoted 2 times

🗨️ 👤 **Aninina** 2 years, 6 months ago

Selected Answer: A

Answer is A. Option B is similar to A, but it uses a low-cost GPU-compatible EC2 instance rather than a container, which may not be as flexible or scalable as using containers.

upvoted 4 times

🗨️ 👤 **matteocal** 2 years, 11 months ago

Selected Answer: A

Answer is A

<https://aws.amazon.com/blogs/compute/gpu-workloads-on-aws-batch/>

upvoted 2 times

🗨️ 👤 **ovokpus** 3 years ago

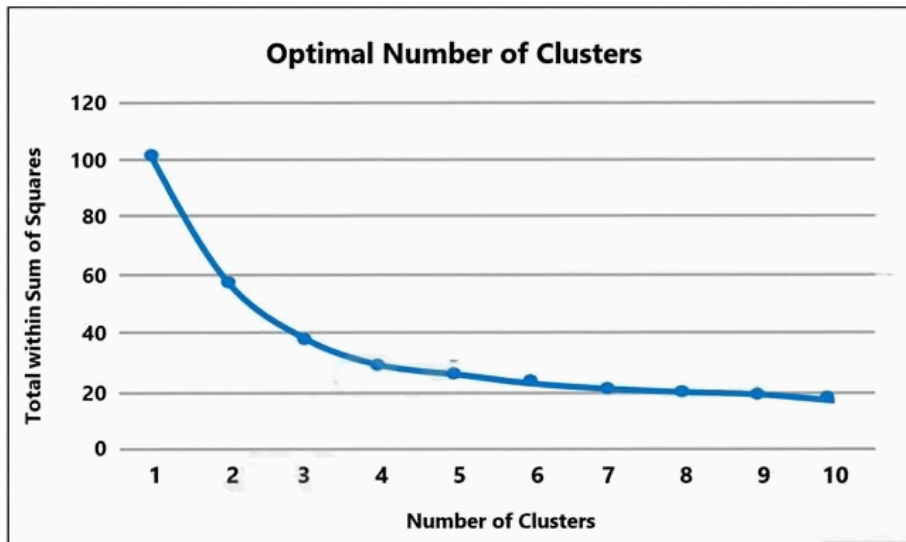
Selected Answer: A

<https://aws.amazon.com/blogs/compute/gpu-workloads-on-aws-batch/>

There you have it

upvoted 1 times

A Machine Learning Specialist prepared the following graph displaying the results of k-means for $k = [1..10]$:



Considering the graph, what is a reasonable selection for the optimal choice of k ?

- A. 1
- B. 4
- C. 7
- D. 10

Suggested Answer: C

Community vote distribution

B (100%)

ksrivastavaSumit Highly Voted 2 years, 9 months ago

B seems correct based on the elbow method
upvoted 16 times

Juka3lj 2 years, 8 months ago

I agree, most likely B.
upvoted 4 times

exam_prep Highly Voted 2 years, 1 month ago

Hi all, I am not able to see the image. It is broken for me. Is it possible for someone to share the image.
upvoted 7 times

endeesa Most Recent 7 months ago

Selected Answer: B

The closest no.to the elbow is clearly 4
upvoted 2 times

cyberfriends 8 months, 1 week ago

Selected Answer: B

B is correct
upvoted 1 times

Mickey321 10 months, 1 week ago

Selected Answer: B

Elbow method
upvoted 1 times

ADVIT 12 months ago

B:
[https://en.wikipedia.org/wiki/Elbow_method_\(clustering\)](https://en.wikipedia.org/wiki/Elbow_method_(clustering))

upvoted 1 times

🗨️ 👤 **Sylzys** 1 year, 4 months ago

Selected Answer: B

Elbow is more visible at 4

upvoted 2 times

🗨️ 👤 **rockingkiran** 1 year, 6 months ago

Selected Answer: B

B seems correct

upvoted 1 times

🗨️ 👤 **Peeking** 1 year, 6 months ago

Selected Answer: B

<https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/>

upvoted 2 times

🗨️ 👤 **Skychaser** 1 year, 11 months ago

Selected Answer: B

Elbow method

upvoted 2 times

🗨️ 👤 **Sadgamaya** 2 years, 3 months ago

. B seems correct.

upvoted 1 times

🗨️ 👤 **mahmoudai** 2 years, 7 months ago

B seems better

upvoted 2 times

🗨️ 👤 **Vita_Rasta84444** 2 years, 7 months ago

number 4 is the elbow of the hand, B is correct

upvoted 4 times

🗨️ 👤 **cnethers** 2 years, 8 months ago

because the elbow method is a heuristic method its open to debate as to where the correct bend in the cluster is. It's a good tool to use with lower computation cost than computing the silhouette score. When looking at <https://www.youtube.com/watch?v=qs8nfzUsW5U> instead of eyeballing where the bend is, he calculates where the difference between scores is smaller than the 90th percentile

upvoted 2 times

A media company with a very large archive of unlabeled images, text, audio, and video footage wishes to index its assets to allow rapid identification of relevant content by the Research team. The company wants to use machine learning to accelerate the efforts of its in-house researchers who have limited machine learning expertise.


Which is the FASTEST route to index the assets?

- A. Use Amazon Rekognition, Amazon Comprehend, and Amazon Transcribe to tag data into distinct categories/classes.
- B. Create a set of Amazon Mechanical Turk Human Intelligence Tasks to label all footage.
- C. Use Amazon Transcribe to convert speech to text. Use the Amazon SageMaker Neural Topic Model (NTM) and Object Detection algorithms to tag data into distinct categories/classes.
- D. Use the AWS Deep Learning AMI and Amazon EC2 GPU instances to create custom models for audio transcription and topic modeling, and use object detection to tag data into distinct categories/classes.

Suggested Answer: A

Community vote distribution

A (100%)

  **SophieSu** Highly Voted 3 years, 9 months ago

A. Fastest route must Amazon Services.

upvoted 27 times

  **AShahine21** 3 years, 8 months ago

Amazon Mechanical Turk is an Amazon service

upvoted 2 times

  **Ioict** Most Recent 1 year, 9 months ago

Selected Answer: A

A. YES - Rekognition with built-in labels for images & video, Transcribe to convert sound to text and Comprehend for Topic Modeling

B. NO - complicated

C. NO - complicated

D. NO - complicated



upvoted 1 times

  **Mickey321** 1 year, 10 months ago

Selected Answer: A

AWS services for fastest route

upvoted 1 times

  **angus** 2 years, 1 month ago

why not C?


C. Use Amazon Transcribe to convert speech to text. Use the Amazon SageMaker Neural Topic Model (NTM) and Object Detection algorithms to tag data into distinct categories/classes.

upvoted 1 times

  **winstonmcgee69** 1 year ago



this takes time and u need to have atleast some technical ML expertise

upvoted 1 times

  **Flysun** 2 years, 3 months ago

I will choose B <https://aws.amazon.com/cn/getting-started/hands-on/machine-learning-tutorial-label-training-data/>



upvoted 1 times

  **Valcilio** 2 years, 3 months ago

Selected Answer: A

Mechanical Turk is the most accurate, but the three services in letter A is the fastest!

upvoted 1 times

  **Ajose0** 2 years, 4 months ago

Selected Answer: A

A. Use Amazon Rekognition, Amazon Comprehend, and Amazon Transcribe to tag data into distinct categories/classes is the fastest route to index the assets. These AWS services provide pre-built machine learning models that can be used to tag the content in the archive without the need for building custom models from scratch.

This option would be faster than using custom models with the AWS Deep Learning AMI and Amazon EC2 GPU instances, or using Amazon Mechanical Turk for human labeling.

Additionally, the use of pre-built models reduces the need for machine learning expertise, aligning with the company's goal of accelerating efforts by its in-house researchers.

upvoted 2 times

🗳️ 👤 **VinceCar** 2 years, 7 months ago

A. Option B is for those without ML experience. But "researchers who have limited machine learning expertise", so A is better.

upvoted 2 times

🗳️ 👤 **hess** 3 years, 5 months ago

A. The most straight forward use of services.

upvoted 2 times

🗳️ 👤 **YJ4219** 3 years, 8 months ago

I would have said B, but in B it says "label footage" which means it ignored the rest of the data, so i'd go with A

upvoted 1 times

🗳️ 👤 **Madwyn** 3 years, 8 months ago

The question said "a very large archive" meaning a lot of money to pay for labour. B won't be as fast as machine, plus you only label the footage, ignored other stuff.

upvoted 2 times

🗳️ 👤 **AjithkumarSL** 3 years, 8 months ago

Would go for A

upvoted 1 times

🗳️ 👤 **AShahine21** 3 years, 9 months ago

I will go with B

upvoted 1 times

🗳️ 👤 **Juka3lj** 3 years, 9 months ago

Correct answer is A

upvoted 3 times

🗳️ 👤 **ksrivastavaSumit** 3 years, 9 months ago

B. as no one in-house is an expert and It probably is the fastest way to get there

upvoted 2 times

🗳️ 👤 **AhmedAbuMusa** 3 years, 8 months ago

Take into consideration that it is "a very large archive"

upvoted 1 times

A Machine Learning Specialist is working for an online retailer that wants to run analytics on every customer visit, processed through a machine learning pipeline.

The data needs to be ingested by Amazon Kinesis Data Streams at up to 100 transactions per second, and the JSON data blob is 100 KB in size. What is the MINIMUM number of shards in Kinesis Data Streams the Specialist should use to successfully ingest this data?

- A. 1 shards
- B. 10 shards
- C. 100 shards
- D. 1,000 shards

Suggested Answer: B



Community vote distribution

B (100%)

  **[Removed]**  3 years, 2 months ago

Agreed, B it is. See <https://medium.com/slalom-data-analytics/amazon-kinesis-data-streams-auto-scaling-the-number-of-shards-105dc967bed5>

One shard can ingest 1 MB/second or 1,000 records/second. So $100 \text{ KB} \times 100 = 10 \text{ MB}$ (10 shards required)
upvoted 26 times

  **james2033**  9 months, 3 weeks ago

Selected Answer: B
 $100 \text{ KB} \times 100 = 10 \text{ MB}$
 1 MB/second
 $10 / 1 = 10 \text{ shards}$.
 upvoted 1 times

  **Mickey321** 1 year, 4 months ago

Selected Answer: B
 Each shard in Amazon Kinesis Data Streams can support up to 1,000 transactions per second.
 The data needs to be ingested at up to 100 transactions per second, so we need at least 1 shard.
 However, we also need to consider the size of the JSON data blob. Each JSON data blob is 100 KB in size, and each shard can only store up to 1 MB of data.
 This means that we need to have at least 10 shards, so that each shard can store 100 KB of data.
 upvoted 1 times



  **PHTR** 1 year, 11 months ago

B - Max. ingestion per shard = 1000 KB/s


 $\rightarrow 100 \text{ Records} \times 100 \text{ KB} = 10.000 \text{ KB}$
 $\rightarrow 10.000 \text{ KB} / 1000 \text{ KB/per Shard} = 10 \text{ Shards}$
 upvoted 3 times

  **Shailendraa** 2 years, 3 months ago



10 should be correct.
 upvoted 1 times

  **tgaos** 2 years, 6 months ago

Selected Answer: B
 B is correct
 upvoted 2 times

  **mahmoudai** 3 years, 1 month ago

$100 \text{ kb} \times 100 \text{ t/second} = 10000 \text{ kb} = 10 \text{ mb}$
 $10 \text{mb} / \text{max_threshold_per_shard} (1 \text{ mb}) = 10 \text{ shards}$
 upvoted 1 times

  **benson2021** 3 years, 2 months ago

Reference: <https://docs.aws.amazon.com/streams/latest/dev/service-sizes-and-limits.html>

upvoted 2 times

A Machine Learning Specialist is deciding between building a naive Bayesian model or a full Bayesian network for a classification problem. The Specialist computes the Pearson correlation coefficients between each feature and finds that their absolute values range between 0.1 to 0.95. Which model describes the underlying data in this situation?

- A. A naive Bayesian model, since the features are all conditionally independent.
- B. A full Bayesian network, since the features are all conditionally independent.
- C. A naive Bayesian model, since some of the features are statistically dependent.
- D. A full Bayesian network, since some of the features are statistically dependent.

Suggested Answer: C

Community vote distribution

D (100%)

🗳️ **[Removed]** Highly Voted 2 years, 9 months ago

I would say D, because of correlations and dependencies between features. See <https://towardsdatascience.com/basics-of-bayesian-network-79435e11ae7b> and <https://www.quora.com/Whats-the-difference-between-a-naive-Bayes-classifier-and-a-Bayesian-network?share=1>
upvoted 24 times

🗳️ **Juka3lj** 2 years, 8 months ago

I agree, makes most sense
upvoted 1 times

🗳️ **Vita_Rasta84444** Highly Voted 2 years, 8 months ago

It should be D. Naive Bayes is called naive because it assumes that each input variable is independent. This is a strong assumption and unrealistic for real data; however, the technique is very effective on a large range of complex problems.
upvoted 9 times

🗳️ **Mickey321** Most Recent 10 months, 1 week ago

Selected Answer: D

In this case, the absolute values of the Pearson correlation coefficients range between 0.1 to 0.95. This means that some of the features are statistically dependent. Therefore, a full Bayesian network is a better model for the underlying data than a naive Bayesian model.
upvoted 1 times

🗳️ **Ajose0** 1 year, 4 months ago

Selected Answer: D

In a full Bayesian network, features are connected to each other by edges that represent their conditional dependence relationships. A full Bayesian network is useful when the relationships between the features are complex, non-linear or when they are not conditionally independent.

In this situation, where the Pearson correlation coefficients range between 0.1 and 0.95, it suggests that there are dependencies between the features, indicating that a full Bayesian network would be appropriate to capture the relationships between the features and model the data.
upvoted 5 times

🗳️ **ystotest** 1 year, 7 months ago

Selected Answer: D

distinction between Bayes theorem and Naive Bayes is that Naive Bayes assumes conditional independence where Bayes theorem does not. This means the relationship between all input features are independent. The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between -1 and 1 that measures the strength and direction of the relationship between two variables.
upvoted 3 times

🗳️ **Shailendraa** 1 year, 10 months ago

A naive Bayesian model, since some of the features, are statistically dependent.
upvoted 2 times

🗳️ **SophieSu** 2 years, 8 months ago

D. Naive bayes - features are independent given the class.
upvoted 6 times

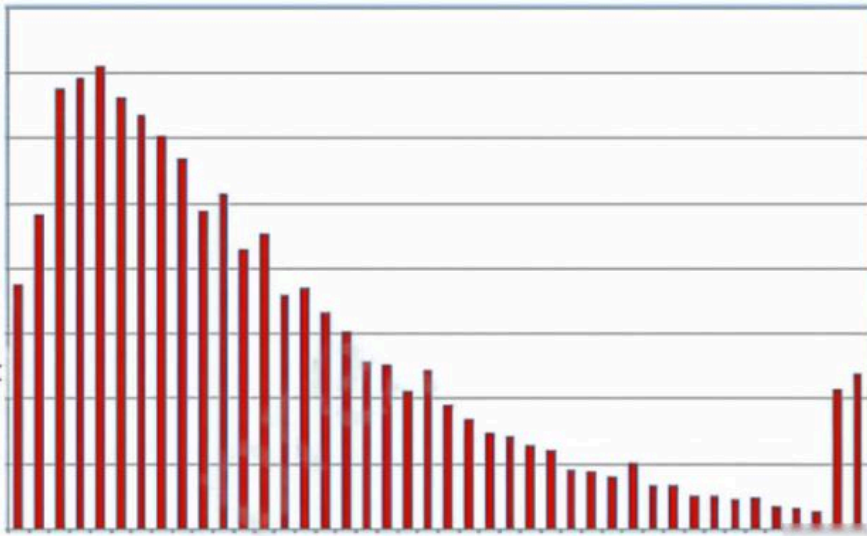
🗳️ **astonm13** 2 years, 8 months ago

I would say, B. Naive Bayes assumes conditional independence and not statistical
upvoted 2 times

🗨️ 👤 **abdohanfi** 2 years, 7 months ago
you mean (a) naive bayes not (b)
upvoted 1 times

🗨️ 👤 **cnethers** 2 years, 8 months ago
This is also a good source of information to help build your understanding <https://www.simplypsychology.org/correlation.html>
upvoted 1 times

A Data Scientist is building a linear regression model and will use resulting p-values to evaluate the statistical significance of each coefficient. Upon inspection of the dataset, the Data Scientist discovers that most of the features are normally distributed. The plot of one feature in the dataset is shown in the graphic.



What transformation should the Data Scientist apply to satisfy the statistical assumptions of the linear regression model?

- A. Exponential transformation
- B. Logarithmic transformation
- C. Polynomial transformation
- D. Sinusoidal transformation

Suggested Answer: A

Community vote distribution

B (100%)

astonm13 Highly Voted 3 years, 3 months ago

I would say B. Logarithmic transformation converts skewed distributions towards normal
upvoted 20 times

AIWave Most Recent 10 months, 2 weeks ago

I would go with B. For right skewed distributions -> Logarithmic transformation
For left skewed distributions -> exponential transformations=
upvoted 1 times

Mickey321 1 year, 4 months ago

Selected Answer: B

The linear regression model assumes that the errors are normally distributed. The plot of the feature shows that the errors are not normally distributed.

The logarithmic transformation can be used to transform the errors to be normally distributed.

The exponential transformation, polynomial transformation, and sinusoidal transformation cannot be used to transform the errors to be normally distributed.

upvoted 1 times

ccpmad 1 year, 5 months ago

Selected Answer: B

B

when the feature data is not normally distributed, applying a logarithmic transformation can help to normalize the data and satisfy the assumptions of the linear regression model.

upvoted 1 times

cpal012 1 year, 8 months ago

'A' would make it considerably worse.

upvoted 1 times

🗨️ **goku58** 1 year, 6 months ago

Exponential transformation would make it exponentially worse. :D

upvoted 3 times

🗨️ **Zhechen0912** 1 year, 9 months ago

Selected Answer: B

Log Normal Distribution => Log() => Normal Distribution

upvoted 2 times

🗨️ **sqavi** 1 year, 10 months ago

Selected Answer: B

B is correct answer

upvoted 1 times

🗨️ **vetaal** 2 years, 11 months ago

Selected Answer: B

This is B, as this feature seems skewed while others have a regular distribution according to the question. The log transformation will reduce this features skewness.

upvoted 2 times

🗨️ **YJ4219** 3 years, 2 months ago

I think it's B.

reference: <https://corporatefinanceinstitute.com/resources/knowledge/other/positively-skewed-distribution/#:~:text=For%20positively%20skewed%20distributions%2C%20the,each%20value%20in%20the%20dataset.>

"For positively skewed distributions, the most popular transformation is the log transformation. The log transformation implies the calculations of the natural logarithm for each value in the dataset. The method reduces the skew of a distribution. Statistical tests are usually run only when the transformation of the data is complete."

upvoted 2 times

🗨️ **konradL** 3 years, 3 months ago

I would also go for B, as Log transformation is often mentioned, when we are talking about right (positive) skewness.

upvoted 3 times

A Machine Learning Specialist is assigned to a Fraud Detection team and must tune an XGBoost model, which is working appropriately for test data. However, with unknown data, it is not working as expected. The existing parameters are provided as follows.

```
param = {
    'eta': 0.05, # the training step for each iteration
    'silent': 1, # logging mode - quiet
    'n_estimators': 2000,
    'max_depth': 30,
    'min_child_weight': 3,
    'gamma': 0,
    'subsample': 0.8,
    'objective': 'multi:softprob', # error evaluation for multiclass training
    'num_class': 201} # the number of classes that exist in this dataset
num_round = 60 # the number of training iterations
```

Which parameter tuning guidelines should the Specialist follow to avoid overfitting?

- A. Increase the max_depth parameter value.
- B. Lower the max_depth parameter value.
- C. Update the objective to binary:logistic.
- D. Lower the min_child_weight parameter value.

Suggested Answer: B

Community vote distribution

B (100%)

🗳️ 👤 **SophieSu** Highly Voted 2 years, 8 months ago

B lower max_depth is the correct answer.

D min_child_weight means something like "stop trying to split once your sample size in a node goes below a given threshold"

Lower min_child_weight, the tree becomes more deep and complex.

Increase min_child_weight, the tree will have less branches and less complexity.

upvoted 17 times

🗳️ 👤 **Mickey321** Most Recent 10 months, 1 week ago

Selected Answer: B

The max_depth parameter controls the maximum depth of the decision trees in the XGBoost model. A higher max_depth value will result in more complex decision trees, which can lead to overfitting.

upvoted 1 times

🗳️ 👤 **ccpmad** 11 months ago

Selected Answer: B

Overfitting occurs when a model performs well on the training data but poorly on unseen or test data. In the context of XGBoost, reducing the max_depth parameter helps prevent overfitting. The max_depth parameter controls the maximum depth of the trees in the ensemble. A smaller max_depth value limits the complexity of the trees, making them less likely to memorize the noise in the training data and improve generalization to unseen data.

upvoted 1 times

🗳️ 👤 **gcaria** 1 year ago

Selected Answer: B

It is B

upvoted 1 times

🗳️ 👤 **vbal** 1 year ago

B: overfitting problem.

upvoted 1 times

🗳️ 👤 **Shailendraa** 1 year, 9 months ago

12-Sep Exam.

upvoted 1 times

🗳️ 👤 **[Removed]** 2 years, 7 months ago

Selected Answer: B

When a model overfits, the solutions are:

1. Reduce model flexibility and complexity
2. Reduce the number of feature combinations
3. Decrease n-grams size
4. Decrease the number of numeric attribute bins
5. Increase the amount of regularization
6. Add dropout

upvoted 1 times

🗳️ 👤 **Dr_Kiko** 2 years, 8 months ago

B. 30-deep tree is crazy; normally it's 6-7 no more

upvoted 2 times

🗳️ 👤 **cnethers** 2 years, 8 months ago

A. Increase the max_depth parameter value. (This would increase the complexity resulting in overfitting)

B. Lower the max_depth parameter value. (This would reduce the complexity and minimize overfitting)

C. Update the objective to binary:logistic. it depends on what the target(s) generally you would have a binary classification for fraud detection but there is nothing to say you can't have a multi class so there is not enough information given.

D. Lower the min_child_weight parameter value. (This would reduce the complexity and minimize overfitting)

I find that there are 2 correct answers to this question which does not help B & D

upvoted 2 times

🗳️ 👤 **arulrajayaraj** 2 years, 8 months ago

Ans : B , Lower values avoid over-fitting.

No for D - Larger values avoid over-fitting.

upvoted 8 times

🗳️ 👤 **cnethers** 2 years, 9 months ago

Thus, those parameters can be used to control the complexity of the trees. It is important to tune them together in order to find a good trade-off between model bias and variance

upvoted 2 times

🗳️ 👤 **cnethers** 2 years, 9 months ago

min_child_weight is the minimum weight (or number of samples if all samples have a weight of 1) required in order to create a new node in the tree. A smaller min_child_weight allows the algorithm to create children that correspond to fewer samples, thus allowing for more complex trees, but again, more likely to overfit.

upvoted 1 times

🗳️ 👤 **cnethers** 2 years, 9 months ago

max_depth is the maximum number of nodes allowed from the root to the farthest leaf of a tree. Deeper trees can model more complex relationships by adding more nodes, but as we go deeper, splits become less relevant and are sometimes only due to noise, causing the model to overfit.

upvoted 2 times

A data scientist is developing a pipeline to ingest streaming web traffic data. The data scientist needs to implement a process to identify unusual web traffic patterns as part of the pipeline. The patterns will be used downstream for alerting and incident response. The data scientist has access to unlabeled historic data to use, if needed.

The solution needs to do the following:

⇒ Calculate an anomaly score for each web traffic entry.

Adapt unusual event identification to changing web patterns over time.

▪

Which approach should the data scientist implement to meet these requirements?

- A. Use historic web traffic data to train an anomaly detection model using the Amazon SageMaker Random Cut Forest (RCF) built-in model. Use an Amazon Kinesis Data Stream to process the incoming web traffic data. Attach a preprocessing AWS Lambda function to perform data enrichment by calling the RCF model to calculate the anomaly score for each record.
- B. Use historic web traffic data to train an anomaly detection model using the Amazon SageMaker built-in XGBoost model. Use an Amazon Kinesis Data Stream to process the incoming web traffic data. Attach a preprocessing AWS Lambda function to perform data enrichment by calling the XGBoost model to calculate the anomaly score for each record.
- C. Collect the streaming data using Amazon Kinesis Data Firehose. Map the delivery stream as an input source for Amazon Kinesis Data Analytics. Write a SQL query to run in real time against the streaming data with the k-Nearest Neighbors (kNN) SQL extension to calculate anomaly scores for each record using a tumbling window.
- D. Collect the streaming data using Amazon Kinesis Data Firehose. Map the delivery stream as an input source for Amazon Kinesis Data Analytics. Write a SQL query to run in real time against the streaming data with the Amazon Random Cut Forest (RCF) SQL extension to calculate anomaly scores for each record using a sliding window.

Suggested Answer: A

Community vote distribution

D (100%)

🗳️ **jiadong** Highly Voted 2 years, 9 months ago

I think the answer is D - RCF works together with Data Analytics, and sliding window helped on new information
upvoted 24 times

🗳️ **SophieSu** 2 years, 8 months ago

better to say "RCF is a built-in algorithm/function in Kinesis Data Analytics"
upvoted 3 times

🗳️ **Mickey321** Most Recent 10 months, 1 week ago

Selected Answer: D

D uses the built-in RCF algorithm, which is designed for anomaly detection on streaming data and can adapt to changing patterns over time. It does not require any training data or preprocessing steps, as the RCF algorithm can learn from the streaming data directly.

It uses a sliding window, which allows for continuous updating of the anomaly scores based on the most recent data points.

It leverages the Amazon Kinesis Data Analytics service, which provides a scalable and managed platform for running SQL queries on streaming data.

Option A requires training an RCF model on historic data, which may not reflect the current web traffic patterns. It also adds complexity and latency by invoking a Lambda function for each record.

upvoted 4 times

🗳️ **Mickey321** 10 months, 1 week ago

Selected Answer: D

Answer D

upvoted 1 times

🗳️ **kaike_reis** 11 months ago

Selected Answer: D

Letra B está descartada, pois trás um modelo supervisionado de classificação para um problema não supervisionado. Letra C trás outro modelo que não é recomendado também, em comparação ao RCF. A solução mais fácil de implementar e que atinge os critérios pedidos é a Letra D. Letra A está errada, pois usamos KDS para ingestão apenas.

upvoted 1 times

🗳️ 👤 **ccpmad** 11 months ago

Selected Answer: D

the data scientist needs to identify unusual web traffic patterns in real-time and adapt to changing web patterns over time. Amazon Kinesis Data Analytics provides real-time analytics capabilities on streaming data. The Amazon Random Cut Forest (RCF) SQL extension is designed for anomaly detection in streaming data, which fits the requirement to calculate an anomaly score for each web traffic entry.

upvoted 2 times

🗳️ 👤 **Sidekick** 1 year, 10 months ago

Answer is D

"The algorithm starts developing the machine learning model using current records in the stream when you start the application. The algorithm does not use older records in the stream for machine learning, nor does it use statistics from previous executions of the application."

<https://docs.aws.amazon.com/kinesisanalytics/latest/sqlref/sqlrf-random-cut-forest.html>

upvoted 4 times

🗳️ 👤 **apprehensive_scar** 2 years, 4 months ago

Selected Answer: D

D it is. easy one

upvoted 2 times

🗳️ 👤 **vetaal** 2 years, 5 months ago

Selected Answer: D

RCF is dynamic and adapts with time. D seems more appropriate.

upvoted 3 times

🗳️ 👤 **hess** 2 years, 5 months ago

It is A. The only way to handle the historic data is using sagemaker and you can preprocess a data stream using a lambda.

upvoted 2 times

🗳️ 👤 **ZSun** 1 year, 2 months ago

But, the question does not require using historical data. BTW, it only has unlabeled historic data, and unlabeled data is not really useful training a detection model.

upvoted 1 times

🗳️ 👤 **AMEJack** 2 years, 7 months ago

Definitely D, Data Analytics is using RCF, Using window for selecting data with SQL

upvoted 1 times

🗳️ 👤 **Huy** 2 years, 7 months ago

One more reason to select D, not A, is there is no Lambda function to preprocess record in Kinesis Data Stream.

upvoted 1 times

🗳️ 👤 **DimLam** 8 months, 2 weeks ago

That's not true: <https://docs.aws.amazon.com/kinesisanalytics/latest/dev/lambda-preprocessing.html>

upvoted 1 times

🗳️ 👤 **gbrnq** 2 years, 7 months ago

"Adapt unusual event identification to changing web patterns over time." -> option A does not satisfy this, only mentions build the model once

upvoted 2 times

🗳️ 👤 **randomnamer** 2 years, 8 months ago

The data scientist has access to unlabeled historic data to use, if needed. D has no mention of this. Also, A says the lambda function provides data enrichment. For me it's A.

upvoted 4 times

🗳️ 👤 **seanLu** 2 years, 8 months ago

A and D both seems to works. But A does not satisfy requirement 2, adapt to patterns over time. Since the model is only trained on old data. So D may be better.

upvoted 3 times

🗳️ 👤 **astonm13** 2 years, 9 months ago

It is definitely D

upvoted 1 times

A Data Scientist received a set of insurance records, each consisting of a record ID, the final outcome among 200 categories, and the date of the final outcome.

Some partial information on claim contents is also provided, but only for a few of the 200 categories. For each outcome category, there are hundreds of records distributed over the past 3 years. The Data Scientist wants to predict how many claims to expect in each category from month to month, a few months in advance.

What type of machine learning model should be used?

- A. Classification month-to-month using supervised learning of the 200 categories based on claim contents.
- B. Reinforcement learning using claim IDs and timestamps where the agent will identify how many claims in each category to expect from month to month.
- C. Forecasting using claim IDs and timestamps to identify how many claims in each category to expect from month to month.
- D. Classification with supervised learning of the categories for which partial information on claim contents is provided, and forecasting using claim IDs and timestamps for all other categories.

Suggested Answer: D

Community vote distribution

C (100%)

JBX2010 Highly Voted 2 years, 9 months ago

I think it should be C as the final outcome among 200 categories is already know. No need to build a classification model. It's pure forecasting problem.

upvoted 23 times

abdohanfi 2 years, 8 months ago

he said for a few what about the unclassified many i think we need to make classification for the rest first as it will help us with forecasting later with month to month forecasting

upvoted 2 times

SophieSu Highly Voted 2 years, 9 months ago

C is my answer.

No need to do classification. Because you know whether the insurance has a claim or not in the dataset. The claim contents do not provide additional information.

upvoted 7 times

Mickey321 Most Recent 10 months, 1 week ago

Selected Answer: C

forecasting

upvoted 2 times

kaike_reis 11 months ago

Selected Answer: C

It's pure forecasting problem.

upvoted 1 times

Chelseajcole 1 year, 4 months ago

I would say no machine learning model needed at all. Just using count group by categories SQL is enough

upvoted 1 times

drcok87 1 year, 4 months ago

FinalOutcome

1

2

.

200

RecordID, FinalOutcome, Date, ClaimContents

1

2

.

100000

Note: claim content has partial information, only for few of 200 categories

predict how many claims to expect in each category from month to month, a few months in advance

We dont need the claim contents, we have all we need from first 3 columns to train a forecast model

c

upvoted 1 times

🗳️ 👤 **AjoseO** 1 year, 4 months ago

Selected Answer: C

Forecasting using claim IDs and timestamps to identify how many claims in each category to expect from month to month.

The problem requires the prediction of the number of claims in each category for each month, which is a time series forecasting problem. The timestamps and record IDs can be used to model the underlying patterns in the data, and the model can be trained to predict the number of claims in each category for future months based on these patterns.

While the claim contents might provide additional information, the fact that partial information is only available for a few categories suggests that this information might not be enough to build a robust model, and that it might not be possible to apply supervised learning to all 200 categories. Instead, the model should be trained on the time series data (claim IDs and timestamps) for all categories, and the claim contents can be used to improve the accuracy of the model only for the categories for which such information is available.

upvoted 4 times

🗳️ 👤 **informatica** 1 year, 5 months ago

how can a forecasting/classification model can be based on the claim ID? (that should be unique)

upvoted 1 times

🗳️ 👤 **matteocal** 1 year, 11 months ago

Selected Answer: C

it's a forecasting problem, not a classification one

upvoted 2 times

🗳️ 👤 **Dr_Kiko** 2 years, 8 months ago

predict how many claims to expect in each category from month to month, a few months in advance

C is the only one mentioning forecasting

upvoted 2 times

🗳️ 👤 **kezzzzz** 2 years, 8 months ago

D is correct. Multi-label classification to impute the missing claim contents, then forecasting what we want. C is missing the imputation part.

upvoted 5 times

🗳️ 👤 **f4bi4n** 2 years ago

The question is, can we get something useful out of the handful of 200s and will this impact the forecast as we could forecast the numbers without...

upvoted 1 times

🗳️ 👤 **randomnamer** 2 years, 8 months ago

It is true that the final outcome is known. But C does not use the partial information from the 200 categories. Reinforcement learning currently is state of the art in stock prediction and other time series. Why waste valuable information? For me it's B.

upvoted 2 times

🗳️ 👤 **cnethers** 2 years, 9 months ago

This is a supervised learning approach:

Supervised learning problems can be further grouped into regression and classification problems.

Classification: A classification problem is when the output variable is a category, such as "red" and "blue" or "disease" and "no disease."

Regression: A regression problem is when the output variable is a real value, such as "dollars" or "weight."

upvoted 2 times

A company that promotes healthy sleep patterns by providing cloud-connected devices currently hosts a sleep tracking application on AWS. The application collects device usage information from device users. The company's Data Science team is building a machine learning model to predict if and when a user will stop utilizing the company's devices. Predictions from this model are used by a downstream application that determines the best approach for contacting users.

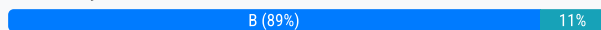
The Data Science team is building multiple versions of the machine learning model to evaluate each version against the company's business goals. To measure long-term effectiveness, the team wants to run multiple versions of the model in parallel for long periods of time, with the ability to control the portion of inferences served by the models.

Which solution satisfies these requirements with MINIMAL effort?

- A. Build and host multiple models in Amazon SageMaker. Create multiple Amazon SageMaker endpoints, one for each model. Programmatically control invoking different models for inference at the application layer.
- B. Build and host multiple models in Amazon SageMaker. Create an Amazon SageMaker endpoint configuration with multiple production variants. Programmatically control the portion of the inferences served by the multiple models by updating the endpoint configuration.
- C. Build and host multiple models in Amazon SageMaker Neo to take into account different types of medical devices. Programmatically control which model is invoked for inference based on the medical device type.
- D. Build and host multiple models in Amazon SageMaker. Create a single endpoint that accesses multiple models. Use Amazon SageMaker batch transform to control invoking the different models through the single endpoint.

Suggested Answer: D

Community vote distribution



SophieSu Highly Voted 3 years, 2 months ago

B is the correct answer.

A/B testing with Amazon SageMaker is required in the Exam.

In A/B testing, you test different variants of your models and compare how each variant performs.

Amazon SageMaker enables you to test multiple models or model versions behind the `same endpoint` using `production variants`.

Each production variant identifies a machine learning (ML) model and the resources deployed for hosting the model.

To test multiple models by `distributing traffic` between them, specify the `percentage of the traffic` that gets routed to each model by specifying the `weight` for each `production variant` in the endpoint configuration.

upvoted 43 times

[Removed] Highly Voted 3 years, 2 months ago

I would answer B, it seems similar to this AWS example: <https://docs.aws.amazon.com/sagemaker/latest/dg/model-ab-testing.html#model-testing-target-variant>

upvoted 10 times

Mickey321 Most Recent 1 year, 4 months ago

Selected Answer: B

Option B

upvoted 1 times

Ajose0 1 year, 10 months ago

Selected Answer: B

This solution allows the Data Science team to build and host multiple models in Amazon SageMaker, which is a fully managed service for training, deploying, and managing machine learning models.

The team can then create an endpoint configuration with multiple production variants, which are different versions of the models. By programmatically updating the endpoint configuration, the team can control the portion of inferences served by the different models.

This allows them to evaluate the models against their business goals and measure their long-term effectiveness without having to make changes at the application layer.

upvoted 3 times

🗳️ 👤 **Morsa** 2 years, 5 months ago

Selected Answer: D

Answer D as it is said "the team intends to run numerous versions in parallel for extended periods of time," so batch transform

upvoted 1 times

🗳️ 👤 **cpal012** 1 year, 8 months ago

How can you create a single endpoint for batch transforms? this answer is nonsensical.

upvoted 1 times

🗳️ 👤 **VR10** 10 months, 1 week ago

It is possible to create a single endpoint for AWS Batch transforms. Here are the key steps:

Create an interface endpoint for AWS Batch in your VPC using the AWS CLI or console. The endpoint service name will be in the format of `com.amazonaws.<region>.batch`

.

When creating the endpoint, assign an IAM role with necessary permissions to make calls to the Batch API.

You can then submit batch transform jobs to AWS Batch referencing resources in both public and private subnets of the VPC. The endpoint ensures private connectivity to Batch.

The single endpoint allows chaining multiple transforms together in a pipeline efficiently without needing internet access. New transforms can be added without redeploying the endpoint.

AWS Batch will automatically provision the required compute environments like EC2 instances or containers to run the transforms and scale as needed based on job requirements.

upvoted 1 times

🗳️ 👤 **[Removed]** 2 years, 6 months ago

it says,"host a sleep monitoring application", it is the host which means online, not batch, b is correct

upvoted 1 times

🗳️ 👤 **John_Pongthorn** 2 years, 10 months ago

Selected Answer: B

The possibility to alter the percentage of inferences supplied by the models.

Which method achieves these criteria with the LEAST amount of effort?

upvoted 4 times

🗳️ 👤 **apprehensive_scar** 2 years, 11 months ago

B. Easy

upvoted 2 times

🗳️ 👤 **anttan** 3 years ago

Think anser is D, below is from the Sagemaker doc.

"<https://docs.aws.amazon.com/sagemaker/latest/dg/batch-transform.html>"

Use Batch Transform to Test Production Variants

To test different models or various hyperparameter settings, create a separate transform job for each new model variant and use a validation dataset. For each transform job, specify a unique model name and location in Amazon S3 for the output file. To analyze the results, use Inference Pipeline Logs and Metrics.

upvoted 4 times

🗳️ 👤 **[Removed]** 3 years ago

The question talks about the LEAST amount of effort. In this case, there will be as many transform jobs required to be built as there are variants.

That may not be the least amount of effort.

upvoted 2 times

An agricultural company is interested in using machine learning to detect specific types of weeds in a 100-acre grassland field. Currently, the company uses tractor-mounted cameras to capture multiple images of the field as 10×10 grids. The company also has a large training dataset that consists of annotated images of popular weed classes like broadleaf and non-broadleaf docks.

The company wants to build a weed detection model that will detect specific types of weeds and the location of each type within the field. Once the model is ready, it will be hosted on Amazon SageMaker endpoints. The model will perform real-time inferencing using the images captured by the cameras.

Which approach should a Machine Learning Specialist take to obtain accurate predictions?

- A. Prepare the images in RecordIO format and upload them to Amazon S3. Use Amazon SageMaker to train, test, and validate the model using an image classification algorithm to categorize images into various weed classes.
- B. Prepare the images in Apache Parquet format and upload them to Amazon S3. Use Amazon SageMaker to train, test, and validate the model using an object- detection single-shot multibox detector (SSD) algorithm.
- C. Prepare the images in RecordIO format and upload them to Amazon S3. Use Amazon SageMaker to train, test, and validate the model using an object- detection single-shot multibox detector (SSD) algorithm.
- D. Prepare the images in Apache Parquet format and upload them to Amazon S3. Use Amazon SageMaker to train, test, and validate the model using an image classification algorithm to categorize images into various weed classes.

Suggested Answer: C

Community vote distribution

C (100%)

🗳️ 👤 **SophieSu** Highly Voted 3 years, 8 months ago

C is my answer.

Pay attention that the question is asking for 2 things:

1. detect specific types of weeds
2. detect the location of each type within the field.

Image Classification can only classify images.

Object detection algorithm:

1. identifies all instances of objects within the image scene.
2. its location and scale in the image are indicated by a rectangular bounding box.

Data format for Computer Vision algorithms in SageMaker:

Recommend to use RecordIO.

upvoted 33 times

🗳️ 👤 **MultiCloudIronMan** Most Recent 9 months, 1 week ago

Selected Answer: C

RecordIO Format: This format is efficient for storing and processing large datasets, which is beneficial for training deep learning models.

Object Detection SSD Algorithm: This algorithm is designed to detect and locate multiple objects within an image, making it ideal for identifying and pinpointing various types of weeds in the field

upvoted 1 times

🗳️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: C

Record IO preferred and also object detection due to several types of weeds

upvoted 1 times

🗳️ 👤 **Ajose0** 2 years, 4 months ago

Selected Answer: C

The goal is to detect specific types of weeds and their locations within a field, which is a task that requires object detection, rather than image classification. Object detection algorithms are designed to identify objects and their locations within an image, whereas image classification algorithms only categorize an entire image into various classes.

Single-shot multibox detectors (SSD) are a type of object detection algorithm that are well-suited for real-time inferencing and have been shown to be

effective for a variety of object detection tasks.

By preparing the images in RecordIO format and using Amazon SageMaker, the company can easily train, test, and validate the model, making it easier to deploy the model in a scalable and secure environment.

upvoted 3 times

🗳️ 👤 **apprehensive_scar** 3 years, 4 months ago

C is the right answer. you need to detect location

upvoted 3 times

🗳️ 👤 **AShahine21** 3 years, 8 months ago

C

You can detect the type of weeds and the location within the field.

upvoted 1 times

🗳️ 👤 **cnethers** 3 years, 8 months ago

If they had an answer with "Faster R-CNN" then it would be different.

This is a good article talking about SSD, Faster R-CNN, R-FCN and others which is a good read.

<https://jonathan-hui.medium.com/object-detection-speed-and-accuracy-comparison-faster-r-cnn-r-fcn-ssd-and-yolo-5425656ae359>

upvoted 2 times

🗳️ 👤 **cnethers** 3 years, 9 months ago

I would go with answer C ..

SSD are new architectures faster than the old CNN

<https://towardsdatascience.com/understanding-ssd-multibox-real-time-object-detection-in-deep-learning-495ef744fab>

upvoted 2 times

🗳️ 👤 **[Removed]** 3 years, 9 months ago

I would select answer A, situation is very similar to this one: <https://aws.amazon.com/blogs/machine-learning/building-a-lawn-monitor-and-weed-detection-solution-with-aws-machine-learning-and-iot-services/>

upvoted 3 times

🗳️ 👤 **crispogioele** 3 years, 8 months ago

I think it's better go with C, since the question also ask for the location of the weed on the field, while the example you posted is just a classifier.

upvoted 1 times

🗳️ 👤 **attaraya** 3 years, 7 months ago

since field is divided in to 10x10 grid I felt A is more suitable.

upvoted 1 times

🗳️ 👤 **cpal012** 2 years, 2 months ago

So you expect precisely one weed per grid (total 100) of an entire field? If a field is a hectare, then each grid would be 100m²

upvoted 1 times

🗳️ 👤 **Bala1212081** 3 years, 8 months ago

The link says clearly only classification and not talking about object detection. Answer should be C

upvoted 1 times

A manufacturer is operating a large number of factories with a complex supply chain relationship where unexpected downtime of a machine can cause production to stop at several factories. A data scientist wants to analyze sensor data from the factories to identify equipment in need of preemptive maintenance and then dispatch a service team to prevent unplanned downtime. The sensor readings from a single machine can include up to 200 data points including temperatures, voltages, vibrations, RPMs, and pressure readings.

To collect this sensor data, the manufacturer deployed Wi-Fi and LANs across the factories. Even though many factory locations do not have reliable or high-speed internet connectivity, the manufacturer would like to maintain near-real-time inference capabilities.


Which deployment architecture for the model will address these business requirements?

- A. Deploy the model in Amazon SageMaker. Run sensor data through this model to predict which machines need maintenance.
- B. Deploy the model on AWS IoT Greengrass in each factory. Run sensor data through this model to infer which machines need maintenance.
- C. Deploy the model to an Amazon SageMaker batch transformation job. Generate inferences in a daily batch report to identify machines that need maintenance.
- D. Deploy the model in Amazon SageMaker and use an IoT rule to write data to an Amazon DynamoDB table. Consume a DynamoDB stream from the table with an AWS Lambda function to invoke the endpoint.

Suggested Answer: A

Community vote distribution

B (100%)




  **[Removed]**  2 years, 9 months ago

I would select B, based on the following AWS examples:

<https://aws.amazon.com/blogs/iot/industrial-iot-from-condition-based-monitoring-to-predictive-quality-to-digitize-your-factory-with-aws-iot-services/>

<https://aws.amazon.com/blogs/iot/using-aws-iot-for-predictive-maintenance/>

upvoted 26 times

  **SophieSu**  2 years, 9 months ago

B is my answer.

For latency-sensitive use cases and for use-cases that require analyzing large amounts of streaming data, it may not be possible to run ML inference in the cloud. Besides, cloud-connectivity may not be available all the time.

For these use cases, you need to deploy the ML model close to the data source.

SageMaker Neo + IoT Greengrass

To design and push something to edge:

1. design something to do the job, say TF model
2. compile it for the edge device using SageMaker Neo, say Nvidia Jetson
3. run it on the edge using IoT Greengrass

upvoted 17 times

  **Mickey321**  10 months, 1 week ago

Selected Answer: B

without relying on internet connectivity.

upvoted 2 times

  **Peeking** 1 year, 6 months ago

Selected Answer: B

The described solution will be solved by an edge solution as internet reliability is low. IoT Greengrass is the best solution for the edge inference.

upvoted 3 times

  **ovokpus** 2 years ago

Selected Answer: B

This is an edge solution, having as little traffic with AWS resources in regions. For this, start thinking IoT Greengrass and Sagemaker Neo, and you'll be halfway there.

Answer is B, no doubt

upvoted 3 times

🗨️ 👤 **apprehensive_scar** 2 years, 4 months ago

B is the answer, obviously

upvoted 1 times

🗨️ 👤 **[Removed]** 2 years, 7 months ago

Selected Answer: B

This solution requires edge capabilities and to be able to run the inference models in near real-time. SageMaker Neo is a deployable unit on the edge architecture (IoT Greengrass) which can host the runtime inference model.

upvoted 4 times

🗨️ 👤 **mahmoudai** 2 years, 8 months ago

A: not a complete solution a lot of details is missed

C: daily batch training is huge defect in this solution

D: writing to dynamoDB and invoking endpoint make this solution slower than using an IoT Green Grass

Answer: B

upvoted 1 times

🗨️ 👤 **Vita_Rasta84444** 2 years, 8 months ago

I would choose B because IoT reduce latency because they work on local machine

upvoted 1 times

🗨️ 👤 **astonm13** 2 years, 9 months ago

I would choose B

upvoted 1 times

A Machine Learning Specialist is designing a scalable data storage solution for Amazon SageMaker. There is an existing TensorFlow-based model implemented as a train.py script that relies on static training data that is currently stored as TFRecords.

Which method of providing training data to Amazon SageMaker would meet the business requirements with the LEAST development overhead?

- A. Use Amazon SageMaker script mode and use train.py unchanged. Point the Amazon SageMaker training invocation to the local path of the data without reformatting the training data.
- B. Use Amazon SageMaker script mode and use train.py unchanged. Put the TFRecord data into an Amazon S3 bucket. Point the Amazon SageMaker training invocation to the S3 bucket without reformatting the training data.
- C. Rewrite the train.py script to add a section that converts TFRecords to protobuf and ingests the protobuf data instead of TFRecords.
- D. Prepare the data in the format accepted by Amazon SageMaker. Use AWS Glue or AWS Lambda to reformat and store the data in an Amazon S3 bucket.

Suggested Answer: D

Community vote distribution

B (100%)

  **[Removed]**  3 years, 8 months ago

I would select B. Based on the following AWS documentation it appears this is the right approach:

https://sagemaker.readthedocs.io/en/stable/frameworks/tensorflow/using_tf.html

<https://github.com/aws-samples/amazon-sagemaker-script-mode/blob/master/tf-horovod-inference-pipeline/train.py>

upvoted 21 times

  **SophieSu**  3 years, 8 months ago

B is my answer.

Reading Data

```
filenames = ["s3://bucketname/path/to/file1.tfrecord",
"s3://bucketname/path/to/file2.tfrecord"]
```

```
dataset = tf.data.TFRecordDataset(filenames)
```


upvoted 12 times

  **MultiCloudIronMan**  9 months, 1 week ago

Selected Answer: B

This approach leverages the existing TFRecord format and minimizes changes to the current setup, ensuring a smooth transition to using Amazon SageMaker with minimal development effort.

upvoted 1 times

  **Mickey321** 1 year, 10 months ago

Selected Answer: B

option B

upvoted 1 times

  **kaike_reis** 1 year, 11 months ago

Selected Answer: B

Letters C and D need code development and are therefore discarded. As we want a scalable data storage, it is recommended to use the Letter B, since S3 is scalable. Letter A is wrong as your personal computer is not scalable.

upvoted 2 times

  **ashton777** 2 years ago



Where had the Capslock Donald gone? I kinda miss his answers

upvoted 8 times

  **himanshu10k** 2 years, 2 months ago

Internet connectivity issue: then how IOT can be a solution? (Correct answer should be A)

upvoted 2 times



  **AjoseO** 2 years, 3 months ago

Selected Answer: B

Amazon SageMaker script mode enables training a machine learning model using a script that you provide. By using the unchanged train.py script and putting the TFRecord data into an Amazon S3 bucket, you can easily point the Amazon SageMaker training invocation to the S3 bucket without reformatting the training data.

This option avoids the need to rewrite the train.py script or to prepare the data in a different format. It also leverages the scalability and cost-effectiveness of Amazon S3 for storing large amounts of data, which is important for training machine learning models.

upvoted 3 times

  **ccpmad** 1 year, 11 months ago



thank you chatgpt

upvoted 1 times

  **apprehensive_scar** 3 years, 4 months ago

B, obviously

upvoted 1 times

  **KM226** 3 years, 5 months ago

Selected Answer: B

I like answer B

upvoted 2 times

  **Zhubajie** 3 years, 7 months ago


Why not A? Why can't we train it from local?

upvoted 1 times

  **AddiWei** 3 years, 4 months ago


Sagemaker to my understanding requires the data to be in S3.

upvoted 5 times

  **Huy** 3 years, 7 months ago

B. <https://aws.amazon.com/about-aws/whats-new/2019/01/amazon-sagemaker-batch-transform-now-supports-tfrecord-format/>

upvoted 2 times

  **cnethers** 3 years, 8 months ago

Unfortunately you can't use the script unchanged, there are some things that need to be added:

1. Make sure your script can handle --model_dir as an additional command line argument. If you did not specify a location when you created the TensorFlow estimator, an S3 location under the default training job bucket is used. Distributed training with parameter servers requires you to use the tf.estimator.train_and_evaluate API and to provide an S3 location as the model directory during training.
2. Load input data from the input channels. The input channels are defined when fit is called.

https://sagemaker.readthedocs.io/en/stable/frameworks/tensorflow/using_tf.html

Beause of the pre-rec Ans A and B are an easy disqualification.

There is no need to change the training format so option C is a red herring

Ans is D


Not the most obvious answer

upvoted 4 times

  **SophieSu** 3 years, 8 months ago

according your explanation, the correct answer should be B

upvoted 2 times

  **akgarg00** 1 year, 7 months ago

It mentions using sagemaker in "script mode" Which is different from working on Sagemaker using python SDK.

upvoted 1 times

The chief editor for a product catalog wants the research and development team to build a machine learning system that can be used to detect whether or not individuals in a collection of images are wearing the company's retail brand. The team has a set of training data. Which machine learning algorithm should the researchers use that BEST meets their requirements?

- A. Latent Dirichlet Allocation (LDA)
- B. Recurrent neural network (RNN)
- C. K-means
- D. Convolutional neural network (CNN)

Suggested Answer: D

Community vote distribution

D (100%)

☐ **SophieSu** Highly Voted 3 years, 8 months ago

D. CNN - image
upvoted 36 times

☐ **Chiquitabandita** Most Recent 1 year ago

yeah, nothing creepy about a company wanting to do this :)
upvoted 2 times

☐ **Mickey321** 1 year, 10 months ago

Selected Answer: D
D - image
upvoted 2 times

☐ **kaike_reis** 1 year, 11 months ago

Selected Answer: D
D is the correct
Well, I did those exams topics questions for 2 weeks and still easier compared to Maarek exams that simulate AWS MLS. Is that correct? Can I expect the exam to be easier as it's in exams topics?
upvoted 3 times

☐ **ccpmad** 1 year, 11 months ago

Selected Answer: D
Convolutional neural networks (CNNs) are specifically designed for image recognition tasks and have been highly successful in detecting patterns and features within images.

CNNs are particularly effective in capturing spatial patterns and visual features from images
upvoted 2 times

☐ **ryuhei** 2 years, 9 months ago

Selected Answer: D
Answer is "D"
upvoted 1 times

A retail company is using Amazon Personalize to provide personalized product recommendations for its customers during a marketing campaign. The company sees a significant increase in sales of recommended items to existing customers immediately after deploying a new solution version, but these sales decrease a short time after deployment. Only historical data from before the marketing campaign is available for training. How should a data scientist adjust the solution?

- A. Use the event tracker in Amazon Personalize to include real-time user interactions.
- B. Add user metadata and use the HRNN-Metadata recipe in Amazon Personalize.
- C. Implement a new solution using the built-in factorization machines (FM) algorithm in Amazon SageMaker.
- D. Add event type and event value fields to the interactions dataset in Amazon Personalize.

Suggested Answer: D

Community vote distribution

A (100%)

🗳️ 👤 **SophieSu** Highly Voted 2 years, 8 months ago

A is the correct answer. Because in this case, it is not the problem with the existing historical data (event value, event type(click or not)), the sales do not keep growing and now you need to obtain more recent interactive data. An event tracker specifies a destination dataset group for new event data.
upvoted 38 times

🗳️ 👤 **AjithkumarSL** 2 years, 8 months ago

I agree.. A is the right choice.. The model need the real time data to adjust to create recommendations..
upvoted 3 times

🗳️ 👤 **ovokpus** Highly Voted 2 years ago

Selected Answer: A

here is the receipt:

<https://docs.aws.amazon.com/personalize/latest/dg/maintaining-relevance.html>

upvoted 6 times

🗳️ 👤 **Mickey321** Most Recent 10 months, 1 week ago

Selected Answer: A

A real time data

upvoted 1 times

🗳️ 👤 **vingidost** 1 year, 3 months ago

A is the right choice.

upvoted 1 times

🗳️ 👤 **AjoseO** 1 year, 4 months ago

Selected Answer: A

A. Use the event tracker in Amazon Personalize to include real-time user interactions.

By using the event tracker in Amazon Personalize, the data scientist can collect real-time user interactions, including clicks, views, and purchases, and use these interactions to update the model and generate more accurate recommendations. This could help address the decrease in sales after deploying a new solution version, as the model can be updated to reflect the latest customer behavior. Additionally, including real-time user interactions can help the model better respond to changes in customer behavior and provide more relevant and personalized recommendations, which can help increase sales.

upvoted 2 times

🗳️ 👤 **apprehensive_scar** 2 years, 4 months ago

Selected Answer: A

easy one. A

upvoted 2 times

🗳️ 👤 **Huy** 2 years, 7 months ago

A. <https://docs.aws.amazon.com/personalize/latest/dg/recording-events.html>

upvoted 3 times

A machine learning (ML) specialist wants to secure calls to the Amazon SageMaker Service API. The specialist has configured Amazon VPC with a VPC interface endpoint for the Amazon SageMaker Service API and is attempting to secure traffic from specific sets of instances and IAM users. The VPC is configured with a single public subnet.

Which combination of steps should the ML specialist take to secure the traffic? (Choose two.)

- A. Add a VPC endpoint policy to allow access to the IAM users.
- B. Modify the users' IAM policy to allow access to Amazon SageMaker Service API calls only.
- C. Modify the security group on the endpoint network interface to restrict access to the instances.
- D. Modify the ACL on the endpoint network interface to restrict access to the instances.
- E. Add a SageMaker Runtime VPC endpoint interface to the VPC.


Suggested Answer: AC

Reference:


<https://aws.amazon.com/blogs/machine-learning/private-package-installation-in-amazon-sagemaker-running-in-internet-free-mode/>

Community vote distribution

AC (100%)

 **mona_mansour** Highly Voted 2 years, 8 months ago

A&C...><https://aws.amazon.com/blogs/machine-learning/securing-all-amazon-sagemaker-api-calls-with-aws-privatelink/>
upvoted 15 times

 **wisoxe8356** Highly Voted 1 year, 6 months ago

Selected Answer: AC

A - VPC endpoint policy can limit the access to specific group of user/roles
Not B - setting iam user policy can limit user access other aws service but not secure the traffic
C - "specific" sets of instances - means security rules in instance level
Not D - ACL (access control list) allows or denies specific inbound or outbound traffic at the subnet level.
Not E - VPC is configured with public subnet, adding interface without limit the traffic means not secure
upvoted 7 times

 **loict** Most Recent 9 months, 2 weeks ago

Selected Answer: AC

A. YES - for users
B. NO - the users should access more than just SageMaker
C. YES - for instances
D. NO - ACL are not supported for SageMaker endpoint (only S3, RDS, EKS, etc.)
E. NO - endpoint is already there
upvoted 1 times

 **ccpmad** 11 months ago

Selected Answer: AC

A. Add a VPC endpoint policy to allow access to the IAM users: This will specify the permissions for the IAM users to access the Amazon SageMaker Service API through the VPC endpoint.

C. Modify the security group on the endpoint network interface to restrict access to the instances: By configuring the security group, the specialist can control which instances are allowed to communicate with the SageMaker Service API through the VPC endpoint.
upvoted 1 times

 **venimus_vidimus_vicimus** 1 year, 7 months ago

Should be A & D n0?

We want to configure the endpoint - first to allow IAM users, second to control access to instances.

Since Security Groups are attached to instances (not VPCs) and only allow allow rules - it should be D.

upvoted 3 times

 **exam_prep** 2 years, 1 month ago

Yes, A & D are correct.


A> This will limit access to only names IAM users. It is like defining all for given principals as below:

```
{
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": "*",
      "Action": "*",
      "Resource": "*"
    }
  ]
}
```

D-> To restrict access to certain instances or IP address you define deny rule at NACL level. Here VPC Interface endpoint is in subnet (the only subnet in VPC). So modify NACL configurations at this subnet level.

Security group are only for allowing the traffic not for deny so C is incorrect.

upvoted 1 times

  **yj123** 2 years, 4 months ago

security group cannot restrict access explicitly, C?

upvoted 1 times

  **yj123** 2 years, 4 months ago

i mean A, D

upvoted 1 times

  **[Removed]** 2 years, 7 months ago

Selected Answer: AC

Security Group controls instance level access. The question requires instance level access.

The VPC endpoint is already set up. It needs a policy attachment for particular IAM Users. I would have preferred this to be IAM Roles instead of Users, as a more appropriate question. Nevertheless, answer is A & C.

upvoted 2 times

  **Madwyn** 2 years, 8 months ago

A say allow access TO the IAM users? That's wired, why to the IAM users? How do you access them?

upvoted 1 times

  **msamory** 2 years, 8 months ago

The VPC endpoint is already available waiting to be configured. No need to add one. A and E are out.

Furthermore if an IAM endpoint is not set, a default one will be provided and you can't have more than 1 IAM policy but can modify the one that's available.

-Restrict access to only calls coming from the VPC, then modify the security group to give access to user group or roles that need access to that notebook.

I think the answer is B and C

upvoted 4 times

  **Madwyn** 2 years, 8 months ago

A says add a VPC endpoint policy, not add an endpoint.

upvoted 2 times

  **cnethers** 2 years, 8 months ago

<https://docs.aws.amazon.com/vpc/latest/userguide/vpc-endpoints-access.html>

<https://docs.aws.amazon.com/sagemaker/latest/dg/notebook-interface-endpoint.html#nbi-private-link-policy>

<https://docs.aws.amazon.com/vpc/latest/userguide/integrated-services-vpce-list.html>

upvoted 2 times

An e commerce company wants to launch a new cloud-based product recommendation feature for its web application. Due to data localization regulations, any sensitive data must not leave its on-premises data center, and the product recommendation model must be trained and tested using nonsensitive data only. Data transfer to the cloud must use IPsec. The web application is hosted on premises with a PostgreSQL database that contains all the data. The company wants the data to be uploaded securely to Amazon S3 each day for model retraining. How should a machine learning specialist meet these requirements?

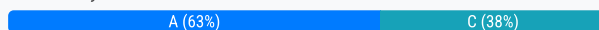
- A. Create an AWS Glue job to connect to the PostgreSQL DB instance. Ingest tables without sensitive data through an AWS Site-to-Site VPN connection directly into Amazon S3.
- B. Create an AWS Glue job to connect to the PostgreSQL DB instance. Ingest all data through an AWS Site-to-Site VPN connection into Amazon S3 while removing sensitive data using a PySpark job.
- C. Use AWS Database Migration Service (AWS DMS) with table mapping to select PostgreSQL tables with no sensitive data through an SSL connection. Replicate data directly into Amazon S3.
- D. Use PostgreSQL logical replication to replicate all data to PostgreSQL in Amazon EC2 through AWS Direct Connect with a VPN connection. Use AWS Glue to move data from Amazon EC2 to Amazon S3.

Suggested Answer: C

Reference:

https://docs.aws.amazon.com/dms/latest/userguide/CHAP_Source.PostgreSQL.html

Community vote distribution



cnethers Highly Voted 3 years, 8 months ago

ASK : Extract Data over IPsec

So we need an ETL + Site to site VPN

GLUE is an ETL service but can it connect to PostgreSQL? yes

<https://docs.aws.amazon.com/glue/latest/dg/aws-glue-programming-etl-connect.html#aws-glue-programming-etl-connect-jdbc>

How to connect Glue to an on-site DB

<https://aws.amazon.com/blogs/big-data/how-to-access-and-analyze-on-premises-data-stores-using-aws-glue/>

My Answer would be A

Answer C only makes a 443 (SSL) connection so does not meet the IPsec requirement

upvoted 27 times

ksrivastavaSumit Highly Voted 3 years, 9 months ago

A? IPsec needs to be covered as well

upvoted 8 times

StelSen 3 years, 8 months ago

Yes. <https://aws.amazon.com/blogs/big-data/how-to-access-and-analyze-on-premises-data-stores-using-aws-glue/>. 'A' is the correct answer.

upvoted 3 times

MultiCloudIronMan Most Recent 8 months, 2 weeks ago

Selected Answer: A

It's 'A' because IPsec is required.

upvoted 1 times

sachin80 1 year, 1 month ago

A: <https://medium.com/awsblackbelt/loading-on-prem-postgres-data-into-amazon-s3-with-server-side-filtering-c13bcee8b769>

upvoted 1 times

VR10 1 year, 3 months ago

Selected Answer: A

B - Doesn't take care of only nonsensitive data being allowed to leave the on-premise.

C - Uses SSL and not IPsec.

D - like B transfers all data.

Hence the correct answer is A

upvoted 1 times

🗨️ **AIWave** 1 year, 4 months ago

I will go with B

Site to Site VPN -> IPsec requirement

AWS Glue -> connect and catalog PostgreSQL

Pyspark -> remove sensitive information. AWS glue supports pyspark

upvoted 1 times

🗨️ **kyuhuck** 1 year, 4 months ago

Selected Answer: C

The best option is to use AWS Database Migration Service (AWS DMS) with table mapping to select PostgreSQL tables with no sensitive data through an SSL connection. Replicate data directly into Amazon S3. This option meets the following requirements: It ensures that only nonsensitive data is transferred to the cloud by using table mapping to filter out the tables that contain sensitive data¹. It uses IPsec to secure the data transfer by enabling SSL encryption for the AWS DMS endpoint². It uploads the data to Amazon S3 each day for model retraining by using the ongoing replication feature of AWS DMS³

upvoted 3 times

🗨️ **LeoD** 6 months, 1 week ago

IPsec and SSL are two different things. Using SSL does not necessarily mean option C has IPsec implemented, which is required.

upvoted 1 times

🗨️ **Reju** 1 year, 9 months ago

but glue can not filter out the data during the ingestion and hence option A wouldn't be the right one! I would go for B

upvoted 1 times

🗨️ **LeoD** 6 months, 1 week ago

I think A is saying only to ingest tables that don't contain sensitive data, meaning while configuring Glue, the specialist will only select the tables that don't contain sensitive data for ingestion.

upvoted 1 times

🗨️ **jopaca1216** 1 year, 9 months ago

B

Both A and C are not correct... due to the question is not talking about tables with no sensitive data... and that DMS typically act on the data on AWS side, the right answer is B

AWS Glue connects to the PostgreSQL database, allowing the removal of sensitive data using a PySpark job BEFORE securely ingesting the data into Amazon S3, thus aligning with the requirements.

upvoted 3 times

🗨️ **Hybrid_Cloud_boy** 1 year, 6 months ago

I think the issue with this answer would be that the data actually leaves the DC and enters the glue service before sensitive data is redacted. - Which makes me lean A

upvoted 3 times

🗨️ **Mickey321** 1 year, 10 months ago

Selected Answer: C

Option c

upvoted 1 times

🗨️ **ADVIT** 1 year, 12 months ago

A:

<https://aws.amazon.com/blogs/big-data/doing-data-preparation-using-on-premises-postgresql-databases-with-aws-glue-databrew/>

upvoted 2 times

🗨️ **Ajose0** 2 years, 4 months ago

Selected Answer: A

A. Create an AWS Glue job to connect to the PostgreSQL DB instance. Ingest tables without sensitive data through an AWS Site-to-Site VPN connection directly into Amazon S3.

This solution meets the requirements of data localization regulations and secure data transfer. By creating an AWS Glue job to connect to the PostgreSQL DB instance, the machine learning specialist can extract tables without sensitive data. By using a Site-to-Site VPN connection, the data

can be securely transferred from the on-premises data center to Amazon S3, where it can be used for model retraining. This solution ensures that any sensitive data remains in the on-premises data center, and that only non-sensitive data is uploaded to the cloud.

upvoted 2 times

🗲️ 👤 **matteocal** 2 years, 11 months ago

Selected Answer: A

IPSec means VPN

upvoted 4 times

🗲️ 👤 **geekgirl007** 3 years, 5 months ago

Answer is A. IPsec is not the same as SSL. Site to site VPN is for IPsec: <https://aws.amazon.com/vpn/site-to-site-vpn/>

Also Glue can directly connect to Postgres and upload to S3: <https://aws.amazon.com/blogs/big-data/how-to-access-and-analyze-on-premises-data-stores-using-aws-glue/>

upvoted 3 times

🗲️ 👤 **Deepsachin** 3 years, 7 months ago

A is the answer

upvoted 1 times

🗲️ 👤 **Dr_Kiko** 3 years, 8 months ago

Between A and C, I pick A because IPSec requires VPN; otherwise DMS is a better option

upvoted 3 times

🗲️ 👤 **Dr_Kiko** 3 years, 8 months ago

Between A and C, I pick A because IPSec requires VPN; otherwise DMS is a better option

upvoted 1 times

A logistics company needs a forecast model to predict next month's inventory requirements for a single item in 10 warehouses. A machine learning specialist uses

Amazon Forecast to develop a forecast model from 3 years of monthly data. There is no missing data. The specialist selects the DeepAR+ algorithm to train a predictor. The predictor means absolute percentage error (MAPE) is much larger than the MAPE produced by the current human forecasters.

Which changes to the CreatePredictor API call could improve the MAPE? (Choose two.)

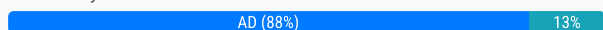
- A. Set PerformAutoML to true.
- B. Set ForecastHorizon to 4.
- C. Set ForecastFrequency to W for weekly.
- D. Set PerformHPO to true.
- E. Set FeaturizationMethodName to filling.

Suggested Answer: CD

Reference:

<https://docs.aws.amazon.com/forecast/latest/dg/forecast.dg.pdf>

Community vote distribution



scuzzy2010 Highly Voted 2 years, 8 months ago

I would choose A and D, however both of them is not possible at the same time. The question is ambiguous, it could mean which two options, but not necessarily both.

A - If you want Amazon Forecast to evaluate each algorithm and choose the one that minimizes the objective function, set PerformAutoML to true.

D - The following algorithms support HPO: -> DeepAR+.

upvoted 17 times

Oscaaar 2 years, 8 months ago

If custom forecast types are specified, Forecast evaluates metrics at those specified forecast types, and takes the averages of those metrics to determine the optimal outcomes during HPO and AutoML.

For both AutoML and HPO, Forecast chooses the option that minimizes the average losses over the forecast types. During HPO, Forecast uses the first backtest window to find the optimal hyperparameter values. During AutoML, Forecast uses the averages across all backtest windows and the optimal hyperparameters values from HPO to find the optimal algorithm. <https://docs.aws.amazon.com/forecast/latest/dg/metrics.html>

upvoted 3 times

Vita_Rasta84444 Highly Voted 2 years, 8 months ago

It is A and D, there are no weekly data, they have only monthly data and can not switch horizon to 4

upvoted 7 times

loict Most Recent 9 months, 2 weeks ago

Selected Answer: AC

A. YES - DeepAR+ most likely to be chosen, but worth a try

B. NO - increasing the forecast horizon is not likely to improve the 3 months we want

C. NO - we want monthly, not weekly

D. YES

E. NO - there are no missing values

upvoted 1 times

Mickey321 10 months, 1 week ago

Selected Answer: AD

The changes to the CreatePredictor API call that could improve the MAPE are option A and option D. By setting PerformAutoML to true, you can enable Amazon Forecast to automatically explore different algorithms and choose the best one for your data and business problem. By setting PerformHPO to true, you can enable Amazon Forecast to perform hyperparameter optimization (HPO) and tune the algorithm parameters to improve the accuracy of the predictor. These options can help you find the optimal configuration for your forecast model without manually specifying the algorithm or the hyperparameters.

upvoted 2 times

🗳️ 👤 **Ajose0** 1 year, 4 months ago

Selected Answer: AD

- A. Set PerformAutoML to true.
- D. Set PerformHPO to true.

Setting PerformAutoML to true will enable Amazon Forecast to automatically select the best algorithm and hyperparameters for your data and problem. This can help improve the MAPE by finding the optimal combination of algorithm and hyperparameters that minimize prediction error.

Setting PerformHPO to true will enable Amazon Forecast to perform a hyperparameter optimization search to find the best combination of hyperparameters that result in the best prediction performance. This can help improve the MAPE by finding the optimal combination of hyperparameters that minimize prediction error.

upvoted 4 times

🗳️ 👤 **yemaurocio** 1 year, 6 months ago

Selected Answer: AD

- A. Looking for better algorithms performance
- D. Hyperparameters optimization

upvoted 1 times

🗳️ 👤 **Shailendraa** 1 year, 9 months ago

12-sep exam

upvoted 3 times

🗳️ 👤 **vanluigi** 2 years, 1 month ago

Why are not B and C? The question asks about modifications that increase MAPE (thats bad):

- B - If FH is larger, error will increase
- C - Data is based on months, change that will make erros on forecasting values
- E - There is no data gap so is useless
- A - Selec best between all should DECREASE MAPE
- D - Tunning hyperparms will DECREASE MAPE

upvoted 5 times

🗳️ 👤 **mona_mansour** 2 years, 8 months ago

A&D...>By default, Amazon Forecast uses the 0.1 (P10), 0.5 (P50), and 0.9 (P90) quantiles for hyperparameter tuning during hyperparameter optimization (HPO) and for model selection during AutoML. If you specify custom forecast types when creating a predictor, Forecast uses those forecast types during HPO and AutoML.

If custom forecast types are specified, Forecast evaluates metrics at those specified forecast types, and takes the averages of those metrics to determine the optimal outcomes during HPO and AutoML.

For both AutoML and HPO, Forecast chooses the option that minimizes the average losses over the forecast types. During HPO, Forecast uses the first backtest window to find the optimal hyperparameter values. During AutoML, Forecast uses the averages across all backtest windows and the optimal hyperparameters values from HPO to find the optimal algorithm.

upvoted 4 times

🗳️ 👤 **SophieSu** 2 years, 8 months ago

- C. ForecastFrequency
- M- MONTHLY
- W- WEEKLY

- D. PerformHPO

Whether to perform hyperparameter optimization (HPO). HPO finds optimal hyperparameter values for your training data. The process of performing HPO is known as running a hyperparameter tuning job.

The default value is false. In this case, Amazon Forecast uses default hyperparameter values from the chosen algorithm.

- E. FeaturizationMethodName

The name of the method. The "filling" method is the only supported method.

upvoted 2 times

🗳️ 👤 **seanLu** 2 years, 8 months ago

But for option C, according to the Developer Guide, The forecast frequency must be greater than or equal to the TARGET_TIME_SERIES dataset frequency. and the training data is monthly data, so ForecastFrequency can not be less than Monthly.

upvoted 6 times

  **SophieSu** 2 years, 9 months ago

ABE can be excluded. CD is my answer.

A. PerformAutoML

If you want Amazon Forecast to evaluate each algorithm and choose the one that minimizes the objective function, set PerformAutoML to true. The objective function is defined as the mean of the weighted losses over the forecast types. By default, these are the p10, p50, and p90 quantile losses.

When AutoML is enabled, the following properties are disallowed:

AlgorithmArn

HPOConfig

PerformHPO

TrainingParameters

B. ForecastHorizon

Specifies the number of time-steps that the model is trained to predict. The forecast horizon is also called the prediction length.

For example, if you configure a dataset for daily data collection (using the DataFrequency parameter of the CreateDataset operation) and set the forecast horizon to 10, the model returns predictions for 10 days.

The maximum forecast horizon is the lesser of 500 time-steps or 1/3 of the TARGET_TIME_SERIES dataset length.

upvoted 3 times

A data scientist wants to use Amazon Forecast to build a forecasting model for inventory demand for a retail company. The company has provided a dataset of historic inventory demand for its products as a .csv file stored in an Amazon S3 bucket. The table below shows a sample of the dataset.

timestamp	item_id	demand	category	lead_time
2019-12-14	uni_000736	120	hardware	90
2020-01-31	uni_003429	98	hardware	30
2020-03-04	uni_000211	234	accessories	10

How should the data scientist transform the data?

- A. Use ETL jobs in AWS Glue to separate the dataset into a target time series dataset and an item metadata dataset. Upload both datasets as .csv files to Amazon S3.
- B. Use a Jupyter notebook in Amazon SageMaker to separate the dataset into a related time series dataset and an item metadata dataset. Upload both datasets as tables in Amazon Aurora.
- C. Use AWS Batch jobs to separate the dataset into a target time series dataset, a related time series dataset, and an item metadata dataset. Upload them directly to Forecast from a local machine.
- D. Use a Jupyter notebook in Amazon SageMaker to transform the data into the optimized protobuf recordIO format. Upload the dataset in this format to Amazon S3.

Suggested Answer: B

Community vote distribution

A (100%)

[Removed] Highly Voted 3 years, 3 months ago

I would answer A. Target and metadata must be in two files and loaded from S3, based on documentation:

<https://docs.aws.amazon.com/forecast/latest/dg/dataset-import-guidelines-troubleshooting.html>

upvoted 26 times

ZSun 1 year, 8 months ago

1. I cannot find any evidence support the seperate file defination.
2. A,B,C all seperate datasets, this explanation is weak.

upvoted 1 times

Ajose0 Highly Voted 1 year, 9 months ago

Selected Answer: A

Amazon Forecast requires the input data to be separated into a target time series dataset and an item metadata dataset.

The target time series dataset should include the time series data that you want to use for forecasting, such as inventory demand in this case. The item metadata dataset should include the metadata that describes the items in the time series, such as product IDs, categories, and attributes.

Therefore, the data scientist should use ETL jobs in AWS Glue to separate the dataset into a target time series dataset and an item metadata dataset. Both datasets should be uploaded as .csv files to Amazon S3, which is a suitable storage option for input data to Amazon Forecast.

upvoted 9 times

ccpmad 1 year, 5 months ago

thank you chatgpt

upvoted 2 times

AIWave Most Recent 10 months, 2 weeks ago

I would go with A

Input formats for forecast -> Json, CSV and paraquet (Selects A & eliminates B, C, D)

Data needs to be split in target time series dataset and an item metadata dataset

upvoted 1 times

Mickey321 1 year, 4 months ago

Selected Answer: A

Target and metadata must be in two files

upvoted 1 times

🗨️ 👤 **kaike_reis** 1 year, 4 months ago

Selected Answer: A

Letter A is correct, as it uses a specific transformation service (AWS Glue) and saves it in a cloud database for AWS Forecast to access. By default in ML, our storage option will be AWS S3 (unless caveats or issue specifications). That said, we discard B and C. Letter D is discarded due to the format requested by AWS Forecast being csv.

upvoted 1 times

🗨️ 👤 **ystotest** 2 years, 1 month ago

Selected Answer: A

I would vote for A

upvoted 3 times

🗨️ 👤 **tgaos** 2 years, 7 months ago

The answer is A.

According to the <https://docs.aws.amazon.com/forecast/latest/dg/forecast.dg.pdf> , page 51.

Target Time Series Dataset:

Required: timestamp, item_id, demand

Additional: lead_time

Item Metadata Dataset:

item_id, category

upvoted 3 times

🗨️ 👤 **tgaos** 2 years, 6 months ago

You can find the same question with the picture at <https://ccnav7.net/a-data-scientist-wants-to-use-amazon-forecast-to-build-a-forecasting-model-for-inventory-demand-for-a-retail-company/>

upvoted 1 times

🗨️ 👤 **DSJingguo** 3 years, 2 months ago

The correct answer is A

"Forecast supports only the comma-separated values (CSV) file format. You can't separate values using tabs, spaces, colons, or any other characters.

Guideline: Convert your dataset to CSV format (using only commas as your delimiter) and try importing the file again."

upvoted 1 times

🗨️ 👤 **achiko** 3 years, 2 months ago

lead time belongs to related time series, as its not a target variable

upvoted 1 times

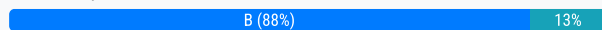
A machine learning specialist is running an Amazon SageMaker endpoint using the built-in object detection algorithm on a P3 instance for real-time predictions in a company's production application. When evaluating the model's resource utilization, the specialist notices that the model is using only a fraction of the GPU.

Which architecture changes would ensure that provisioned resources are being utilized effectively?

- A. Redeploy the model as a batch transform job on an M5 instance.
- B. Redeploy the model on an M5 instance. Attach Amazon Elastic Inference to the instance.
- C. Redeploy the model on a P3dn instance.
- D. Deploy the model onto an Amazon Elastic Container Service (Amazon ECS) cluster using a P3 instance.

Suggested Answer: D

Community vote distribution



[Removed] 3 years, 9 months ago

B is correct. Redeploy with CPU and add elastic inference to reduce costs. See: <https://aws.amazon.com/machine-learning/elastic-inference/>
upvoted 25 times

Togy 3 months ago

Selected Answer: D

The Amazon EC2 M5 instance family is designed for general-purpose workloads, and they are CPU-optimized. Therefore, M5 instances do not come with GPUs. The only option that talks about optimising the use of already provisioned resources is option D, So that must be the answer.
upvoted 2 times

MultiCloudIronMan 9 months, 1 week ago

Selected Answer: B

This solution allows you to use a more cost-effective instance type while leveraging Elastic Inference to provide the necessary GPU acceleration
upvoted 1 times

GS_77 9 months, 4 weeks ago

Selected Answer: C

redeploying the model on a P3dn instance is the best approach to ensure the provisioned GPU resources are being utilized effectively.
upvoted 2 times

AIWave 1 year, 4 months ago

My vote is B

Elastic inference

- provides cheaper acceleration than full GPU

- works with M class machines

- Works with Tensorflow, MXNet, pytorch, image classification and object detection algorithms

upvoted 1 times

sukye 1 year, 7 months ago

Elastic Inference has been deprecated since Apr 2023.

upvoted 3 times

Mickey321 1 year, 10 months ago

Selected Answer: B

can reduce the cost and improve the resource utilization of your model, as Amazon Elastic Inference allows you to attach low-cost GPU-powered acceleration to Amazon EC2 and Amazon SageMaker instances to run inference workloads with a fraction of the compute resources. You can also choose the right amount of inference acceleration that suits your needs, and scale it up or down as needed.

upvoted 2 times

Ajose0 2 years, 4 months ago

Selected Answer: B

Amazon Elastic Inference allows you to attach low-cost GPU-powered acceleration to EC2 and SageMaker instances, to reduce the cost of running deep learning inference.

You can choose any CPU instance that is best suited to the overall compute and memory needs of your application, and then separately configure the right amount of GPU-powered inference acceleration. This would allow you to efficiently utilize resources and reduce costs.

upvoted 3 times

🗲️ 👤 **Peeking** 2 years, 6 months ago

Selected Answer: B

Elastic inference enables GPU only when load increases. With 50% utilisation there is no need to deploy P3 as the base inference machine.

upvoted 1 times

🗲️ 👤 **ystotest** 2 years, 7 months ago

Selected Answer: B

Agreed with B

upvoted 1 times

🗲️ 👤 **Shailendraa** 2 years, 9 months ago

12-sep exam

upvoted 2 times

🗲️ 👤 **SriAkula** 3 years, 3 months ago

Answer: B

Explanation: <https://aws.amazon.com/machine-learning/elastic-inference/>

upvoted 2 times

🗲️ 👤 **mahmoudai** 3 years, 8 months ago

B: production mostly needs CPU with EI rather than GPU machines

upvoted 1 times

🗲️ 👤 **mona_mansour** 3 years, 8 months ago

B..>Amazon Elastic Inference (EI) is a resource you can attach to your Amazon EC2 CPU instances to accelerate your deep learning (DL) inference workloads. Amazon EI accelerators come in multiple sizes and are a cost-effective method to build intelligent capabilities into applications running on Amazon EC2 instances.

upvoted 3 times

🗲️ 👤 **Vita_Rasta84444** 3 years, 8 months ago

B is correct

upvoted 1 times

A data scientist uses an Amazon SageMaker notebook instance to conduct data exploration and analysis. This requires certain Python packages that are not natively available on Amazon SageMaker to be installed on the notebook instance.

How can a machine learning specialist ensure that required packages are automatically available on the notebook instance for the data scientist to use?

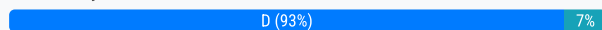
- A. Install AWS Systems Manager Agent on the underlying Amazon EC2 instance and use Systems Manager Automation to execute the package installation commands.
- B. Create a Jupyter notebook file (.ipynb) with cells containing the package installation commands to execute and place the file under the /etc/init directory of each Amazon SageMaker notebook instance.
- C. Use the conda package manager from within the Jupyter notebook console to apply the necessary conda packages to the default kernel of the notebook.
- D. Create an Amazon SageMaker lifecycle configuration with package installation commands and assign the lifecycle configuration to the notebook instance.

Suggested Answer: B

Reference:

<https://towardsdatascience.com/automating-aws-sagemaker-notebooks-2dec62bc2c84>

Community vote distribution



🗳️ **[Removed]** **Highly Voted** 👍 2 years, 9 months ago

I would select D. See AWS documentation: <https://docs.aws.amazon.com/sagemaker/latest/dg/nbi-add-external.html>

upvoted 20 times

🗳️ **u_b** **Most Recent** 🕒 7 months, 2 weeks ago

by excluding wrong options:

A you might not have access to the EC2 instance => out

B no automation => out

C only the default kernel, which limits the DS => out

=> D

upvoted 2 times

🗳️ **Mickey321** 10 months, 1 week ago

Selected Answer: D

option D

upvoted 1 times

🗳️ **ADVIT** 12 months ago

D.

<https://docs.aws.amazon.com/sagemaker/latest/dg/nbi-add-external.html>

upvoted 1 times

🗳️ **CKS1210** 1 year ago

Selected Answer: D

Install custom environments and kernels on the notebook instance's Amazon EBS volume. This ensures that they persist when you stop and restart the notebook instance, and that any external libraries you install are not updated by SageMaker. To do that, use a lifecycle configuration that includes both a script that runs when you create the notebook instance (on-create) and a script that runs each time you restart the notebook instance (on-start).

upvoted 1 times

🗳️ **31Rishab** 1 year, 9 months ago

Selected Answer: D

Even the link given suggest Option D

upvoted 3 times

🗨️ 👤 **ovokpus** 2 years ago

Selected Answer: D

Please ignore my previous comment, the answer is D
upvoted 2 times

🗨️ 👤 **ovokpus** 2 years ago

Selected Answer: B

Key word here is how can the developer "guarantee"??
He guarantees that by including the install commands as part of the notebook.

So, against the grain, I stand with B
upvoted 1 times

🗨️ 👤 **ovokpus** 2 years ago

Scratch that. The Answer is D
upvoted 1 times

🗨️ 👤 **ayatkhrisat** 2 years, 1 month ago

Selected Answer: D

D should be the answer
upvoted 3 times

🗨️ 👤 **vetaal** 2 years, 5 months ago

Selected Answer: D

<https://docs.aws.amazon.com/sagemaker/latest/dg/nbi-add-external.html>
upvoted 1 times
upvoted 2 times

🗨️ 👤 **geekgirl007** 2 years, 5 months ago

Selected Answer: D

D "automatically" is the key here and using lifecycle configuration <https://docs.aws.amazon.com/sagemaker/latest/dg/nbi-add-external.html>
upvoted 2 times

🗨️ 👤 **rafaelo** 2 years, 6 months ago

It is D, although is not even the best answer in my opinion. Although by default conda packages are installed in ephemeral storage, you can change that default behaviour. I did that in my last project and we created our own conda environment that persisted between shutdowns.
upvoted 1 times

🗨️ 👤 **abdohanfi** 2 years, 8 months ago

based on the reference given under the answer its D not B
upvoted 1 times

🗨️ 👤 **mona_mansour** 2 years, 8 months ago

You can install packages using the following methods:

1-Lifecycle configuration scripts

2-Notebooks – The following commands are supported.

%conda install

%pip install

3-The Jupyter terminal – You can install packages using pip and conda directly.
upvoted 2 times

🗨️ 👤 **mona_mansour** 2 years, 8 months ago

NOT B ...>/etc/init contains configuration files used by Upstart.
ANS...>D
upvoted 1 times

🗨️ 👤 **astonm13** 2 years, 8 months ago

Its for sure D
upvoted 1 times

🗨️ 👤 **ksrivastavaSumit** 2 years, 8 months ago

D <https://docs.aws.amazon.com/sagemaker/latest/dg/nbi-add-external.html>
upvoted 3 times

A data scientist needs to identify fraudulent user accounts for a company's ecommerce platform. The company wants the ability to determine if a newly created account is associated with a previously known fraudulent user. The data scientist is using AWS Glue to cleanse the company's application logs during ingestion.

Which strategy will allow the data scientist to identify fraudulent accounts?

- A. Execute the built-in FindDuplicates Amazon Athena query.
- B. Create a FindMatches machine learning transform in AWS Glue.
- C. Create an AWS Glue crawler to infer duplicate accounts in the source data.
- D. Search for duplicate accounts in the AWS Glue Data Catalog.

Suggested Answer: B

Reference:

<https://docs.aws.amazon.com/glue/latest/dg/machine-learning.html>

Community vote distribution

B (100%)

🗳️ **mona_mansour** **Highly Voted** 👍 2 years, 8 months ago

B ,You can use the FindMatches transform to find duplicate records in the source data. A labeling file is generated or provided to help teach the transform.

upvoted 13 times

🗳️ **[Removed]** **Highly Voted** 👍 2 years, 9 months ago

B it is. Reasonable explanation.

upvoted 6 times

🗳️ **Mickey321** **Most Recent** 🕒 10 months, 1 week ago

Selected Answer: B

option B Find matches

upvoted 1 times

🗳️ **Valcilio** 1 year, 3 months ago

Selected Answer: B

It's B, how it's using Glue to clean the data the easiest way will be use Glue's ML FindMatches extension to do this too.

upvoted 1 times

🗳️ **Tomatoteacher** 1 year, 5 months ago

Selected Answer: B

It is B.

upvoted 1 times

🗳️ **omar_bahrain** 2 years, 9 months ago

Agree. Please refer to:

<https://aws.amazon.com/blogs/big-data/integrate-and-deduplicate-datasets-using-aws-lake-formation-findmatches/>

upvoted 2 times

A Data Scientist is developing a machine learning model to classify whether a financial transaction is fraudulent. The labeled data available for training consists of 100,000 non-fraudulent observations and 1,000 fraudulent observations.

The Data Scientist applies the XGBoost algorithm to the data, resulting in the following confusion matrix when the trained model is applied to a previously unseen validation dataset. The accuracy of the model is 99.1%, but the Data Scientist needs to reduce the number of false negatives.

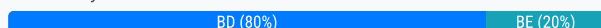
	Predicted 0	Predicted 1
Actual 0	99,966	34
Actual 1	877	123

Which combination of steps should the Data Scientist take to reduce the number of false negative predictions by the model? (Choose two.)

- A. Change the XGBoost eval_metric parameter to optimize based on Root Mean Square Error (RMSE).
- B. Increase the XGBoost scale_pos_weight parameter to adjust the balance of positive and negative weights.
- C. Increase the XGBoost max_depth parameter because the model is currently underfitting the data.
- D. Change the XGBoost eval_metric parameter to optimize based on Area Under the ROC Curve (AUC).
- E. Decrease the XGBoost max_depth parameter because the model is currently overfitting the data.

Suggested Answer: DE

Community vote distribution



LydiaGom Highly Voted 3 years, 1 month ago

B and D

upvoted 12 times

ovokpus Highly Voted 3 years ago

Selected Answer: BD

Compensate for imbalance and optimize on AUC. This is a class imbalance problem, not an overfitting problem.

upvoted 8 times

rb39 2 years, 9 months ago

totally right, overfitting has nothing to do so there is no need to reduce tree depth

upvoted 1 times

MJSY Most Recent 9 months ago

Selected Answer: BD

the question didnt show the model performance on training data, so the overfitting issues is not correct.

upvoted 1 times

loict 1 year, 9 months ago

Selected Answer: BD

- A. NO - that will not address FN specifically but also FP
- B. YES - changing weight is best practice for class imbalance
- C. NO - there is no underfitting at 99.1% accuracy
- D. YES - AUC will address recall, which takes into account FN rate
- E. NO - there is no overfitting at 99.1% accuracy

upvoted 3 times

Mickey321 1 year, 10 months ago

Selected Answer: BD

Step B: Increasing the XGBoost scale_pos_weight parameter to adjust the balance of positive and negative weights can help the model deal with the imbalanced dataset. According to the XGBoost documentation, this parameter controls the balance of positive and negative weights, and is useful for unbalanced classes. A typical value to consider is $\text{sum}(\text{negative instances}) / \text{sum}(\text{positive instances})$. In this case, since there are 100 times more non-fraudulent transactions than fraudulent ones, setting scale_pos_weight to 100 can make the model more sensitive to the minority class and reduce false negatives.

Step D: Changing the XGBoost `eval_metric` parameter to optimize based on Area Under the ROC Curve (AUC) can help the model focus on improving the true positive rate and the true negative rate, which are both important for fraud detection. According to the XGBoost

upvoted 2 times

🗳️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: BD

Step B: Increasing the XGBoost `scale_pos_weight` parameter to adjust the balance of positive and negative weights can help the model deal with the imbalanced dataset. According to the XGBoost documentation, this parameter controls the balance of positive and negative weights, and is useful for unbalanced classes. A typical value to consider is $\text{sum}(\text{negative instances}) / \text{sum}(\text{positive instances})$. In this case, since there are 100 times more non-fraudulent transactions than fraudulent ones, setting `scale_pos_weight` to 100 can make the model more sensitive to the minority class and reduce false negatives.

Step D: Changing the XGBoost `eval_metric` parameter to optimize based on Area Under the ROC Curve (AUC) can help the model focus on improving the true positive rate and the true negative rate, which are both important for fraud detection.

upvoted 1 times

🗳️ 👤 **MIlb** 2 years, 2 months ago

Selected Answer: BE

I have some doubts about D and E.

Precision-Recall AUC is better than AUC curve in imbalanced classes. Then, I choose E

upvoted 2 times

🗳️ 👤 **Ajose0** 2 years, 3 months ago

Selected Answer: BD

Option A and Option E are unlikely to help reduce false negatives.

Option C, increasing `max_depth`, may lead to overfitting, which could make the model worse.

Option D, changing the `eval_metric` to optimize based on AUC, could help improve the model's ability to discriminate between the two classes.

Option B, increasing the `scale_pos_weight` parameter to adjust the balance of positive and negative weights, can help the model better handle imbalanced datasets, which is the case here. By increasing the weight of positive examples, the model will learn to prioritize correctly classifying them, which should reduce the number of false negatives.

upvoted 1 times

🗳️ 👤 **Tomatoteacher** 2 years, 5 months ago

Selected Answer: BD

BD, I have done this before, but it would be better to use Average Precision(AP) instead of AUC, but it is better than other answers.

upvoted 1 times

🗳️ 👤 **Shailendraa** 2 years, 9 months ago

12-sep exam

upvoted 1 times

🗳️ 👤 **ovokpus** 3 years ago

Selected Answer: BE

Compensate for imbalance and overwriting.

upvoted 1 times

🗳️ 👤 **NeverMinda** 3 years ago

Selected Answer: BE

B and E

upvoted 1 times

🗳️ 👤 **MLGuru** 3 years, 1 month ago

B. Increase the XGBoost `scale_pos_weight` parameter to adjust the balance of positive and negative weights is the correct answer.

According to https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost_hyperparameters.html, `scale_pos_weight` controls the balance of positive and negative weights. It's useful for unbalanced classes.

upvoted 3 times

A data scientist has developed a machine learning translation model for English to Japanese by using Amazon SageMaker's built-in seq2seq algorithm with 500,000 aligned sentence pairs. While testing with sample sentences, the data scientist finds that the translation quality is reasonable for an example as short as five words. However, the quality becomes unacceptable if the sentence is 100 words long. Which action will resolve the problem?

- A. Change preprocessing to use n-grams.
- B. Add more nodes to the recurrent neural network (RNN) than the largest sentence's word count.
- C. Adjust hyperparameters related to the attention mechanism.
- D. Choose a different weight initialization type.

Suggested Answer: B

Community vote distribution

C (100%)

 **cnethers** Highly Voted 3 years, 3 months ago

I agree with an answer of C

Attention mechanism. The disadvantage of an encoder-decoder framework is that model performance decreases as and when the length of the source sequence increases because of the limit of how much information the fixed-length encoded feature vector can contain. To tackle this problem, in 2015, Bahdanau et al. proposed the attention mechanism. In an attention mechanism, the decoder tries to find the location in the encoder sequence where the most important information could be located and uses that information and previously decoded words to predict the next token in the sequence.

upvoted 27 times

 **AIWave** Most Recent 10 months, 2 weeks ago

Selected Answer: C

C. By tuning attention-related hyperparameters (such as attention type, attention layer size, and dropout), the model can focus on relevant parts of the input sequence during translation.

upvoted 1 times

 **loict** 1 year, 3 months ago

Selected Answer: C

- A. NO - n-grams are more the opposite, it is to capture local information
- B. NO - it could help, but not best
- C. YES - best practice (<https://docs.aws.amazon.com/sagemaker/latest/dg/seq-2-seq-hyperparameters.html>)
- D. NO - weight are for vectorized words, they do not relate to sequences


upvoted 1 times

 **Mickey321** 1 year, 4 months ago

Selected Answer: C

Action C: Adjusting hyperparameters related to the attention mechanism can help improve the translation quality for long sentences, because the attention mechanism allows the decoder to focus on the most relevant parts of the source sentence at each time step. According to the Amazon SageMaker documentation, the seq2seq algorithm supports several types of attention mechanisms, such as dot, general, concat, and location. The data scientist can experiment with different values of the hyperparameters `attention_type`, `attention_coverage_type`, and `attention_num_hidden` to find the optimal configuration for the translation task.

upvoted 1 times

 **Ajose0** 1 year, 10 months ago

Selected Answer: C

C. Adjust hyperparameters related to the attention mechanism.

The seq2seq algorithm uses an attention mechanism to dynamically focus on relevant parts of the input sequence for each output sequence element. Increasing the attention mechanism's ability to learn dependencies between long input and output sequences might help improve the translation quality for long sentences.

The data scientist could try adjusting relevant hyperparameters such as attention depth or attention scale, or try a different attention mechanism such as scaled dot-product attention, to see if that improves the translation quality for long sentences.

upvoted 4 times

🗨️ 👤 **peterfish** 2 years, 5 months ago

Selected Answer: C

i go with C

upvoted 4 times

🗨️ 👤 **SriAkula** 2 years, 9 months ago

Ans: C

Explanation: <https://docs.aws.amazon.com/sagemaker/latest/dg/seq-2-seq-howitworks.html>

upvoted 2 times

🗨️ 👤 **AddiWei** 2 years, 10 months ago

This is such a niche question for a niche market. Geared towards someone who specializes in NLP.

upvoted 3 times

🗨️ 👤 **Juka3lj** 3 years, 2 months ago

c is correct

<https://docs.aws.amazon.com/sagemaker/latest/dg/seq-2-seq-howitworks.html>

upvoted 3 times

🗨️ 👤 **rajesriv** 3 years, 3 months ago

I believe the answer is C

<https://docs.aws.amazon.com/sagemaker/latest/dg/seq-2-seq-howitworks.html>

upvoted 3 times

A financial company is trying to detect credit card fraud. The company observed that, on average, 2% of credit card transactions were fraudulent. A data scientist trained a classifier on a year's worth of credit card transactions data. The model needs to identify the fraudulent transactions (positives) from the regular ones (negatives). The company's goal is to accurately capture as many positives as possible. Which metrics should the data scientist use to optimize the model? (Choose two.)

- A. Specificity
- B. False positive rate
- C. Accuracy
- D. Area under the precision-recall curve
- E. True positive rate

Suggested Answer: AB

Community vote distribution

DE (82%)

BD (18%)

littlewat Highly Voted 2 years, 8 months ago

D, E is the answer. we need to make the recall rate(not precision) high.
upvoted 37 times

[Removed] Highly Voted 2 years, 9 months ago

To maximize detection of fraud in real-world, imbalanced datasets, D and E should always be applied.

https://en.wikipedia.org/wiki/Sensitivity_and_specificity

<https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/>

upvoted 13 times

[Removed] 2 years, 9 months ago

Note, True positive rate = Sensitivity = Recall
upvoted 5 times

cnethers 2 years, 9 months ago

that is not correct unfortunately

Recall is = Sensitivity = False Negative which is a Type II error

Precision = specificity = False Positive which is a Type I error

I do agree that in the real world you would focus on Recall/sensitivity ie. reducing type II errors.

However, in the question, they want to reduce the False Positives so you would need to focus on precision and specificity minimizing type I errors

upvoted 5 times

yummytaco 2 years, 8 months ago

recall = sensitivity = TRUE POSITIVE RATE

[https://www.google.com/url?](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjyduhdndjwAhXtzDgGHVsSBacQFjABegQIBRAD&url=https%3A%2F%2Fwww)

[sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjyduhdndjwAhXtzDgGHVsSBacQFjABegQIBRAD&url=https%3A%2F%2Fwww](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjyduhdndjwAhXtzDgGHVsSBacQFjABegQIBRAD&url=https%3A%2F%2Fwww)

[positive-rate%2F&usg=AOvVaw10zzmY-IDlhboUTwEMnqw](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjyduhdndjwAhXtzDgGHVsSBacQFjABegQIBRAD&url=https%3A%2F%2Fwww)

upvoted 2 times

seanLu 2 years, 8 months ago

This is incorrect. The goal is to capture as many positive as possible, so false positive is not a concern. suppose we have 100 samples, 2 are positive. have two models: A has TP = 2, TN = 48, FP = 50, FN = 0. B has TP = 1, TN = 88, FP = 10, FN = 1. Model A has higher false positive rate (50/98 vs 10/88), since it captures all TP. I will go with D, E.

upvoted 6 times

loict Most Recent 9 months, 2 weeks ago

Selected Answer: DE

"accurately capture positives" means maximize TPR.

- A. NO - Specificity = $TN / (TN + FP)$ is a measure of negative cases
- B. NO - $FPR = FP / Total$
- C. NO - given the class imbalance, overall accuracy would not help
- D. YES - not sure if we need that on top of E, but other options are eliminated anyway
- E. YES - $TPR = TP / Total$, what we want

upvoted 3 times

🗳️ 👤 **teka112233** 10 months ago

Selected Answer: DE

The data scientist should use True positive rate and Area under the precision-recall curve to optimize the model.

The true positive rate (TPR) is the proportion of actual positives that are correctly identified as such. It is also known as sensitivity or recall. In this case, it is important to capture as many fraudulent transactions as possible, so the TPR should be maximized.

The area under the precision-recall curve (AUPRC) is a measure of how well the model is able to distinguish between positive and negative classes. It is a good metric to use when the classes are imbalanced, as in this case where only 2% of transactions are fraudulent. The AUPRC summarizes the trade-off between precision and recall across all possible thresholds.

Accuracy and specificity are not good metrics to use when the classes are imbalanced because they can be misleading. The false positive rate (FPR) is also not a good metric to use because it does not take into account the number of true negatives.

upvoted 1 times

🗳️ 👤 **Mickey321** 10 months, 1 week ago

Selected Answer: DE

Metric D: Area under the precision-recall curve (AUPRC) is a good metric to use for imbalanced classification problems, where the positive class is much less frequent than the negative class. Precision is the proportion of positive predictions that are correct, and recall (or true positive rate) is the proportion of positive cases that are detected. AUPRC summarizes the trade-off between precision and recall for different decision thresholds, and a higher AUPRC means that the model can achieve both high precision and high recall. Since the company's goal is to accurately capture as many positives as possible, AUPRC can help them evaluate how well the model performs on the minority class.

Metric E: True positive rate (TPR) is another good metric to use for imbalanced classification problems, as it measures the sensitivity of the model to the positive class. TPR is the same as recall, and it is the proportion of positive cases that are detected by the model. A higher TPR means that the model can identify more fraudulent transactions, which is the company's goal.

upvoted 1 times

🗳️ 👤 **Ajose0** 1 year, 4 months ago

Selected Answer: DE

The goal is to accurately capture as many fraudulent transactions (positives) as possible. To optimize the model towards this goal, the data scientist should focus on metrics that emphasize the true positive rate and the area under the precision-recall curve.

True positive rate (TPR or sensitivity) is the proportion of actual positive cases that are correctly identified as positive by the model. A higher TPR means that more fraudulent transactions are being captured.

The precision-recall curve is a graph that shows the trade-off between precision and recall for different thresholds.

upvoted 1 times

🗳️ 👤 **Ajose0** 1 year, 4 months ago

Precision is the fraction of correctly identified positive instances among all instances the model has classified as positive. Recall, also known as the true positive rate, is the fraction of positive instances that are correctly identified as positive by the model. A higher area under the precision-recall curve indicates that the model is making fewer false positive predictions and more true positive predictions, which aligns with the goal of the financial company to accurately capture as many fraudulent transactions as possible.

upvoted 1 times

🗳️ 👤 **ystotest** 1 year, 7 months ago

Selected Answer: DE

agreed with DE

upvoted 4 times

🗳️ 👤 **f4bi4n** 2 years ago

Why not A and D?

- Specificity shows us how the FNR is

- AUC PR includes Precision and Recall which shows us the ratio of TP to TP/FP and TP to TP / FN

upvoted 2 times

🗨️ 👤 **SriAkula** 2 years, 3 months ago

Answer: D&E

upvoted 1 times

🗨️ 👤 **KM226** 2 years, 6 months ago

I meant say D&E not BD

upvoted 1 times

🗨️ 👤 **KM226** 2 years, 6 months ago

Selected Answer: BD

I believe the answer is B&D, which equals F1. F1 combines precision and Sensitivity.

upvoted 2 times

🗨️ 👤 **mahmoudai** 2 years, 7 months ago

D&E is the only choices that takes False Negatives into considration

upvoted 1 times

🗨️ 👤 **f4bi4n** 2 years ago

TPR is already included in the AUC PR

TNR is not included in all others besides A

upvoted 1 times

🗨️ 👤 **AShahine21** 2 years, 7 months ago

Recall and TPR

D and E

upvoted 2 times

🗨️ 👤 **kawow** 2 years, 8 months ago

AB is the answer

upvoted 1 times

A machine learning specialist is developing a proof of concept for government users whose primary concern is security. The specialist is using Amazon SageMaker to train a convolutional neural network (CNN) model for a photo classifier application. The specialist wants to protect the data so that it cannot be accessed and transferred to a remote host by malicious code accidentally installed on the training container. Which action will provide the MOST secure protection?

- A. Remove Amazon S3 access permissions from the SageMaker execution role.
- B. Encrypt the weights of the CNN model.
- C. Encrypt the training and validation dataset.
- D. Enable network isolation for training jobs.

Suggested Answer: D

Community vote distribution

D (100%)

  **AShahine21** Highly Voted 3 years, 2 months ago

I will go with D, "cannot be accessed and transferred to a remote host by malicious code accidentally installed on the training container"

Based on the following link: <https://aws.amazon.com/blogs/security/secure-deployment-of-amazon-sagemaker-resources/>

"EnableNetworkIsolation – Set this to true when creating training, hyperparameter tuning, and inference jobs to prevent situations like malicious code being accidentally installed and transferring data to a remote host."

upvoted 18 times

  **achiko** Highly Voted 3 years, 3 months ago

If you enable network isolation, the containers can't make any outbound network calls, even to other AWS services such as Amazon S3. Additionally, no AWS credentials are made available to the container runtime environment. In the case of a training job with multiple instances, network inbound and outbound traffic is limited to the peers of each training container. SageMaker still performs download and upload operations against Amazon S3 using your SageMaker execution role in isolation from the training or inference container.

upvoted 8 times

  **Dr_Kiko** 3 years, 1 month ago

ahaha this link literally contains the answer

For example, a malicious user or code that you accidentally install on the container (in the form of a publicly available source code library) could access your data and transfer it to a remote host.

upvoted 1 times

  **james2033** Most Recent 9 months, 3 weeks ago

Selected Answer: D

'network isolation' make sense



upvoted 1 times

  **loict** 1 year, 3 months ago

Selected Answer: D

- A. NO - Remove Amazon S3 access permissions from the SageMaker execution role.
- B. NO - Encrypting the weights has nothing to do with protecting the training data
- C. NO - If the dataset is encrypted, one may still hack SageMaker instance and get access to unencrypted data
- D. YES - Enable network isolation for training jobs, data is protected end-to-end

upvoted 2 times

  **Mickey321** 1 year, 4 months ago

Selected Answer: D

Network isolation

upvoted 2 times

  **Valcilio** 1 year, 9 months ago

Selected Answer: D



It's D, not C because encrypted can be stole.

upvoted 3 times

  **halfway** 3 years, 1 month ago

I choose D. More document about it: <https://docs.aws.amazon.com/sagemaker/latest/dg/mkt-algo-model-internet-free.html>

upvoted 2 times

  **benson2021** 3 years, 2 months ago

Answer is D.


<https://aws.amazon.com/blogs/security/secure-deployment-of-amazon-sagemaker-resources/>
search for 'isolation' and there is a security parameter : EnableNetworkIsolation talking about this.

upvoted 4 times

  **Vita_Rasta84444** 3 years, 2 months ago

I would choose C

upvoted 1 times

  **omar_bahrain** 3 years, 3 months ago

most likely it is C.

<https://docs.aws.amazon.com/sagemaker/latest/dg/data-protection.html>

upvoted 5 times

  **Dr_Kiko** 3 years, 1 month ago

incorrect; you CAN transfer encrypted files even w/o a key

D is a better option

upvoted 1 times

A medical imaging company wants to train a computer vision model to detect areas of concern on patients' CT scans. The company has a large collection of unlabeled CT scans that are linked to each patient and stored in an Amazon S3 bucket. The scans must be accessible to authorized users only. A machine learning engineer needs to build a labeling pipeline.

Which set of steps should the engineer take to build the labeling pipeline with the LEAST effort?

- A. Create a workforce with AWS Identity and Access Management (IAM). Build a labeling tool on Amazon EC2 Queue images for labeling by using Amazon Simple Queue Service (Amazon SQS). Write the labeling instructions.
- B. Create an Amazon Mechanical Turk workforce and manifest file. Create a labeling job by using the built-in image classification task type in Amazon SageMaker Ground Truth. Write the labeling instructions.
- C. Create a private workforce and manifest file. Create a labeling job by using the built-in bounding box task type in Amazon SageMaker Ground Truth. Write the labeling instructions.
- D. Create a workforce with Amazon Cognito. Build a labeling web application with AWS Amplify. Build a labeling workflow backend using AWS Lambda. Write the labeling instructions.

Suggested Answer: B

Community vote distribution

C (100%)

🗳️ 👤 **[Removed]** Highly Voted 2 years, 9 months ago

I would answer C, because of the requirement that authorized users should only have access. These users will comprise the private workforce of AWS Ground Truth. See documentation: <https://docs.aws.amazon.com/sagemaker/latest/dg/sms-workforce-private.html>
upvoted 17 times

🗳️ 👤 **CharlesChiang** 2 years, 8 months ago

Agree C
upvoted 3 times

🗳️ 👤 **astonm13** 2 years, 8 months ago

Yes it is C
upvoted 1 times

🗳️ 👤 **cnethers** 2 years, 9 months ago

agree C
upvoted 2 times

🗳️ 👤 **benson2021** Highly Voted 2 years, 7 months ago

Answer is C. The question mentions that "to detect *areas* of concern on patients' CT scans", that can be achieved by bounding box instead of image classification.

bounding box: <https://docs.aws.amazon.com/sagemaker/latest/dg/sms-bounding-box.html>

image classification: <https://docs.aws.amazon.com/sagemaker/latest/dg/sms-image-classification.html>

upvoted 8 times

🗳️ 👤 **ZSun** 1 year, 2 months ago

The concern here is not about "object detection" or "image classification". it is about using "ground truth" and "private workforce"
upvoted 1 times

🗳️ 👤 **jopaca1216** Most Recent 9 months, 2 weeks ago

The principal key is that Mechanical Turk workforce does not ensure privacy of the CT Scans, and Ground Truth does.
upvoted 3 times

🗳️ 👤 **Ajose0** 1 year, 4 months ago

Selected Answer: C

This option would allow the medical imaging company to create a private workforce, which can ensure that only authorized users have access to the scans, and to use Amazon SageMaker Ground Truth to create a labeling job, which would simplify the labeling pipeline process.

upvoted 3 times

🗳️ 👤 **Shailendraa** 1 year, 9 months ago

12-sep exam

upvoted 1 times

  **ovokpus** 2 years ago

Selected Answer: C

C - GroundTruth and privacy concerns

upvoted 2 times

A company is using Amazon Textract to extract textual data from thousands of scanned text-heavy legal documents daily. The company uses this information to process loan applications automatically. Some of the documents fail business validation and are returned to human reviewers, who investigate the errors. This activity increases the time to process the loan applications.

What should the company do to reduce the processing time of loan applications?

- A. Configure Amazon Textract to route low-confidence predictions to Amazon SageMaker Ground Truth. Perform a manual review on those words before performing a business validation.
- B. Use an Amazon Textract synchronous operation instead of an asynchronous operation.
- C. Configure Amazon Textract to route low-confidence predictions to Amazon Augmented AI (Amazon A2I). Perform a manual review on those words before performing a business validation.
- D. Use Amazon Rekognition's feature to detect text in an image to extract the data from scanned images. Use this information to process the loan applications.


Suggested Answer: C

Community vote distribution

C (100%)

  **[Removed]**  2 years, 2 months ago




I agree with C, given we are evaluating model inferences (predictions). See <https://aws.amazon.com/augmented-ai/> and <https://aws.amazon.com/blogs/machine-learning/automated-monitoring-of-your-machine-learning-models-with-amazon-sagemaker-model-monitor-and-sending-predictions-to-human-review-workflows-using-amazon-a2i/>
upvoted 21 times

  **Dr_Kiko** 2 years, 1 month ago

yeap, it literally says it there

Loan or mortgage applications, tax forms, and many other financial documents contain millions of data points which need to be processed and extracted quickly and effectively. Using Amazon Textract and Amazon A2I you can extract critical data from these forms

upvoted 5 times

  **alp_ileri**  9 months, 3 weeks ago

why not a?

upvoted 1 times

  **ZSun** 8 months, 1 week ago

the differences rely on the function of these two service. Ground Truth is used for "labeling" typically, text or image label: if the service cannot automatically label the data, it send to ground truth and wait for human to label it.

but A2I is for validate prediction. The model already predict the results and human then add views to it.



upvoted 5 times

  **ZSun** 7 months, 4 weeks ago

<https://docs.aws.amazon.com/textract/latest/dg/a2i-textract.html>


<https://aws.amazon.com/blogs/machine-learning/using-amazon-textract-with-amazon-augmented-ai-for-processing-critical-documents/>

upvoted 2 times

  **alp_ileri** 9 months, 3 weeks ago

i think ground truth can do same task instead of Augmented AI



upvoted 2 times

  **Valcilio** 9 months, 3 weeks ago

Selected Answer: C

The answer is C, Augmented AI is made for review ML predictions!

upvoted 2 times



  **Ajose0** 10 months, 2 weeks ago

Selected Answer: C

By routing the low-confidence predictions to Amazon Augmented AI, the company can reduce the time to process the loan applications by leveraging human intelligence to review and validate the predictions. This way, the company can quickly address any errors or mistakes that Amazon Textract

might make, reducing the time to process loan applications.

upvoted 1 times

  **Juka3lj** 2 years, 2 months ago

correct is C

upvoted 2 times

A company ingests machine learning (ML) data from web advertising clicks into an Amazon S3 data lake. Click data is added to an Amazon Kinesis data stream by using the Kinesis Producer Library (KPL). The data is loaded into the S3 data lake from the data stream by using an Amazon Kinesis Data Firehose delivery stream. As the data volume increases, an ML specialist notices that the rate of data ingested into Amazon S3 is relatively constant. There also is an increasing backlog of data for Kinesis Data Streams and Kinesis Data Firehose to ingest. Which next step is MOST likely to improve the data ingestion rate into Amazon S3?


- A. Increase the number of S3 prefixes for the delivery stream to write to.
- B. Decrease the retention period for the data stream.
- C. Increase the number of shards for the data stream.
- D. Add more consumers using the Kinesis Client Library (KCL).

Suggested Answer: C

Community vote distribution

C (63%)

A (38%)

 **SophieSu** Highly Voted 3 years, 3 months ago

C is the correct answer. # of shard is determined by:

1. # of transactions per second times

2. data blob eg. 100 KB in size

3. One shard can Ingest 1 MB/second
upvoted 38 times

 **dolorez** Highly Voted 2 years, 7 months ago

the answer should be A - the reason why shards are not the right answer is the lack of ProvisionedThroughputExceeded exceptions that occur when a KDS has delivery into S3 and a rising backlog of data (which indicates KDS stream is still able to ingest data) in the stream, hence the S3 write limit per prefix is at fault

[https://www.amazonaws.cn/en/kinesis/data-](https://www.amazonaws.cn/en/kinesis/data-streams/faqs/#:~:text=Q%3A%20What%20happens%20if%20the%20capacity%20limits%20of%20a%20Kinesis%20stream%20are%20exceeded%20while%20th)

[streams/faqs/#:~:text=Q%3A%20What%20happens%20if%20the%20capacity%20limits%20of%20a%20Kinesis%20stream%20are%20exceeded%20while%20th](https://www.amazonaws.cn/en/kinesis/data-streams/faqs/#:~:text=Q%3A%20What%20happens%20if%20the%20capacity%20limits%20of%20a%20Kinesis%20stream%20are%20exceeded%20while%20th)

<https://docs.aws.amazon.com/AmazonS3/latest/userguide/optimizing-performance.html>

upvoted 13 times

 **u_b** 1 year, 1 month ago

from https://aws.amazon.com/kinesis/data-firehose/faqs?nc1=h_ls

Q: How often does Kinesis Data Firehose read data from my Kinesis stream?

A: Kinesis Data Firehose calls Kinesis Data Streams GetRecords() once every second for each Kinesis shard.

// and the number of records per GetRecords() is at most 10.000

=> having n shards you will get at most 10.000n records to firehose per sec. => hence firehose instead of s3 could be the limiting factor.

=> i'd also go with inc shards as the first choice (to not having to change the S3 consumers)

upvoted 1 times

 **rav009** Most Recent 7 months, 2 weeks ago

Selected Answer: A

shards is a concept in kinesis data stream.

But here the topic mention "There also is an increasing backlog of data for Kinesis Data Streams and Kinesis Data Firehose to ingest"

So even firehose has large backlogs, which means the limit comes from the S3.

So A.

upvoted 1 times

🗨️ 👤 **pupsik** 10 months, 1 week ago

Selected Answer: A

The bottle neck is not at data ingestion (i.e. Kinesis shards), but in write to S3, which throughput is bound by prefixes used.
upvoted 2 times

🗨️ 👤 **wendaz** 1 year, 2 months ago

A is not solving the issue, the bottleneck locate not in S3 but in the KDS, so we should solve the problem at the KDS, the Shards
upvoted 1 times

🗨️ 👤 **loict** 1 year, 3 months ago

I think the question is very ambiguous. "There also is an increasing backlog of data for Kinesis Data Streams and Kinesis Data Firehose to ingest.", that suggest the backlog is on the client-side (even before reaching KDS). Any component down the chain can be a bottleneck (KDS shrad, Firehose, S3). There is just no way to know in my opinion, but increasing shard is certainly the easiest to try without impact the storage structure in S3 and possibly breaking the app.
upvoted 4 times

🗨️ 👤 **teka112233** 1 year, 4 months ago

Selected Answer: C

this is my key word to solve this problem :
There also is an increasing backlog of data for Kinesis Data Streams and Kinesis Data Firehose to ingest.
so increasing the shards to ingest is the solution
upvoted 1 times

🗨️ 👤 **Mickey321** 1 year, 4 months ago

Selected Answer: C

no of shards
upvoted 1 times

🗨️ 👤 **daidaidai** 1 year, 7 months ago

Selected Answer: C

A is not correct, because "There also is an increasing backlog of data for Kinesis Data Streams and Kinesis Data Firehose to ingest", the backlog is totally not caused by S3 performance, but the shard issue.
upvoted 2 times

🗨️ 👤 **MIlb** 1 year, 8 months ago

Selected Answer: C

To increase ingest
upvoted 3 times

🗨️ 👤 **Ajose0** 1 year, 9 months ago

Selected Answer: C

The increasing backlog of data for Kinesis Data Streams and Kinesis Data Firehose indicates that the ingestion rate is slower than the data production rate. Therefore, the next step to improve the data ingestion rate into Amazon S3 is to increase the capacity of Kinesis Data Streams by increasing the number of shards. This will increase the parallelism of data processing, allowing for a higher throughput rate. Option C is the correct answer.

Option A is incorrect because increasing the number of S3 prefixes for the delivery stream will not directly affect the ingestion rate into S3.
upvoted 4 times

🗨️ 👤 **Aninina** 1 year, 12 months ago

Selected Answer: C

To improve the data ingestion rate into Amazon S3, the ML specialist should consider increasing the number of shards for the Kinesis data stream. A Kinesis data stream is made up of one or more shards, and each shard provides a fixed amount of capacity for ingesting and storing data. By increasing the number of shards, the specialist can increase the overall capacity of the data stream and improve the rate at which data is ingested.
upvoted 3 times

🗨️ 👤 **GauravLahotiML** 2 years, 1 month ago

Selected Answer: C



C is the correct answer
upvoted 3 times

🗨️ 👤 **aScientist** 2 years, 1 month ago

Selected Answer: C

Clearly S3 is a bottleneck. S3 has parallel perfomance acrtoss prefixes, thus increasing throughput

upvoted 3 times

  **niopio** 2 years, 3 months ago

Selected Answer: A

It seems S3 is the bottleneck. Adding more prefixes will help:



<https://docs.aws.amazon.com/AmazonS3/latest/userguide/optimizing-performance.html>

upvoted 6 times

  **Shailendraa** 2 years, 3 months ago

12-sep exam

upvoted 1 times

  **V_B_** 2 years, 4 months ago

The question seems to indicate the problem in the ability of S3 to load the data. Therefore, I think the answer is A.

<https://docs.aws.amazon.com/firehose/latest/dev/dynamic-partitioning.html>

upvoted 2 times

A data scientist must build a custom recommendation model in Amazon SageMaker for an online retail company. Due to the nature of the company's products, customers buy only 4-5 products every 5-10 years. So, the company relies on a steady stream of new customers. When a new customer signs up, the company collects data on the customer's preferences. Below is a sample of the data available to the data scientist.

timestamp	user_id	product_id	preference_1	...	preference_10
2020-03-04	90	25	0	...	0.374
2020-03-04	90	61	0	...	0.374
2020-02-21	203	56	1	...	0.098

How should the data scientist split the dataset into a training and test set for this use case?

- A. Shuffle all interaction data. Split off the last 10% of the interaction data for the test set.
- B. Identify the most recent 10% of interactions for each user. Split off these interactions for the test set.
- C. Identify the 10% of users with the least interaction data. Split off all interaction data from these users for the test set.
- D. Randomly select 10% of the users. Split off all interaction data from these users for the test set.

Suggested Answer: D

Community vote distribution

D (54%)

B (46%)

[Removed] 3 years, 3 months ago

I would select B, straight from this AWS example: <https://aws.amazon.com/blogs/machine-learning/building-a-customized-recommender-system-in-amazon-sagemaker/>

upvoted 26 times

ttsun 3 years, 1 month ago

the blog didn't mentioned anything about sample selection. how is B arrived?

upvoted 3 times

NicZ1111 3 years, 1 month ago

I think the answer is D because customers buy only 4-5 products every 5-10 years so it doesn't make sense to get 10% interactions for each user as a test set.

upvoted 9 times

jrff 2 years, 2 months ago

Yes, agree. Answer should be D

upvoted 2 times

VinceCar 2 years, 1 month ago

B. Recommendation should use the historical to predict the future action. B is using the older records to predict the newer records. D is using 90% user to predict other 10%, 90% is irrelevant to other 10%.

upvoted 2 times

kawaimahiro 7 months, 1 week ago

There is no difference between A and D, so I prefer B as the answer

upvoted 3 times

kyuhuck 10 months, 3 weeks ago

Selected Answer: D

The best way to split the dataset into a training and test set for this use case is to randomly select 10% of the users and split off all interaction data from these users for the test set. This is because the company relies on a steady stream of new customers, so the test set should reflect the behavior of new customers who have not been seen by the model before. The other options are not suitable because they either mix old and new customers in the test set (A and B), or they bias the test set towards users with less interaction data. References:

Amazon SageMaker Developer Guide: Train and Test Datasets

Amazon Personalize Developer Guide: Preparing and Importing Data

upvoted 2 times

🗨️ 👤 **praveenaws** 11 months, 3 weeks ago

Selected Answer: D

Primary concern is to evaluate the model's performance on completely new users then option D would be more appropriate.

upvoted 3 times

🗨️ 👤 **u_b** 1 year, 1 month ago

I'd also take time into consideration, since even for such long-lived products there might be trends or regulations or whatever that make customers prefer one over the other. => A,D are out

C will not give you a test set of desired size => out

=> B

upvoted 2 times

🗨️ 👤 **sonoluminescence** 1 year, 1 month ago

Selected Answer: D

If the primary concern is to evaluate the model's performance on completely new users (which seems to be the case for the company in question), then option D would be more appropriate.

upvoted 2 times

🗨️ 👤 **DimLam** 1 year, 2 months ago

Selected Answer: D

I would choose D.

According to the question, because of the product nature, the company doesn't rely on customer-product historical interactions for recommendations. It relies on customer explicit preferences, which are gathered on the first sign-up.

The company wants to make recommendations for these new users. It is the main source of revenue for the company.

To conduct thorough testing company needs to simulate the new users, not existing ones.

To do it we need to randomly choose some percentage of users and remove all of their transactions from the train set. And use their transactions only in test.

upvoted 3 times

🗨️ 👤 **Reju** 1 year, 3 months ago

Selected Answer: B

By selecting the most recent interactions for each user, you are simulating the scenario of having new customers in your test set. This method allows you to assess how well the model generalizes to both existing and new users.

upvoted 3 times

🗨️ 👤 **Ioict** 1 year, 3 months ago

Selected Answer: D

A. NO - the data is denormalized and users' preferences are present in multiple rows in the interactions; if we split off interactions, we introduce leakage as the same user will be present in train & test

A. NO - the data is denormalized and users' preferences are present in multiple rows in the interactions; if we split off based on the interaction, we introduce leakage as the same user will be present in train & test

C. NO - bias

D. YES - no bias and user based

upvoted 3 times

🗨️ 👤 **Ioict** 1 year, 3 months ago

Selected Answer: B

A NO introduces a bias in the training set (old interactions) vs. test set (new interactions)

C NO will have a very sparse test set

B NO the same user will be present in the training and test set; we want a user-based model, not an interaction-based one, so a user should belong to only one set

D YES - last remaining option.

upvoted 3 times

🗨️ 👤 **Mickey321** 1 year, 4 months ago

Selected Answer: B

Changing to B

upvoted 2 times

🗨️ 👤 **Mickey321** 1 year, 4 months ago

Selected Answer: D

Between B and D but the issue is 4-5 transaction every 5-10 years. Hence last 10% transaction is difficult. So going for D

upvoted 2 times

🗨️ 👤 **AmitGSL** 1 year, 6 months ago

Selected Answer: B

I would select B as it is time series data. Order might be important. So for each user, last 10% of transactions ordered by date could be a good answer.

upvoted 2 times

🗨️ 👤 **cox1960** 1 year, 8 months ago

Selected Answer: D

You want different users in training and in testing datasets, which is C or D. In addition, B is wrong since you cannot take 10% of 4-5 transactions per customer. Actually, between B, C and D, only in D you can get exactly 10%.

upvoted 3 times

🗨️ 👤 **Ajose0** 1 year, 10 months ago

Selected Answer: B

This method is appropriate because it takes into account the unique buying behavior of each customer and is likely to reflect the latest preferences of the customer. It ensures that the test set contains a representative sample of the most recent customer preferences, which is important in this use case where customer preferences change infrequently over time.

upvoted 1 times

🗨️ 👤 **aScientist** 2 years, 1 month ago

Selected Answer: B

B makes the most business sense. Since customers buy products every 4-5 years, it makes sense to be able to predict future sales from really old data. splitting the test set to be only recent interactions is the best way to test model performance from historically 'recent' data

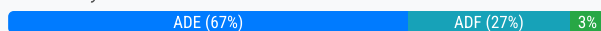
upvoted 1 times

A financial services company wants to adopt Amazon SageMaker as its default data science environment. The company's data scientists run machine learning (ML) models on confidential financial data. The company is worried about data egress and wants an ML engineer to secure the environment. Which mechanisms can the ML engineer use to control data egress from SageMaker? (Choose three.)

- A. Connect to SageMaker by using a VPC interface endpoint powered by AWS PrivateLink.
- B. Use SCPs to restrict access to SageMaker.
- C. Disable root access on the SageMaker notebook instances.
- D. Enable network isolation for training jobs and models.
- E. Restrict notebook presigned URLs to specific IPs used by the company.
- F. Protect data with encryption at rest and in transit. Use AWS Key Management Service (AWS KMS) to manage encryption keys.

Suggested Answer: BDF

Community vote distribution



SophieSu Highly Voted 3 years, 8 months ago

ADF - the concepts in ADF are explained in detail on the official Amazon Exam Readiness Exam Readiness: AWS Certified Machine Learning - Specialty. Amazon official materials do not mention other concepts in BCE.

upvoted 37 times

khchan123 1 year, 7 months ago

ADE for sure. F is for encryption and not data egress.

upvoted 2 times

scuzzy2010 3 years, 8 months ago

I agree with ADF. SCP is to control access to a service, it's not related to securing data.

upvoted 3 times

rahulw230 Highly Voted 3 years, 8 months ago

As per official document only 4 ways to do data egress Enforcing deployment in VPC, Enforcing network isolation, Restricting notebook pre-signed URLs to IPs, Disabling internet access

Correct Ans - ADE

Read Controlling data egress section

Link - <https://aws.amazon.com/blogs/machine-learning/millennium-management-secure-machine-learning-using-amazon-sagemaker/>

upvoted 28 times

Carpediem78 Most Recent 3 months, 1 week ago

Selected Answer: ADF

ADF : O

BCE : X

upvoted 1 times

ef12052 3 months ago

<https://aws.amazon.com/blogs/machine-learning/millennium-management-secure-machine-learning-using-amazon-sagemaker/>

upvoted 1 times

Togy 3 months, 2 weeks ago

Selected Answer: ABD

Correct Choices and Reasoning:

- A. Connect to SageMaker by using a VPC interface endpoint powered by AWS PrivateLink: Keeps traffic within the VPC.
- B. Use SCPs to restrict access to SageMaker: Limits authorized actions and services.
- D. Enable network isolation for training jobs and models: Prevents network access during training and inference.

Therefore, the three mechanisms that the ML engineer can use to control data egress from SageMaker are A. Connect to SageMaker by using a VPC interface endpoint powered by AWS PrivateLink, B. Use SCPs to restrict access to SageMaker, and D. Enable network isolation for training jobs and models.

upvoted 1 times

🗳️ 👤 **KarinaAsh** 7 months, 1 week ago

Selected Answer: ADF

A. Connect to SageMaker by using a VPC interface endpoint powered by AWS PrivateLink

PrivateLink ensures that communication between SageMaker and other AWS services happens entirely within the AWS network, avoiding exposure to the public internet.

This reduces the risk of unintended data egress.

D. Enable network isolation for training jobs and models

Enabling network isolation ensures that containers used for training jobs and models cannot make outbound network connections.

This prevents accidental or malicious data egress.

F. Protect data with encryption at rest and in transit. Use AWS Key Management Service (AWS KMS) to manage encryption keys

Encrypting data ensures its security even if it is inadvertently accessed or stored improperly.

KMS allows centralized and secure management of encryption keys.

upvoted 1 times

🗳️ 👤 **rookiee1111** 1 year, 2 months ago

Selected Answer: ADE

F - it takes care of data sitting in sagemaker env which is encrypted but E ensures that the services or its resources cannot be accessed outside of the allowed IP's

upvoted 1 times

🗳️ 👤 **vkajoria** 1 year, 3 months ago

My vote for ADF

upvoted 1 times

🗳️ 👤 **vkajoria** 1 year, 2 months ago

I changed my selection It is truly ADE. I read the link provided by rahulw230

upvoted 1 times

🗳️ 👤 **AIWave** 1 year, 4 months ago

Selected Answer: ABD

A = VPC endpoints are well known safety mechanism in SM so traffic doesn't leave AWS

B = service control policy can restrict access at org level

D = Network isolation limits training model access only to S3

upvoted 1 times

🗳️ 👤 **kyuhuck** 1 year, 4 months ago

Selected Answer: ADF

To control data egress from SageMaker, the ML engineer can use the following mechanisms:

Connect to SageMaker by using a VPC interface endpoint powered by AWS PrivateLink. This allows the ML engineer to access SageMaker services and resources without exposing the traffic to the public internet. This reduces the risk of data leakage and unauthorized access1 Enable network isolation for training jobs and models.

upvoted 1 times

🗳️ 👤 **sonoluminescence** 1 year, 8 months ago

Question is wrong A, B, E and D are all valid to a point.

upvoted 1 times

🗳️ 👤 **jyrajan69** 1 year, 9 months ago

The more I see it, the more likely I will go with ABD, the only answers that address the data egress issue

upvoted 1 times

🗳️ 👤 **jyrajan69** 1 year, 9 months ago

For those who are sure that is E, please explain how you can use pre-signed URLs to restrict IP's, from my understanding it is a time based access to your S3 objects, you can use policies to control access, like SCP (Service Control Policy), Isolation is definitely one option so that leaves F (Encrypting in transit and Encrypting objects) as the only possible solution as BDF

upvoted 1 times

🗨️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: ADF

A and D are for sure. The challenge between E and F. E restrict access to the notebook hence indirectly control who access it and can access data but encrypting the data is more direct way to protect the egress of the data. hence leaning more towards F

upvoted 1 times

🗨️ 👤 **mawsman** 2 years, 2 months ago

Selected Answer: ADE

Not F because the question is "to control data egress". F (encryption) is not egress control.

upvoted 3 times

🗨️ 👤 **codehive** 2 years, 2 months ago

Selected Answer: ADF

A, D, F are the mechanisms that the ML engineer can use to control data egress from SageMaker. B, C, and E do not directly control data egress from SageMaker. SCPs restrict access to AWS services, disabling root access on the SageMaker notebook instances improves security, and restricting notebook presigned URLs to specific IPs used by the company adds another layer of security, but none of these mechanisms control data egress from SageMaker.

upvoted 2 times

🗨️ 👤 **MIib** 2 years, 2 months ago

Selected Answer: ADE

<https://aws.amazon.com/blogs/machine-learning/millennium-management-secure-machine-learning-using-amazon-sagemaker/>

upvoted 2 times

🗨️ 👤 **MIib** 2 years, 2 months ago

Selected Answer: ADF

Are the correct

upvoted 2 times

A company needs to quickly make sense of a large amount of data and gain insight from it. The data is in different formats, the schemas change frequently, and new data sources are added regularly. The company wants to use AWS services to explore multiple data sources, suggest schemas, and enrich and transform the data. The solution should require the least possible coding effort for the data flows and the least possible infrastructure management.

Which combination of AWS services will meet these requirements?

A.

- ⇒ Amazon EMR for data discovery, enrichment, and transformation
- ⇒ Amazon Athena for querying and analyzing the results in Amazon S3 using standard SQL
- ⇒ Amazon QuickSight for reporting and getting insights

B.

- ⇒ Amazon Kinesis Data Analytics for data ingestion
- ⇒ Amazon EMR for data discovery, enrichment, and transformation
- ⇒ Amazon Redshift for querying and analyzing the results in Amazon S3

C.

- ⇒ AWS Glue for data discovery, enrichment, and transformation
- ⇒ Amazon Athena for querying and analyzing the results in Amazon S3 using standard SQL
- ⇒ Amazon QuickSight for reporting and getting insights

D.

- ⇒ AWS Data Pipeline for data transfer
- ⇒ AWS Step Functions for orchestrating AWS Lambda jobs for data discovery, enrichment, and transformation
- ⇒ Amazon Athena for querying and analyzing the results in Amazon S3 using standard SQL
- ⇒ Amazon QuickSight for reporting and getting insights

Suggested Answer: A

🗳️ 👤 **knightknt** Highly Voted 3 years, 2 months ago

I would choose C.

upvoted 49 times

🗳️ 👤 **ovokpus** Highly Voted 3 years ago

Answer here is C. Glue, Athena and Quicksight are serverless and need little code (only SQL)

upvoted 12 times

🗳️ 👤 **MultiCloudIronMan** Most Recent 9 months ago

C is the right answer

upvoted 2 times

🗳️ 👤 **ArunRav** 1 year, 1 month ago

Answer is C, all serverless

upvoted 2 times

🗳️ 👤 **Noname3562** 1 year, 2 months ago

I woul choose C as well

upvoted 2 times

🗳️ 👤 **endeesa** 1 year, 7 months ago

In the presence of AWS Glue, with a goal to minimise coding efforts. C is the correct answer

upvoted 1 times

🗳️ 👤 **u_b** 1 year, 7 months ago

I also chose C.

A has code/infra overhead of EMR.

B is wrong b/c you dont query S3 with redshift

D is overhead from orchestrating lambda jobs with step funcs

upvoted 3 times

🗳️ 👤 **qsergii** 1 year, 7 months ago

AWS Glue CRAWLER for data discovery

upvoted 1 times

🗨️ 👤 **Snape** 1 year, 8 months ago

C is correct

upvoted 2 times

🗨️ 👤 **jopaca1216** 1 year, 9 months ago

The correct is C

upvoted 1 times

🗨️ 👤 **Mickey321** 1 year, 10 months ago

Why no voting option?

It is option C

upvoted 4 times

🗨️ 👤 **kazivebtak** 1 year, 11 months ago

C is correct

upvoted 2 times

🗨️ 👤 **ADVIT** 1 year, 12 months ago

I think it's C

upvoted 1 times

🗨️ 👤 **mixonfreddy** 2 years ago

Answer is C, all serverless

upvoted 1 times

🗨️ 👤 **Ahmedhadi_** 2 years, 2 months ago

answer is c as data sources varies alot so requires glue crawler

upvoted 1 times

🗨️ 👤 **mite_gvg** 2 years, 2 months ago

C Is correct, you use Glue for ingestion

upvoted 2 times

🗨️ 👤 **codehive** 2 years, 2 months ago

Option C is the most suitable choice to meet the given requirements. AWS Glue is a fully managed extract, transform, and load (ETL) service that allows users to discover, enrich, and transform data easily, without the need for extensive coding. It supports different data sources, schema detection, and schema evolution, which makes it an ideal choice for the given scenario. Amazon Athena, a serverless interactive query service, allows users to run standard SQL queries against data stored in Amazon S3, which makes it easy to analyze the enriched and transformed data. Amazon QuickSight is a cloud-based business intelligence service that can connect to various data sources, including Amazon Athena, to create interactive dashboards and reports, which makes it a suitable choice for gaining insights from the data.

upvoted 1 times

🗨️ 👤 **codehive** 2 years, 2 months ago

Option A is not an ideal choice because Amazon EMR is a heavy-weight service and requires more infrastructure management than AWS Glue.

upvoted 1 times

A company is converting a large number of unstructured paper receipts into images. The company wants to create a model based on natural language processing

(NLP) to find relevant entities such as date, location, and notes, as well as some custom entities such as receipt numbers.

The company is using optical character recognition (OCR) to extract text for data labeling. However, documents are in different structures and formats, and the company is facing challenges with setting up the manual workflows for each document type. Additionally, the company trained a named entity recognition (NER) model for custom entity detection using a small sample size. This model has a very low confidence score and will require retraining with a large dataset.

Which solution for text extraction and entity detection will require the LEAST amount of effort?

- A. Extract text from receipt images by using Amazon Textract. Use the Amazon SageMaker BlazingText algorithm to train on the text for entities and custom entities.
- B. Extract text from receipt images by using a deep learning OCR model from the AWS Marketplace. Use the NER deep learning model to extract entities.
- C. Extract text from receipt images by using Amazon Textract. Use Amazon Comprehend for entity detection, and use Amazon Comprehend custom entity recognition for custom entity detection.
- D. Extract text from receipt images by using a deep learning OCR model from the AWS Marketplace. Use Amazon Comprehend for entity detection, and use Amazon Comprehend custom entity recognition for custom entity detection.

Suggested Answer: C

Reference:

<https://aws.amazon.com/blogs/machine-learning/building-an-nlp-powered-search-index-with-amazon-textract-and-amazon-comprehend/>

Community vote distribution

C (100%)

exam_prep Highly Voted 2 years, 7 months ago

C is the correct answer. You definitely need Amazon Textract service which eliminate options B & D. Between A & C - Comprehend will quicker.
upvoted 15 times

vkajoria Most Recent 9 months ago

Selected Answer: C

Textract and Comprehend will do the job
upvoted 1 times

james2033 9 months, 3 weeks ago

Selected Answer: C

Keywords 'Amazon Textract' and 'Amazon Comprehend'
upvoted 1 times

Mickey321 1 year, 4 months ago

Selected Answer: C

C indeed due to least effort
upvoted 1 times

kaike_reis 1 year, 4 months ago

Selected Answer: C

C is correct
upvoted 1 times

ADVIT 1 year, 5 months ago

I think C
upvoted 1 times

alp_ileri 1 year, 9 months ago

Selected Answer: C

I go for C
upvoted 2 times

🗨️ 👤 **Valcilio** 1 year, 9 months ago

Selected Answer: C

C is the best answer, textract is to extract data from documents and comprehend to understand the filling, objective or origin of a file.

upvoted 1 times

🗨️ 👤 **damaldon** 1 year, 11 months ago

C is correct, you can extract Entity information easily with Comprehend.

<https://aws.amazon.com/comprehend/features/>

upvoted 4 times

A company is building a predictive maintenance model based on machine learning (ML). The data is stored in a fully private Amazon S3 bucket that is encrypted at rest with AWS Key Management Service (AWS KMS) CMKs. An ML specialist must run data preprocessing by using an Amazon SageMaker Processing job that is triggered from code in an Amazon SageMaker notebook. The job should read data from Amazon S3, process it, and upload it back to the same S3 bucket.

The preprocessing code is stored in a container image in Amazon Elastic Container Registry (Amazon ECR). The ML specialist needs to grant permissions to ensure a smooth data preprocessing workflow.

Which set of actions should the ML specialist take to meet these requirements?

- A. Create an IAM role that has permissions to create Amazon SageMaker Processing jobs, S3 read and write access to the relevant S3 bucket, and appropriate KMS and ECR permissions. Attach the role to the SageMaker notebook instance. Create an Amazon SageMaker Processing job from the notebook.
- B. Create an IAM role that has permissions to create Amazon SageMaker Processing jobs. Attach the role to the SageMaker notebook instance. Create an Amazon SageMaker Processing job with an IAM role that has read and write permissions to the relevant S3 bucket, and appropriate KMS and ECR permissions.
- C. Create an IAM role that has permissions to create Amazon SageMaker Processing jobs and to access Amazon ECR. Attach the role to the SageMaker notebook instance. Set up both an S3 endpoint and a KMS endpoint in the default VPC. Create Amazon SageMaker Processing jobs from the notebook.
- D. Create an IAM role that has permissions to create Amazon SageMaker Processing jobs. Attach the role to the SageMaker notebook instance. Set up an S3 endpoint in the default VPC. Create Amazon SageMaker Processing jobs with the access key and secret key of the IAM user with appropriate KMS and ECR permissions.

Suggested Answer: D

Community vote distribution


B (52%)

A (48%)

 **spaceexplorer** Highly Voted 3 years, 2 months ago

Selected Answer: A

A; IAM assigned to SageMaker Notebook instance can be passed to other SageMaker jobs like training; processing, automl, etc., upvoted 12 times

 **kukreti18** 1 year, 11 months ago

Why should the IAM permission be assigned to create S3, when the data is already stored in S3? It only require permission to read and write data in S3. I believe A is incorrect. upvoted 3 times

 **MultiCloudIronMan** Most Recent 8 months, 2 weeks ago

Selected Answer: A

Checked this on CoPilot upvoted 1 times

 **MultiCloudIronMan** 8 months ago

I changed my mind its 'B' Option B is generally better because it provides a more secure and controlled approach to managing permissions. By separating the roles, you can ensure that the SageMaker notebook instance has only the permissions it needs to create processing jobs, while the processing job itself has the specific permissions required to access the S3 bucket, KMS, and ECR. This separation of duties enhances security and minimizes the risk of over-permissioning any single role. upvoted 3 times

 **MJSY** 9 months ago

Selected Answer: B

A is not correct, for safety and principle of least privilege, you should decouple the role of each service. upvoted 3 times

 **Chiquitabandita** 1 year ago

Selected Answer: B

based on answers from here upvoted 1 times

🗨️ 👤 **F1Fan** 1 year, 2 months ago

Selected Answer: A

Option A:

The IAM role is created with the necessary permissions to create Amazon SageMaker Processing jobs, read and write data to the relevant S3 bucket, and access the KMS CMKs and ECR container image.

The IAM role is attached to the SageMaker notebook instance, which allows the notebook to assume the role and create the Amazon SageMaker Processing job with the necessary permissions.

The Amazon SageMaker Processing job is created from the notebook, which ensures that the job has the necessary permissions to read data from S3, process it, and upload it back to the same S3 bucket.

Option B is close, but it's not entirely correct. It mentions creating an IAM role with permissions to create Amazon SageMaker Processing jobs, but it doesn't mention attaching the role to the SageMaker notebook instance. This is a crucial step, as it allows the notebook to assume the role and create the Amazon SageMaker Processing job with the necessary permissions.

upvoted 2 times

🗨️ 👤 **kyuhuck** 1 year, 4 months ago

Selected Answer: B

The correct solution for granting permissions for data preprocessing is to use the following steps: Create an IAM role that has permissions to create Amazon SageMaker Processing jobs. Attach the role to the SageMaker notebook instance. This role allows the ML specialist to run Processing jobs from the notebook code. 1. Create an Amazon SageMaker Processing job with an IAM role that has read and write permissions to the relevant S3 bucket, and appropriate KMS and ECR permissions. This role allows the Processing job to access the data in the encrypted S3 bucket, decrypt it with the KMS CMK, and pull the container image from ECR. 2. The other options are incorrect because they either miss some permissions or use unnecessary steps. For example

upvoted 2 times

🗨️ 👤 **CloudHandsOn** 1 year, 5 months ago

Selected Answer: B

Least priv.

upvoted 2 times

🗨️ 👤 **CloudHandsOn** 1 year, 5 months ago

Selected Answer: B

A. Create an IAM role with S3, KMS, ECR permissions and SageMaker Processing job creation permissions. Attach it to the SageMaker notebook instance: This option seems comprehensive as it includes all necessary permissions. However, attaching this role directly to the SageMaker notebook instance would not be sufficient for the Processing job itself. The Processing job needs its own role with appropriate permissions.

B. Create two IAM roles: one for the SageMaker notebook with permissions to create Processing jobs, and another for the Processing job itself with S3, KMS, and ECR permissions: This option is more aligned with best practices. The notebook instance and the Processing job have different roles tailored to their specific needs. This separation ensures that each service has only the permissions necessary for its operation, following the principle of least privilege.

upvoted 3 times

🗨️ 👤 **rav009** 1 year, 5 months ago

The processing job may not run on the notebook instance. AWS will provide resources to execute the job.

So A is wrong.

B.

upvoted 3 times

🗨️ 👤 **endeesa** 1 year, 7 months ago

Selected Answer: B

If we follow the principle of Least Privilege, B is correct. The notebook instance does not need access to S3 and KMS given that it is only needed to trigger the processing Job.

upvoted 2 times

🗨️ 👤 **u_b** 1 year, 7 months ago

Not A b/c it does not indicate perms given to the Job via IAM role.

=> I went with B.

upvoted 1 times

🗨️ 👤 **DimLam** 1 year, 8 months ago

Selected Answer: B

My answer is B. The notebook instance doesn't need access to S3 and ECR. This access is needed for Processing Job only.

And as a best practice of least privilege I'll choose B

upvoted 4 times

🗨️ 👤 **Reju** 1 year, 9 months ago

Selected Answer: B

where permissions are granted to the SageMaker Processing job itself and not to the notebook instance. This approach offers better security and control over permissions, making it the preferred choice for running SageMaker Processing jobs with the required access to S3, KMS, and ECR. (Follows the principle of least privilege and have more control over permissions.

upvoted 3 times

🗨️ 👤 **loict** 1 year, 9 months ago

Selected Answer: A

It says "Amazon SageMaker Processing job that is triggered from code in an Amazon SageMaker notebook." - so A or C. There is no need to create an S3 endpoint (C), that is only to allow traffic over the internet.

So A.

upvoted 1 times

🗨️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: B

Confusing between A and B. Leaning to B The main difference between A and B is the IAM role that is attached to the SageMaker notebook instance. In A, the role has permissions to access the data, the container image, and the KMS CMK. In B, the role only has permissions to create SageMaker Processing jobs. This means that in A, the notebook instance can potentially access or modify the data or the image without using a Processing job, which is not desirable. In B, the notebook instance can only create Processing jobs, and the Processing jobs themselves have a separate IAM role that grants them access to the data, the image, and the KMS CMK. This way, the data and the image are only accessed by the Processing jobs, which are more secure and controlled than the notebook instance.

upvoted 4 times

🗨️ 👤 **kaike_reis** 1 year, 10 months ago

Selected Answer: A

Letters C and D are wrong, as they bring VPC, something that is not mentioned in the problem. Letter A is correct, since Letter B asks for the creation of two different IAM roles.

upvoted 1 times

🗨️ 👤 **DimLam** 1 year, 8 months ago

What is the problem with creating two different IAM roles?

upvoted 1 times

🗨️ 👤 **ccpmad** 1 year, 11 months ago

Selected Answer: A

Option A ensures that the role has the necessary permissions to access the required resources (S3, KMS, ECR) and that the notebook has the ability to create a processing job in SageMaker seamlessly. It also follows the principle of "least privilege" by granting only the necessary permissions to perform the task without exposing more access than required.

upvoted 1 times

A data scientist has been running an Amazon SageMaker notebook instance for a few weeks. During this time, a new version of Jupyter Notebook was released along with additional software updates. The security team mandates that all running SageMaker notebook instances use the latest security and software updates provided by SageMaker.

How can the data scientist meet this requirements?

- A. Call the `CreateNotebookInstanceLifecycleConfig` API operation
- B. Create a new SageMaker notebook instance and mount the Amazon Elastic Block Store (Amazon EBS) volume from the original instance
- C. Stop and then restart the SageMaker notebook instance
- D. Call the `UpdateNotebookInstanceLifecycleConfig` API operation

Suggested Answer: C

Reference:

<https://docs.aws.amazon.com/sagemaker/latest/dg/nbi-software-updates.html>

Community vote distribution

C (100%)

 **cron0001** Highly Voted 2 years, 2 months ago

Selected Answer: C

This is correct according to official documentation.

<https://docs.aws.amazon.com/sagemaker/latest/dg/nbi-software-updates.html>

upvoted 17 times

 **Mickey321** Most Recent 10 months, 1 week ago

Selected Answer: C

Amazon SageMaker periodically tests and releases software that is installed on notebook instances, such as Jupyter Notebook, security patches, AWS SDK updates, and so on. To ensure that you have the most recent software updates, you need to stop and restart your notebook instance, either in the SageMaker console or by calling `StopNotebookInstance`.

upvoted 1 times

 **ccpmad** 11 months ago

Selected Answer: C

By stopping and restarting the SageMaker notebook instance, it will automatically apply the latest security and software updates provided by SageMaker. This process refreshes the underlying infrastructure, ensuring that the notebook instance is running with the most up-to-date software and security patches. It is a simple and effective way to comply with the security team's mandate for using the latest updates.

upvoted 1 times

 **ADVIT** 12 months ago

C per Developer Documentation <https://gmoein.github.io/files/Amazon%20SageMaker.pdf> Page44

upvoted 1 times

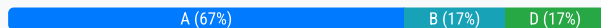
A library is developing an automatic book-borrowing system that uses Amazon Rekognition. Images of library members' faces are stored in an Amazon S3 bucket.

When members borrow books, the Amazon Rekognition CompareFaces API operation compares real faces against the stored faces in Amazon S3. The library needs to improve security by making sure that images are encrypted at rest. Also, when the images are used with Amazon Rekognition, they need to be encrypted in transit. The library also must ensure that the images are not used to improve Amazon Rekognition as a service. How should a machine learning specialist architect the solution to satisfy these requirements?

- A. Enable server-side encryption on the S3 bucket. Submit an AWS Support ticket to opt out of allowing images to be used for improving the service, and follow the process provided by AWS Support.
- B. Switch to using an Amazon Rekognition collection to store the images. Use the IndexFaces and SearchFacesByImage API operations instead of the CompareFaces API operation.
- C. Switch to using the AWS GovCloud (US) Region for Amazon S3 to store images and for Amazon Rekognition to compare faces. Set up a VPN connection and only call the Amazon Rekognition API operations through the VPN.
- D. Enable client-side encryption on the S3 bucket. Set up a VPN connection and only call the Amazon Rekognition API operations through the VPN.

Suggested Answer: B

Community vote distribution



knightknt Highly Voted 3 years, 2 months ago

A Images passed to Amazon Rekognition API operations may be stored and used to improve the service unless you unless you have opted out by visiting the AI services opt-out policy page and following the process explained there
<https://docs.aws.amazon.com/rekognition/latest/dg/security-data-encryption.html>

upvoted 9 times

tgaos 3 years, 1 month ago

So the answer is A

upvoted 1 times

BoroJohn 2 years, 6 months ago

https://docs.aws.amazon.com/organizations/latest/userguide/orgs_manage_policies_ai-opt-out.html

upvoted 2 times

mirik 1 year, 11 months ago

Yes, but server-side encryption doesn't protect at transit. Only client-side encryption can do it.

upvoted 2 times

mirik 1 year, 11 months ago

Ok, I see "encryption in transit" mean HTTPS:

Amazon Rekognition API endpoints only support secure connections over HTTPS. All communication is encrypted with Transport Layer Security (TLS).

upvoted 4 times

ovokpus Highly Voted 3 years ago

Selected Answer: A

Absolutely A. Rekognition API endpoints only support secure connections over HTTPS and all communication is encrypted in transit with TLS

upvoted 6 times

ef12052 Most Recent 3 months, 2 weeks ago

Selected Answer: A

https://aws.amazon.com/rekognition/faqs/?nc1=h_ls

upvoted 1 times

72cc81d 10 months, 2 weeks ago

Selected Answer: B

B is correct one

upvoted 1 times

🗨️ 👤 **AIWave** 1 year, 4 months ago

Selected Answer: A

client-side encryption requires you to manage the encryption and decryption of your data yourself and is an overkill.
Will go with Server side encryption. Recognition already encrypts data in transit

upvoted 2 times

🗨️ 👤 **kpr2022** 1 year, 5 months ago

Selected Answer: B

B
<https://docs.aws.amazon.com/rekognition/latest/dg/collections.html>
You can opt-out of AI data usage of aws through organizations settings.

upvoted 1 times

🗨️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: A

Option A is correct
upvoted 1 times

🗨️ 👤 **mirik** 1 year, 11 months ago

Selected Answer: D

Also, when the images are used with Amazon Rekognition. they need to be encrypted in transit

A server-site encryption doesn't encrypt images in transit. only when they are already uploaded to the S3. Only client-side encryption can encrypt the images before they are moving to AWS cloud.

upvoted 2 times

🗨️ 👤 **kaike_reis** 1 year, 10 months ago

You forgot about removing the possibility of Rekognition training.
upvoted 1 times

🗨️ 👤 **rav009** 1 year, 5 months ago

client side encryption means the key is stored on the client side. AWS has no key, how can they train?
upvoted 1 times

🗨️ 👤 **blanco750** 2 years, 3 months ago

According to Rekognition FAQs, You may opt out of having your image and video inputs used to improve or develop the quality of Amazon Rekognition and other Amazon machine-learning/artificial-intelligence technologies by using an AWS Organizations opt-out policy.
<https://aws.amazon.com/rekognition/faqs/>
upvoted 1 times

🗨️ 👤 **mlcert1** 2 years, 6 months ago

how is it A???
upvoted 1 times

A company is building a line-counting application for use in a quick-service restaurant. The company wants to use video cameras pointed at the line of customers at a given register to measure how many people are in line and deliver notifications to managers if the line grows too long. The restaurant locations have limited bandwidth for connections to external services and cannot accommodate multiple video streams without impacting other operations.

Which solution should a machine learning specialist implement to meet these requirements?

- A. Install cameras compatible with Amazon Kinesis Video Streams to stream the data to AWS over the restaurant's existing internet connection. Write an AWS Lambda function to take an image and send it to Amazon Rekognition to count the number of faces in the image. Send an Amazon Simple Notification Service (Amazon SNS) notification if the line is too long.
- B. Deploy AWS DeepLens cameras in the restaurant to capture video. Enable Amazon Rekognition on the AWS DeepLens device, and use it to trigger a local AWS Lambda function when a person is recognized. Use the Lambda function to send an Amazon Simple Notification Service (Amazon SNS) notification if the line is too long.
- C. Build a custom model in Amazon SageMaker to recognize the number of people in an image. Install cameras compatible with Amazon Kinesis Video Streams in the restaurant. Write an AWS Lambda function to take an image. Use the SageMaker endpoint to call the model to count people. Send an Amazon Simple Notification Service (Amazon SNS) notification if the line is too long.
- D. Build a custom model in Amazon SageMaker to recognize the number of people in an image. Deploy AWS DeepLens cameras in the restaurant. Deploy the model to the cameras. Deploy an AWS Lambda function to the cameras to use the model to count people and send an Amazon Simple Notification Service (Amazon SNS) notification if the line is too long.

Suggested Answer: A

Community vote distribution



spaceexplorer Highly Voted 3 years, 2 months ago

Selected Answer: D

Answer is D:

A is not correct since restaurant has limited bandwidth

B is not correct since cannot enable Rekognition service on DeepLens

C is not correct the same reason as A

upvoted 17 times

muralipr 2 years, 11 months ago

B is correct with Rekognition integrated with Deeplens and no extra configuration needed. (<https://aws.amazon.com/blogs/machine-learning/building-a-smart-garage-door-opener-with-aws-deeplens-and-amazon-rekognition/>)

upvoted 6 times

RLai 2 years, 6 months ago

In this blog, rekognition service is not running on Deeplens. It said "After you deploy the sample object detection project into AWS DeepLens, you need to change the inference (edge) Lambda function to upload image frames to Amazon S3. ... and Rekognition would do its work from the Cloud to image frames on S3."... It would still consume lots of bandwidth. So B is NOT correct.

upvoted 2 times

dunhill 2 years, 7 months ago

I also agree with D. B is incorrect due to that it's no need to do "person is recognized". It just needs to count the number of people.

upvoted 2 times

wjohnny Highly Voted 2 years, 6 months ago

Selected Answer: C

AWS will not recommend to use Deeplense in production. From <https://aws.amazon.com/deeplens/device-terms-of-use/>

upvoted 8 times

alp_ileri 2 years, 3 months ago

aws doesn't allow use in production but in evaluation. can we accept counting number of people as an evaluation?

upvoted 2 times

BTRYING 2 years, 5 months ago

<https://aws.amazon.com/deeplens/device-terms-of-use/>

upvoted 3 times

🗨️ **santi1975** Most Recent 4 months, 2 weeks ago

Selected Answer: A

Sorry guys, not B, C or D. Reasons? Deeplens is a deprecated product, not suitable for being used in real production environment (as clearly stated in its T&C), thus B & D option are out.

Between A & C, the clearly option is A. C implies the creation of a custom ML model. Making a custom model is very expensive, time consuming, error prone and a highly specialized task. Option A uses a well-known, key-in-hand service as AWS Rekognition which implies very little effort in comparison with uses a custom-made one.

I know, this option does not follow the flock, but I think that I am right.

upvoted 3 times

🗨️ **MultiCloudIronMan** 8 months, 2 weeks ago

Selected Answer: B

Yes, Amazon Rekognition can be integrated with AWS DeepLens. You can use AWS DeepLens to capture video and perform initial processing on the device. For more advanced image and video analysis, you can send frames from DeepLens to Amazon Rekognition

upvoted 1 times

🗨️ **kyuhuck** 1 year, 4 months ago

Selected Answer: D

The best solution for building a line-counting application for use in a quick-service restaurant is to use the following steps: Build a custom model in Amazon SageMaker to recognize the number of people in an image. Amazon SageMaker is a fully managed service that provides tools and workflows for building, training, and deploying machine learning models. A custom model can be tailored to the specific use case of line counting and achieve higher accuracy than a generic model. 1. Deploy AWS DeepLens cameras in the restaurant to capture video

upvoted 3 times

🗨️ **CloudHandsOn** 1 year, 5 months ago

Selected Answer: B

B. AWS DeepLens with Local Amazon Rekognition and AWS Lambda: AWS DeepLens is designed for local processing and can run models at the edge (i.e., on the device itself). This setup would enable local analysis of the video feed without the need to stream the video to the cloud, thus conserving bandwidth. Amazon Rekognition and Lambda can then be used to analyze the footage and send notifications. This option aligns well with the bandwidth limitations.

D. Custom Model on AWS DeepLens with AWS Lambda: Deploying a custom model built in SageMaker to AWS DeepLens allows for local processing of video data. This option also avoids the bandwidth issue by processing data on the device. However, developing a custom model might be more complex than using pre-built solutions like Amazon Rekognition.

upvoted 1 times

🗨️ **vikaspd** 1 year, 6 months ago

Selected Answer: D

Rekognition is a managed service. It uses API's and can't be deployed locally on devices. What we need here is local inference on the camera. AWS DeepLens comes pre-installed with a high performance, efficient, optimized inference engine for deep learning using Apache MXNet.

upvoted 2 times

🗨️ **DimLam** 1 year, 8 months ago

Selected Answer: A

I would go with A,

As DeepLens is not for production workloads, we are left with A or C. A requires less effort.

upvoted 1 times

🗨️ **seifskl** 1 year, 8 months ago

Selected Answer: B

B : <https://aws.amazon.com/ko/blogs/machine-learning/building-a-smart-garage-door-opener-with-aws-deeplens-and-amazon-rekognition/>

upvoted 1 times

🗨️ **Mickey321** 1 year, 10 months ago

Selected Answer: D

Based on the requirements, the best solution is option D. This option uses AWS DeepLens cameras to capture video and process it locally on the device, without sending any video streams to external services. This reduces the bandwidth consumption and avoids impacting other operations in the restaurant. The option also uses a custom model built in Amazon SageMaker to recognize the number of people in an image, which can be more

accurate and tailored to the specific use case than a generic face detection model. The option also deploys an AWS Lambda function to the cameras to use the model to count people and send an Amazon Simple Notification Service (Amazon SNS) notification if the line is too long.

upvoted 1 times

🗨️ 👤 **ccpmad** 1 year, 11 months ago

Selected Answer: D

it is D. "The restaurant locations have limited bandwidth for connections to external services and cannot accommodate multiple video streams without impacting other operations."

So, using Amazon Kinesis Video Streams is not a solution here.

Ok, DeepLens disappears in 2024...but this question is for 2022...

In the real world, the restaurant would buy good signal internet and use answer C, which is better solution.

upvoted 1 times

🗨️ 👤 **TQM__9MD** 1 year, 11 months ago

Selected Answer: C

C is Answer

upvoted 1 times

🗨️ 👤 **mirik** 1 year, 11 months ago

Selected Answer: C

AWS DeepLens will reach end-of-life in 31/01/2024 so, I don't think this question will even appear in the exam.

upvoted 5 times

🗨️ 👤 **MIlb** 2 years, 2 months ago

Selected Answer: D

Deeplens + lambda + model inference

upvoted 3 times

🗨️ 👤 **fez_2312** 2 years, 3 months ago

After giving this some thought, I am thinking D. Tricky, my initial answer was C. But D is a better solution - given DeepLens and counting the number of people.

upvoted 1 times

🗨️ 👤 **SANDEEP_AWS** 2 years, 3 months ago

Selected Answer: B

<https://aws.amazon.com/ko/blogs/machine-learning/building-a-smart-garage-door-opener-with-aws-deeplens-and-amazon-rekognition/>

upvoted 3 times

🗨️ 👤 **Amit11011996** 2 years, 3 months ago

According to this link,

Answer should be D, because we can directly deploy model in Deep lense to count the number of people instead a use of rekognition.

upvoted 1 times

🗨️ 👤 **lizlizliz** 2 years, 6 months ago

<https://aws.amazon.com/blogs/machine-learning/optimize-workforce-in-your-store-using-amazon-rekognition/>

B

upvoted 3 times

A company has set up and deployed its machine learning (ML) model into production with an endpoint using Amazon SageMaker hosting services. The ML team has configured automatic scaling for its SageMaker instances to support workload changes. During testing, the team notices that additional instances are being launched before the new instances are ready. This behavior needs to change as soon as possible. How can the ML team solve this issue?

- A. Decrease the cooldown period for the scale-in activity. Increase the configured maximum capacity of instances.
- B. Replace the current endpoint with a multi-model endpoint using SageMaker.
- C. Set up Amazon API Gateway and AWS Lambda to trigger the SageMaker inference endpoint.
- D. Increase the cooldown period for the scale-out activity.

Suggested Answer: A

Reference:

<https://aws.amazon.com/blogs/machine-learning/configuring-autoscaling-inference-endpoints-in-amazon-sagemaker/>

Community vote distribution

D (100%)

 **cron0001** Highly Voted 2 years, 2 months ago

Selected Answer: D

I believe this is a problem to do with scaling out (increasing the number of instances), cooldown period should be increased.

<https://docs.aws.amazon.com/autoscaling/ec2/userguide/Cooldown.html>

upvoted 14 times

 **DimLam** Most Recent 8 months, 2 weeks ago

Selected Answer: D

<https://aws.amazon.com/blogs/machine-learning/configuring-autoscaling-inference-endpoints-in-amazon-sagemaker/>

upvoted 1 times

 **Mickey321** 10 months, 1 week ago

Selected Answer: D

Option D

upvoted 1 times


 **Ajose0** 1 year, 4 months ago

Selected Answer: D

The issue is related to scaling out, specifically the fact that new instances are being launched before the existing ones are ready.

To address this issue, the ML team could consider increasing the minimum number of instances, reducing the target value for CPU utilization, or increasing the warm-up time for the instances. These actions can help to ensure that new instances are not launched until the existing ones have reached a stable state, which can prevent performance issues and ensure the reliability of the service.

upvoted 2 times

 **Ajose0** 1 year, 4 months ago

Option D, which suggests increasing the cooldown period for the scale-out activity, could potentially help to address this issue by ensuring that the new instances are not launched too quickly.

Option A, which suggests decreasing the cooldown period for the scale-in activity and increasing the maximum capacity of instances, is not an appropriate solution to the problem described. Decreasing the cooldown period for scale-in activity would result in instances being terminated too quickly, and increasing the maximum capacity of instances would not necessarily prevent new instances from being launched too quickly.

upvoted 2 times

 **ystotest** 1 year, 7 months ago

Selected Answer: D

Agreed with D. should be increased not decreased

upvoted 1 times

🗨️ 👤 **ryuhei** 1 year, 9 months ago

Selected Answer: D

Answer is "D"

upvoted 1 times

🗨️ 👤 **SDikeman62** 2 years, 1 month ago

Selected Answer: D

Definitely D.

upvoted 2 times

A telecommunications company is developing a mobile app for its customers. The company is using an Amazon SageMaker hosted endpoint for machine learning model inferences.

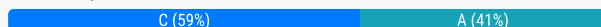
Developers want to introduce a new version of the model for a limited number of users who subscribed to a preview feature of the app. After the new version of the model is tested as a preview, developers will evaluate its accuracy. If a new version of the model has better accuracy, developers need to be able to gradually release the new version for all users over a fixed period of time.

How can the company implement the testing model with the LEAST amount of operational overhead?

- A. Update the ProductionVariant data type with the new version of the model by using the CreateEndpointConfig operation with the InitialVariantWeight parameter set to 0. Specify the TargetVariant parameter for InvokeEndpoint calls for users who subscribed to the preview feature. When the new version of the model is ready for release, gradually increase InitialVariantWeight until all users have the updated version.
- B. Configure two SageMaker hosted endpoints that serve the different versions of the model. Create an Application Load Balancer (ALB) to route traffic to both endpoints based on the TargetVariant query string parameter. Reconfigure the app to send the TargetVariant query string parameter for users who subscribed to the preview feature. When the new version of the model is ready for release, change the ALB's routing algorithm to weighted until all users have the updated version.
- C. Update the DesiredWeightsAndCapacity data type with the new version of the model by using the UpdateEndpointWeightsAndCapacities operation with the DesiredWeight parameter set to 0. Specify the TargetVariant parameter for InvokeEndpoint calls for users who subscribed to the preview feature. When the new version of the model is ready for release, gradually increase DesiredWeight until all users have the updated version.
- D. Configure two SageMaker hosted endpoints that serve the different versions of the model. Create an Amazon Route 53 record that is configured with a simple routing policy and that points to the current version of the model. Configure the mobile app to use the endpoint URL for users who subscribed to the preview feature and to use the Route 53 record for other users. When the new version of the model is ready for release, add a new model version endpoint to Route 53, and switch the policy to weighted until all users have the updated version.

Suggested Answer: D

Community vote distribution



ayatkhrisat Highly Voted 3 years, 1 month ago

Selected Answer: A

<https://docs.aws.amazon.com/sagemaker/latest/dg/model-ab-testing.html>

upvoted 19 times

ayatkhrisat 3 years ago

after reviewing it maybe C not A

upvoted 10 times

RLai 2 years, 6 months ago

https://sagemaker-examples.readthedocs.io/en/latest/sagemaker_endpoints/a_b_testing/a_b_testing.html

Should be A

upvoted 2 times

spaceexplorer Highly Voted 3 years, 2 months ago

Selected Answer: C

Answer is C, hosting two models under single endpoint has less operational overheads than two hosting endpoints

upvoted 12 times

606a82e Most Recent 2 weeks, 2 days ago

Selected Answer: A

Need to Create the ProductionVariant before you update the weight

upvoted 1 times

606a82e 2 weeks, 4 days ago

Selected Answer: A

Need to create the variant before you update the weight

upvoted 1 times

🗳️ 👤 **ef12052** 2 months, 3 weeks ago

Selected Answer: C

in option A it's mentioned that we set initial_weight to 0 which isn't true as the value should be 1 -> C

upvoted 1 times

🗳️ 👤 **MultiCloudIronMan** 8 months, 2 weeks ago

Selected Answer: A

While Option C is a viable method, Option A is generally more straightforward and aligns well with common practices for deploying and managing model versions in SageMaker. Supported by Copilot

upvoted 2 times

🗳️ 👤 **ML_2** 10 months, 2 weeks ago

Selected Answer: A

The Answer is A.

The question says "Developers want to introduce a new version of the model for a limited number of users who subscribed to a..." In order to introduce a new production version with least overhead you have to create a production variant by using CreateEndpointConfig operation and set the InitialVariantWeight to 0. You then specify the TargetVariant parameter for InvokeEndpoint calls for users who subscribed to the preview feature and gradually update the weight.

<https://docs.aws.amazon.com/sagemaker/latest/dg/model-ab-testing.html>

preview feature of the app

upvoted 2 times

🗳️ 👤 **AIWave** 1 year, 4 months ago

Selected Answer: A

- CreateEndPointConfig with initial weight to 0 prohibits any traffic to new variant
- TargetVariant Parameter in the endpoint calls made by selected users ensures new variant be used
- Change of InitialWeight causes gradual release of new variant

upvoted 1 times

🗳️ 👤 **giustino98** 1 year, 7 months ago

Selected Answer: C

Obviously C

upvoted 1 times

🗳️ 👤 **DimLam** 1 year, 8 months ago

Selected Answer: C

<https://docs.aws.amazon.com/sagemaker/latest/dg/deployment-best-practices.html>

You can modify an endpoint without taking models that are already deployed into production out of service. For example, you can add new model variants, update the ML Compute instance configurations of existing model variants, or change the distribution of traffic among model variants. To modify an endpoint, you provide a new endpoint configuration. SageMaker implements the changes without any downtime. For more information see, UpdateEndpoint and UpdateEndpointWeightsAndCapacities.

According to this doc, new variants can be deployed with UpdateEndpoint, and weights can be updated with UpdateEndpointWeightsAndCapacities.

Though for using UpdateEndpoint we need to create an endpoint config.

I will go with C

upvoted 1 times

🗳️ 👤 **Shenannigan** 1 year, 10 months ago

Selected Answer: A

The company can implement the testing model with the least amount of operational overhead by using Option A. The developers can update the ProductionVariant data type with the new version of the model by using the CreateEndpointConfig operation with the InitialVariantWeight parameter set to 0. They can specify the TargetVariant parameter for InvokeEndpoint calls for users who subscribed to the preview feature. When the new version of the model is ready for release, they can gradually increase InitialVariantWeight until all users have the updated version

upvoted 1 times

🗳️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: C

The best option for the company to implement the testing model with the least amount of operational overhead is option C. Option C uses the SageMaker feature of production variants, which allows the company to test multiple models on a single endpoint and control the traffic distribution between them. By setting the DesiredWeight parameter to 0 for the new version of the model, the company can ensure that only users who

subscribed to the preview feature will invoke the new version by specifying the TargetVariant parameter. When the new version of the model is ready for release, the company can gradually increase the DesiredWeight parameter until all users have the updated version. This option minimizes the operational overhead by avoiding the need to create and manage additional endpoints, load balancers, or DNS records.

upvoted 1 times

🗨️ 👤 **kukreti18** 2 years ago

C is correct.

The existing model will be updated using parameter DesiredWeightAndCapacity for new production variant and lead to less operational effort.

upvoted 2 times

🗨️ 👤 **dkx** 2 years, 1 month ago

This one is tricky, but I think it is testing the difference between UpdateEndpointWeightsAndCapacities and ProductionVariant

UpdateEndpointWeightsAndCapacities:

Updates variant weight of one or more variants associated with an existing endpoint, or capacity of one variant associated with an existing endpoint

https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_UpdateEndpointWeightsAndCapacities.html

ProductionVariant:

Identifies a model that you want to host and the resources chosen to deploy for hosting it. If you are deploying multiple models, tell SageMaker how to distribute traffic among the models by specifying variant weights.

https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_ProductionVariant.html

So it must be A, because the variant must exist before it is updated

This link gave me confidence to choose A

<https://docs.aws.amazon.com/sagemaker/latest/dg/model-ab-testing.html>

upvoted 4 times

🗨️ 👤 **Zhechen0912** 2 years, 3 months ago

Selected Answer: C

I agree with C.

upvoted 3 times

🗨️ 👤 **SANDEEP_AWS** 2 years, 3 months ago

Selected Answer: C

Please see step 4: <https://docs.aws.amazon.com/sagemaker/latest/dg/model-ab-testing.html> & in option A it's mentioned that we set initial_weight to 0 which isn't true as the value should be 1.

upvoted 3 times

🗨️ 👤 **matteocal** 2 years, 11 months ago

Selected Answer: C

I did not found the InitialVariantWeight, only DesiredWeight, therefore is C:

https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_DesiredWeightAndCapacity.html

upvoted 5 times

A company offers an online shopping service to its customers. The company wants to enhance the site's security by requesting additional information when customers access the site from locations that are different from their normal location. The company wants to update the process to call a machine learning (ML) model to determine when additional information should be requested. The company has several terabytes of data from its existing ecommerce web servers containing the source IP addresses for each request made to the web server. For authenticated requests, the records also contain the login name of the requesting user. Which approach should an ML specialist take to implement the new security feature in the web application?

- A. Use Amazon SageMaker Ground Truth to label each record as either a successful or failed access attempt. Use Amazon SageMaker to train a binary classification model using the factorization machines (FM) algorithm.
- B. Use Amazon SageMaker to train a model using the IP Insights algorithm. Schedule updates and retraining of the model using new log data nightly.
- C. Use Amazon SageMaker Ground Truth to label each record as either a successful or failed access attempt. Use Amazon SageMaker to train a binary classification model using the IP Insights algorithm.
- D. Use Amazon SageMaker to train a model using the Object2Vec algorithm. Schedule updates and retraining of the model using new log data nightly.

Suggested Answer: C

Community vote distribution

B (100%)

🗳️ **knightknt** Highly Voted 2 years, 8 months ago

B? because ip insights algorithm is unsupervised learning that don't need label
upvoted 12 times

🗳️ **spidy20** Highly Voted 2 years, 4 months ago

Selected Answer: B

Answer should be B
upvoted 5 times

🗳️ **delfoxete** Most Recent 10 months, 4 weeks ago

Selected Answer: B

Amazon SageMaker IP Insights is an unsupervised learning algorithm that learns the usage patterns for IPv4 addresses. It is designed to capture associations between IPv4 addresses and various entities, such as user IDs or account numbers. You can use it to identify a user attempting to log into a web service from an anomalous IP address, for example. Or you can use it to identify an account that is attempting to create computing resources from an unusual IP address. Trained IP Insight models can be hosted at an endpoint for making real-time predictions or used for processing batch transforms.
upvoted 1 times

🗳️ **Morsa** 2 years, 5 months ago

Selected Answer: B

Agree with the comments beliw
upvoted 2 times

🗳️ **tgaos** 2 years, 7 months ago

B. <https://docs.aws.amazon.com/sagemaker/latest/dg/ip-insights.html>
upvoted 3 times

🗳️ **spaceexplorer** 2 years, 8 months ago

Selected Answer: B

B; IP Insights for IP address anomaly detection
upvoted 5 times

A retail company wants to combine its customer orders with the product description data from its product catalog. The structure and format of the records in each dataset is different. A data analyst tried to use a spreadsheet to combine the datasets, but the effort resulted in duplicate records and records that were not properly combined. The company needs a solution that it can use to combine similar records from the two datasets and remove any duplicates.

Which solution will meet these requirements?

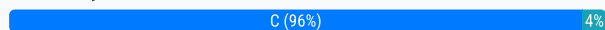
- A. Use an AWS Lambda function to process the data. Use two arrays to compare equal strings in the fields from the two datasets and remove any duplicates.
- B. Create AWS Glue crawlers for reading and populating the AWS Glue Data Catalog. Call the AWS Glue SearchTables API operation to perform a fuzzy- matching search on the two datasets, and cleanse the data accordingly.
- C. Create AWS Glue crawlers for reading and populating the AWS Glue Data Catalog. Use the FindMatches transform to cleanse the data.
- D. Create an AWS Lake Formation custom transform. Run a transformation for matching products from the Lake Formation console to cleanse the data automatically.

Suggested Answer: D

Reference:

<https://aws.amazon.com/lake-formation/features/>

Community vote distribution



spaceexplorer Highly Voted 2 years, 2 months ago

Selected Answer: C

C; Glue can use FindMatches transformation to find duplicates

upvoted 20 times

KlaudYu 2 years ago

It says "Each dataset contains records with a unique structure and format.", so C would not be correct.

upvoted 3 times

f4bi4n 2 years ago

but thats exactly the use of FindMatches:

The FindMatches transform enables you to identify duplicate or matching records in your dataset, even when the records do not have a common unique identifier and no fields match exactly

upvoted 4 times

uninit Highly Voted 1 year, 5 months ago

Selected Answer: C

It is C as described in the tutorial - <https://docs.aws.amazon.com/glue/latest/dg/machine-learning-transform-tutorial.html>

LakeFormation can also invoke a FindMatches algorithm (because it manages Data Ingestion through Glue), but we don't have a data lake in this example. No one would build a whole Data Lake - a process that takes days - only to find some matching records.

upvoted 6 times

Mickey321 Most Recent 10 months, 1 week ago

Selected Answer: C

Option C

upvoted 1 times

adisabeba 1 year, 6 months ago

Selected Answer: D

Lake Formation helps clean and prepare your data for analysis by providing a Machine Learning (ML) Transform called FindMatches for deduplication and finding matching records. For example, use FindMatches to find duplicate records in your database of restaurants, such as when one record lists "Joe's Pizza" at "121 Main St." and another shows "Joseph's Pizzeria" at "121 Main." You don't need to know anything about ML to do this.

FindMatches will simply ask you to label sets of records as either "matching" or "not matching." The system will then learn your criteria for calling a pair of records a match and will build an ML Transform that you can use to find duplicate records within a database or matching records across two

databases.

<https://aws.amazon.com/lake-formation/features/>

upvoted 1 times

  **ogm1** 2 years ago

AWS Lake Formation FindMatches is a new machine learning (ML) transform that enables you to match records across different datasets as well as identify and remove duplicate records, with little to no human intervention

Ans is D

upvoted 2 times

  **ovokpus** 2 years ago

Thing is, FindMatches is not a custom transformation in LakeFormation. And LakeFormation transforms are actually Glue jobs

upvoted 1 times

  **[Removed]** 2 years ago

D is correct

upvoted 2 times

A company provisions Amazon SageMaker notebook instances for its data science team and creates Amazon VPC interface endpoints to ensure communication between the VPC and the notebook instances. All connections to the Amazon SageMaker API are contained entirely and securely using the AWS network.

However, the data science team realizes that individuals outside the VPC can still connect to the notebook instances across the internet.

Which set of actions should the data science team take to fix the issue?

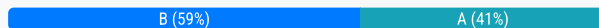
- A. Modify the notebook instances' security group to allow traffic only from the CIDR ranges of the VPC. Apply this security group to all of the notebook instances' VPC interfaces.
- B. Create an IAM policy that allows the `sagemaker:CreatePresignedNotebookInstanceUrl` and `sagemaker:DescribeNotebookInstance` actions from only the VPC endpoints. Apply this policy to all IAM users, groups, and roles used to access the notebook instances.
- C. Add a NAT gateway to the VPC. Convert all of the subnets where the Amazon SageMaker notebook instances are hosted to private subnets. Stop and start all of the notebook instances to reassign only private IP addresses.
- D. Change the network ACL of the subnet the notebook is hosted in to restrict access to anyone outside the VPC.

Suggested Answer: B

Reference:

<https://gmoein.github.io/files/Amazon%20SageMaker.pdf>

Community vote distribution



cron0001 Highly Voted 3 years, 2 months ago

Selected Answer: B

B appears to be correct according to the official source.

<https://docs.aws.amazon.com/sagemaker/latest/dg/notebook-interface-endpoint.html#notebook-private-link-restrict>

upvoted 18 times

KyaJugarHai Most Recent 3 months ago

Selected Answer: A

<https://docs.aws.amazon.com/sagemaker/latest/dg/security.html>

Security group is suffice

upvoted 1 times

2bc8f6c 5 months, 2 weeks ago

Selected Answer: B

<https://docs.aws.amazon.com/sagemaker/latest/dg/notebook-interface-endpoint.html> describes this scenario -To restrict access to only connections made from within your VPC, create an AWS Identity and Access Management policy that restricts access to only calls that come from within your VPC. Then add that policy to every AWS Identity and Access Management user, group, or role used to access the notebook instance.

upvoted 1 times

luccabastos 9 months, 3 weeks ago

Selected Answer: A

Its A.

This solutions works for all users, no more configurations needed.

upvoted 1 times

AIWave 1 year, 4 months ago

Selected Answer: B

Going with B because

- underlying notebook instances are managed by aws and can't apply security groups
- updating IAM policy only restricts connection only from VPC endpoints

upvoted 3 times

kyuhuck 1 year, 4 months ago

Selected Answer: A

The issue is that the notebook instances' security group allows inbound traffic from any source IP address, which means that anyone with the authorized URL can access the notebook instances over the internet. To fix this issue, the data science team should modify the security group to allow traffic only from the CIDR ranges of the VPC, which are the IP addresses assigned to the resources within the VPC. This way, only the VPC interface endpoints and the resources within the VPC can communicate with the notebook instances. The data science team should apply this security group to all of the notebook instances' VPC interfaces, which are the network interfaces that connect the notebook instances to the VPC.

upvoted 1 times

🗳️ 👤 **SVGoogle89** 1 year, 5 months ago

B. notebook instances are controlled by AWS service accounts and hence no access to those instances

upvoted 1 times

🗳️ 👤 **CloudHandsOn** 1 year, 5 months ago

Selected Answer: A

A. Modify the notebook instances' security group: This approach involves adjusting the security group settings to only allow traffic from the VPC's CIDR ranges. By applying this security group to all of the notebook instances' VPC interfaces, it ensures that only traffic originating from within the VPC can access the notebook instances. This is a viable solution because it directly restricts access based on the source of the traffic.

B. Create an IAM policy for VPC endpoint access: This solution involves crafting an IAM policy that restricts certain SageMaker actions to only the VPC endpoints. However, this approach might not fully address the issue of external access to the notebook instances themselves. It's more about controlling who can create or describe notebook instances, rather than restricting network access.

upvoted 2 times

🗳️ 👤 **CloudHandsOn** 1 year, 5 months ago

BUT according to here, it should be A: <https://docs.aws.amazon.com/sagemaker/latest/dg/notebook-interface-endpoint.html>

upvoted 1 times

🗳️ 👤 **CloudHandsOn** 1 year, 5 months ago

should be B*

upvoted 1 times

🗳️ 👤 **rav009** 1 year, 5 months ago

Selected Answer: A

B is talking about a policy to allow. It doesn't ban anything, it's only about allow....

So the answer can't be B.

A

upvoted 1 times

🗳️ 👤 **loict** 1 year, 9 months ago

Selected Answer: B

A. NO - it is not possible the security group of the instances, they are managed by SageMaker and will not appear in the console

B. YES - <https://docs.aws.amazon.com/sagemaker/latest/dg/notebook-interface-endpoint.html#notebook-private-link-restrict>

C. NO - subnets cannot be converted from public to private

D. NO - ACL are for the notebooks, not the network

upvoted 1 times

🗳️ 👤 **teka112233** 1 year, 9 months ago

Selected Answer: A

Based on my search, the answer is A. Modifying the notebook instances' security group to allow traffic only from the CIDR ranges of the VPC is a way to restrict access to anyone outside the VPC1.

Amazon VPC interface endpoints enable you to privately connect your VPC to supported AWS services and VPC endpoint services powered by AWS PrivateLink without requiring an internet gateway, NAT device, VPN connection, or AWS Direct Connect connection2. However, they do not prevent users from accessing the notebook instances using presigned URLs3. Therefore, options B, C and D are not correct.

upvoted 1 times

🗳️ 👤 **Maged_nader12** 1 year, 9 months ago

guys the right answer is B according to this reference:

<https://docs.aws.amazon.com/sagemaker/latest/dg/notebook-interface-endpoint.html#notebook-private-link-restrict>

To restrict access to only connections made from within your VPC, create an AWS Identity and Access Management policy that restricts access to only calls that come from within your VPC. Then add that policy to every AWS Identity and Access Management user, group, or role used to access the notebook instance.

upvoted 1 times

🗨️ 👤 **Shenannigan** 1 year, 10 months ago

Selected Answer: A

This question may be old based on this <https://aws.amazon.com/blogs/machine-learning/customize-your-amazon-sagemaker-notebook-instances-with-lifecycle-configurations-and-the-option-to-disable-internet-access/> but you can still remove all other allowed access and just add the VPC cidrs to the SGs as there is an explicit Deny for anything not explicitly allowed.

upvoted 1 times

🗨️ 👤 **Mickey321** 1 year, 10 months ago

Selected Answer: B

Option B creates an IAM policy that allows the sagemaker:CreatePresignedNotebookInstanceUrl and sagemaker:DescribeNotebookInstance actions from only the VPC endpoints. These actions are required to access the notebook instances through the Amazon SageMaker console or the AWS CLI. By applying this policy to all IAM users, groups, and roles used to access the notebook instances, the data science team can ensure that only authorized users within the VPC can connect to the notebook instances across the internet.

upvoted 1 times

🗨️ 👤 **ccpmad** 1 year, 11 months ago

Selected Answer: A

Modifying the notebook instances' security group to allow traffic only from the CIDR ranges of the VPC ensures that only connections from within the VPC are permitted. This restricts access to the notebook instances from individuals outside the VPC, effectively securing the communication and preventing unauthorized access. On the other hand, Option B, creating an IAM policy for sagemaker:CreatePresignedNotebookInstanceUrl and sagemaker:DescribeNotebookInstance actions from VPC endpoints, does not address the issue of restricting direct internet access to the notebook instances. IAM policies manage permissions for AWS service actions and resources, but they do not control network-level access.

upvoted 1 times

🗨️ 👤 **dkx** 2 years, 1 month ago

Selected Answer: A

"...the data science team realizes that individuals outside the VPC can still connect to the notebook instances across the internet.."

B states - "Create an IAM policy that allows the sagemaker:CreatePresignedNotebookInstanceUrl and sagemaker:DescribeNotebookInstance actions from only the VPC endpoints"

Ok, so now the individuals outside the VPC can't create a CreatePresignedNotebookInstanceUrl or DescribeNotebookInstance, but does that stop them from StopNotebookInstance or DeleteNotebookInstance operations?

For option A, we only allow traffic from the VPC

upvoted 2 times

🗨️ 👤 **ZSun** 2 years, 2 months ago

The problem about A is that "You can specify allow rules, but not deny rules." <https://docs.aws.amazon.com/vpc/latest/userguide/security-group-rules.html#security-group-rule-characteristics>

Therefore, you cannot restrict the unauthorized access

upvoted 1 times

🗨️ 👤 **Chelseajcole** 2 years, 3 months ago

Selected Answer: A

Should be security group thing

upvoted 3 times

A company will use Amazon SageMaker to train and host a machine learning (ML) model for a marketing campaign. The majority of data is sensitive customer data. The data must be encrypted at rest. The company wants AWS to maintain the root of trust for the master keys and wants encryption key usage to be logged.

Which implementation will meet these requirements?

- A. Use encryption keys that are stored in AWS Cloud HSM to encrypt the ML data volumes, and to encrypt the model artifacts and data in Amazon S3.
- B. Use SageMaker built-in transient keys to encrypt the ML data volumes. Enable default encryption for new Amazon Elastic Block Store (Amazon EBS) volumes.
- C. Use customer managed keys in AWS Key Management Service (AWS KMS) to encrypt the ML data volumes, and to encrypt the model artifacts and data in Amazon S3.
- D. Use AWS Security Token Service (AWS STS) to create temporary tokens to encrypt the ML storage volumes, and to encrypt the model artifacts and data in Amazon S3.

Suggested Answer: C

Community vote distribution

C (100%)

🗳️ **exam_prep** Highly Voted 2 years, 7 months ago

C is correct answer. Straight forward to use KMS.
upvoted 8 times

🗳️ **rav009** Most Recent 11 months, 3 weeks ago

Selected Answer: C

"The company wants AWS to maintain the root of trust for the master keys"
The reason A is wrong.
So C
upvoted 1 times

🗳️ **Mickey321** 1 year, 4 months ago

Selected Answer: C

option C
upvoted 1 times

🗳️ **Ajose0** 1 year, 10 months ago

Selected Answer: C

Using customer managed keys in AWS KMS will allow the company to maintain the root of trust for the master keys, and AWS KMS will log key usage. This ensures that the encryption keys used to encrypt the ML data volumes and model artifacts are properly managed and secured. Additionally, using customer managed keys allows the company to have greater control over the encryption process.
upvoted 3 times

🗳️ **mirik** 1 year, 5 months ago

"AWS Security Token Service (AWS STS) to create temporary tokens" - AWS STS also using KMS keys.
upvoted 1 times

🗳️ **Jerry84** 1 year, 11 months ago

Selected Answer: C

<https://docs.aws.amazon.com/kms/latest/developerguide/security-logging-monitoring.html>
upvoted 1 times

A machine learning specialist stores IoT soil sensor data in Amazon DynamoDB table and stores weather event data as JSON files in Amazon S3. The dataset in DynamoDB is 10 GB in size and the dataset in Amazon S3 is 5 GB in size. The specialist wants to train a model on this data to help predict soil moisture levels as a function of weather events using Amazon SageMaker. Which solution will accomplish the necessary transformation to train the Amazon SageMaker model with the LEAST amount of administrative overhead?

- A. Launch an Amazon EMR cluster. Create an Apache Hive external table for the DynamoDB table and S3 data. Join the Hive tables and write the results out to Amazon S3.
- B. Crawl the data using AWS Glue crawlers. Write an AWS Glue ETL job that merges the two tables and writes the output to an Amazon Redshift cluster.
- C. Enable Amazon DynamoDB Streams on the sensor table. Write an AWS Lambda function that consumes the stream and appends the results to the existing weather files in Amazon S3.
- D. Crawl the data using AWS Glue crawlers. Write an AWS Glue ETL job that merges the two tables and writes the output in CSV format to Amazon S3.

Suggested Answer: C

Community vote distribution

D (100%)

🗳️ **cron0001** Highly Voted 2 years, 2 months ago

Selected Answer: D

D. AWS Glue can connect with DynamoDB and join both data sets together via Glue Studio. Requiring minimal overheads
upvoted 19 times

🗳️ **DimLam** Most Recent 8 months, 2 weeks ago

Selected Answer: D

D. AWS Glue can connect with DynamoDB and join both data sets together via Glue Studio. Requiring minimal overheads
upvoted 1 times

🗳️ **ccpmad** 11 months ago

Selected Answer: D

Option D with AWS Glue crawlers and ETL job provides a straightforward and efficient way to merge the data from DynamoDB and Amazon S3 into a format suitable for training the Amazon SageMaker model with minimal administrative overhead.
upvoted 2 times

🗳️ **injoho** 1 year, 2 months ago

D.
<https://aws.amazon.com/blogs/big-data/accelerate-amazon-dynamodb-data-access-in-aws-glue-jobs-using-the-new-aws-glue-dynamodb-elt-connector/>
upvoted 3 times

🗳️ **Shailendraa** 1 year, 9 months ago

12-sep exam
upvoted 1 times

A company sells thousands of products on a public website and wants to automatically identify products with potential durability problems. The company has 1,000 reviews with date, star rating, review text, review summary, and customer email fields, but many reviews are incomplete and have empty fields. Each review has already been labeled with the correct durability result.

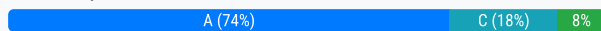
A machine learning specialist must train a model to identify reviews expressing concerns over product durability. The first model needs to be trained and ready to review in 2 days.

What is the MOST direct approach to solve this problem within 2 days?

- A. Train a custom classifier by using Amazon Comprehend.
- B. Build a recurrent neural network (RNN) in Amazon SageMaker by using Gluon and Apache MXNet.
- C. Train a built-in BlazingText model using Word2Vec mode in Amazon SageMaker.
- D. Use a built-in seq2seq model in Amazon SageMaker.

Suggested Answer: B

Community vote distribution



ayatkhrisat Highly Voted 2 years, 7 months ago

Selected Answer: A

A should be the answer
upvoted 15 times

rb39 2 years, 3 months ago

it's a sentiment analysis problem => comprehend
upvoted 4 times

ckkobe24 Highly Voted 2 years, 7 months ago

Selected Answer: C

BlazingText can also do supervised text classification
upvoted 5 times

ef12052 3 months ago

yes but only in TextClassification mode, note W2V mode... so A
upvoted 1 times

F1Fan Most Recent 9 months ago

Built-in BlazingText model using Word2Vec mode in Amazon SageMaker would likely be quicker to set up compared to using Amazon Comprehend for this specific use case. Since the problem statement mentions that the review data is already labeled with the correct durability result, preparing the training data should be relatively straightforward.

Additionally, as a built-in algorithm, BlazingText is optimized and pre-configured for text classification tasks, reducing the need for extensive customization and configuration compared to using Amazon Comprehend for this specific use case.

It's important to note that while BlazingText may be quicker to set up for this particular task, Amazon Comprehend offers a broader range of NLP capabilities and may be more suitable for other NLP tasks or scenarios where more customization and flexibility are required.

However, given the time constraint of 2 days and the specific requirement of identifying product durability concerns from reviews, training a built-in BlazingText model using Word2Vec mode in Amazon SageMaker is likely to be the more direct and quicker approach to get a working solution set up and running.

upvoted 1 times

ef12052 3 months ago

yes but only in TextClassification mode, note W2V mode... so A
upvoted 1 times

3eb0542 10 months, 2 weeks ago

Selected Answer: C

Given the time constraint of 2 days and the need for a quick solution, the most direct approach would be to choose an option that provides a ready-to-use solution without the need for extensive customization or training.

Among the given options, the most direct approach would be:

C. Train a built-in BlazingText model using Word2Vec mode in Amazon SageMaker.

This option allows you to leverage a pre-built model (BlazingText) that is optimized for text classification tasks. Word2Vec mode is suitable for analyzing text data and can quickly provide insights into sentiment or, in this case, concerns over product durability. This approach minimizes the need for extensive data preprocessing and model tuning, allowing you to focus on training and deploying the model within the given timeframe.

upvoted 2 times

🗳️ 👤 **rav009** 11 months, 3 weeks ago

Selected Answer: A

Using an existing model to do the task in 2 days.

A

upvoted 2 times

🗳️ 👤 **DimLam** 1 year, 2 months ago

Selected Answer: A

I would say A

upvoted 2 times

🗳️ 👤 **loict** 1 year, 3 months ago

Selected Answer: A

A. YES - Amazon Comprehend with multi-class mode and Augmented manifest file

B. NO - Gluon is for timeseries

C. NO - still a lot of work after generating embedding

D. NO - seq2seq is to generate text, we want to classify

upvoted 2 times

🗳️ 👤 **teka112233** 1 year, 3 months ago

Selected Answer: A

To solve the problem in 2 days, and dealing with sentiment analysis so A will be the right answer using the comprehend

AWS Comprehend is a natural language processing (NLP) service that uses machine learning to discover insights from text. It provides a range of functionalities, including detecting language and sentiment, extracting named entities and key phrases, and tagging parts of speech⁵. AWS Comprehend can automatically break down concepts like entities, phrases, and syntax in a document, which is particularly helpful for identifying events, organizations, persons, or products referenced in a document

upvoted 2 times

🗳️ 👤 **Mickey321** 1 year, 4 months ago

Selected Answer: A

The most direct approach to solve this problem within 2 days is option A, train a custom classifier by using Amazon Comprehend. By doing so, you can use Amazon Comprehend, a natural language processing (NLP) service that uses machine learning to find insights and relationships in text, to create a custom classifier that can identify reviews expressing concerns over product durability. You can use the labeled reviews as your training data and specify the durability result as the class label. Amazon Comprehend will automatically preprocess the text, extract features, and train the classifier for you. You can also use Amazon Comprehend to evaluate the performance of your classifier and deploy it as an endpoint. This way, you can train a model to solve this problem within 2 days without requiring much coding or infrastructure management.

upvoted 2 times

🗳️ 👤 **vbal** 1 year, 6 months ago

A: You can customize Amazon Comprehend for your specific requirements without the skillset required to build machine learning-based NLP solutions. Using automatic machine learning, or AutoML, Comprehend Custom builds customized NLP models on your behalf, using training data that you provide.

upvoted 1 times

🗳️ 👤 **MIlb** 1 year, 8 months ago

Selected Answer: A

Comprehend can do Custom Classification

upvoted 2 times

🗳️ 👤 **MIlb** 1 year, 8 months ago

Selected Answer: A

Comprehend can do Sentiment Analysis

upvoted 1 times

🗨️ 👤 **fez_2312** 1 year, 9 months ago

The answer is C, because of the amount of data, and the time constraint. C is the most efficient solution. Conventionally A would be the right answer, but given the time constraint the answer is C.

upvoted 1 times

🗨️ 👤 **alp_ileri** 1 year, 9 months ago

I would say blaze text. Cuz comprehend needs custom code, so we have only 2 days.

upvoted 2 times

🗨️ 👤 **ystotest** 2 years, 1 month ago

Selected Answer: A

<https://docs.aws.amazon.com/comprehend/latest/dg/how-document-classification.html>

upvoted 1 times

🗨️ 👤 **cron0001** 2 years, 8 months ago

Selected Answer: D

If the problem needs to be solved in 2 days I would avoid going with any customised solution which would eliminate A and B. As the data is labelled already we don't need an unsupervised algorithm therefore eliminating C. Which leaves us with D

upvoted 3 times

🗨️ 👤 **f4bi4n** 2 years, 6 months ago

its exactly the opposite, because its needs to be ready in 2 day I would use Comprehend ;) You don't need to write code, you have the data already available, so its faster then D

upvoted 8 times

A company that runs an online library is implementing a chatbot using Amazon Lex to provide book recommendations based on category. This intent is fulfilled by an AWS Lambda function that queries an Amazon DynamoDB table for a list of book titles, given a particular category. For testing, there are only three categories implemented as the custom slot types: "comedy," "adventure," and "documentary."

A machine learning (ML) specialist notices that sometimes the request cannot be fulfilled because Amazon Lex cannot understand the category spoken by users with utterances such as "funny," "fun," and "humor." The ML specialist needs to fix the problem without changing the Lambda code or data in DynamoDB.

How should the ML specialist fix the problem?

- A. Add the unrecognized words in the enumeration values list as new values in the slot type.
- B. Create a new custom slot type, add the unrecognized words to this slot type as enumeration values, and use this slot type for the slot.
- C. Use the AMAZON.SearchQuery built-in slot types for custom searches in the database.
- D. Add the unrecognized words as synonyms in the custom slot type.

Suggested Answer: C

Community vote distribution

D (100%)

 **ovokpus** Highly Voted 2 years ago

Selected Answer: D

D is the answer.

The unrecognized words are synonyms for "comedy", so they should be added as synonyms under the comedy slot type

see the excerpt:

"For each intent, you can specify parameters that indicate the information that the intent needs to fulfill the user's request. These parameters, or slots, have a type. A slot type is a list of values that Amazon Lex uses to train the machine learning model to recognize values for a slot. For example, you can define a slot type called "Genres." Each value in the slot type is the name of a genre, "comedy," "adventure," "documentary," etc. You can define a synonym for a slot type value. For example, you can define the synonyms "funny" and "humorous" for the value "comedy."

<https://docs.aws.amazon.com/lex/latest/dg/howitworks-custom-slots.html>

upvoted 12 times

 **knightknt** Highly Voted 2 years, 2 months ago

D? can not be C. Amazon Lex doesn't support the AMAZON.LITERAL or the AMAZON.SearchQuery built-in slot types.


<https://docs.aws.amazon.com/lex/latest/dg/howitworks-builtins-slots.html>

upvoted 9 times

 **cognito_22** 2 years, 1 month ago

<https://docs.aws.amazon.com/lex/latest/dg/howitworks-custom-slots.html>

upvoted 5 times

 **cyberfriends** Most Recent 8 months, 1 week ago

Selected Answer: D

D is the answer.

upvoted 1 times

 **Mickey321** 10 months, 1 week ago

Selected Answer: D

The best way to fix the problem is option D, add the unrecognized words as synonyms in the custom slot type. By doing so, you can map different words that have the same meaning to the same slot value, without changing the Lambda code or data in DynamoDB. For example, you can add "funny", "fun", and "humor" as synonyms for the slot value "comedy". This way, Amazon Lex can understand the category spoken by users and pass it to the Lambda function that queries the DynamoDB table for a list of book titles.

Option A, adding the unrecognized words in the enumeration values list as new values in the slot type, is not a good choice because it would create

new slot values that do not match the existing categories in the DynamoDB table. For example, if you add “funny” as a new value in the slot type, Amazon Lex would pass it to the Lambda function, which would not find any book titles for that category in the DynamoDB table.

upvoted 2 times

🗨️ 👤 **worldboss** 12 months ago

C is the answer

AMAZON.SearchQuery

As you think about what users are likely to ask, consider using a built-in or custom slot type to capture user input that is more predictable, and the AMAZON.SearchQuery slot type to capture less-predictable input that makes up the search query.

The following example shows an intent schema for SearchIntent, which uses the AMAZON.SearchQuery slot type and also includes a CityList slot that uses the AMAZON.City slot type.

Make sure that your skill uses no more than one AMAZON.SearchQuery slot per intent. The Amazon.SearchQuery slot type cannot be combined with another intent slot in sample utterances.

Each sample utterance must include a carrier phrase. The exception is that you can omit the carrier phrase in slot samples. A carrier phrase is the word or words that are part of the utterance, but not the slot, such as “search for” or “find out”.

upvoted 1 times

🗨️ 👤 **AjoseO** 1 year, 4 months ago

Selected Answer: D

The ML specialist should add the unrecognized words as synonyms in the custom slot type. This will allow Amazon Lex to understand the user's intent even if they use synonyms for the predefined slot values. By adding the synonyms, Amazon Lex will recognize them as variations of the predefined slot values and map them to the appropriate slot value. This approach can be a quick and effective way to improve the accuracy of the chatbot's understanding of user requests without having to change the Lambda code or the data in DynamoDB.

upvoted 2 times

🗨️ 👤 **[Removed]** 2 years ago

B is correct

upvoted 2 times

A manufacturing company uses machine learning (ML) models to detect quality issues. The models use images that are taken of the company's product at the end of each production step. The company has thousands of machines at the production site that generate one image per second on average.

The company ran a successful pilot with a single manufacturing machine. For the pilot, ML specialists used an industrial PC that ran AWS IoT Greengrass with a long-running AWS Lambda function that uploaded the images to Amazon S3. The uploaded images invoked a Lambda function that was written in Python to perform inference by using an Amazon SageMaker endpoint that ran a custom model. The inference results were forwarded back to a web service that was hosted at the production site to prevent faulty products from being shipped.

The company scaled the solution out to all manufacturing machines by installing similarly configured industrial PCs on each production machine. However, latency for predictions increased beyond acceptable limits. Analysis shows that the internet connection is at its capacity limit.


How can the company resolve this issue MOST cost-effectively?

- A. Set up a 10 Gbps AWS Direct Connect connection between the production site and the nearest AWS Region. Use the Direct Connect connection to upload the images. Increase the size of the instances and the number of instances that are used by the SageMaker endpoint.
- B. Extend the long-running Lambda function that runs on AWS IoT Greengrass to compress the images and upload the compressed files to Amazon S3. Decompress the files by using a separate Lambda function that invokes the existing Lambda function to run the inference pipeline.
- C. Use auto scaling for SageMaker. Set up an AWS Direct Connect connection between the production site and the nearest AWS Region. Use the Direct Connect connection to upload the images.
- D. Deploy the Lambda function and the ML models onto the AWS IoT Greengrass core that is running on the industrial PCs that are installed on each machine. Extend the long-running Lambda function that runs on AWS IoT Greengrass to invoke the Lambda function with the captured images and run the inference on the edge component that forwards the results directly to the web service.

Suggested Answer: D

Community vote distribution

D (100%)

 **cron0001** Highly Voted 1 year, 8 months ago

Selected Answer: D

D is correct according to official documentation.

<https://docs.aws.amazon.com/greengrass/v1/developerguide/ml-inference.html>

upvoted 17 times

 **rb39** 1 year, 3 months ago

A-C: excluded out, Direct Connect is expensive

upvoted 3 times

 **Ajose0** Highly Voted 10 months, 2 weeks ago

Selected Answer: D

Option D eliminates the need for internet connection since the inference is done on the edge component, and the results are directly forwarded to the web service.

This approach also reduces the need for larger instances and direct connect connections, thus being the most cost-effective solution.

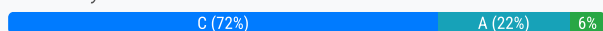
upvoted 6 times

A data scientist is using an Amazon SageMaker notebook instance and needs to securely access data stored in a specific Amazon S3 bucket. How should the data scientist accomplish this?

- A. Add an S3 bucket policy allowing GetObject, PutObject, and ListBucket permissions to the Amazon SageMaker notebook ARN as principal.
- B. Encrypt the objects in the S3 bucket with a custom AWS Key Management Service (AWS KMS) key that only the notebook owner has access to.
- C. Attach the policy to the IAM role associated with the notebook that allows GetObject, PutObject, and ListBucket operations to the specific S3 bucket.
- D. Use a script in a lifecycle configuration to configure the AWS CLI on the instance with an access key ID and secret.

Suggested Answer: C

Community vote distribution



tgaos Highly Voted 2 years, 7 months ago

Agree with the Answer C. Attach the policy to the IAM roal associated with the notebook.

upvoted 9 times

salads Highly Voted 2 years, 4 months ago

Selected Answer: C

c is the right answer

upvoted 8 times

rav009 Most Recent 7 months, 1 week ago

Selected Answer: C

Amazon SageMaker notebook ARN , I don't think there is such a thing.

So A is not right .

So C

upvoted 1 times

CloudHandsOn 11 months, 3 weeks ago

Selected Answer: C

C. Attach policy to IAM role associated with the notebook: This is a standard and recommended approach in AWS. By attaching a policy to the IAM role that the SageMaker notebook instance assumes, you can precisely control the notebook's access to the specific S3 bucket. This method follows the AWS best practice of using IAM roles for managing permissions and also allows for easier management and scalability.

A. Add an S3 bucket policy: This approach involves modifying the S3 bucket policy to grant permissions directly to the SageMaker notebook instance's ARN. While this method can effectively grant access, it is less flexible and scalable compared to using IAM roles. It directly ties the bucket's access policy to a specific resource (the notebook instance), which might not be ideal for managing access in a larger environment.

upvoted 1 times

Mickey321 1 year, 4 months ago

Selected Answer: C

The best way for the data scientist to securely access data stored in a specific Amazon S3 bucket from an Amazon SageMaker notebook instance is option C, attach the policy to the IAM role associated with the notebook that allows GetObject, PutObject, and ListBucket operations to the specific S3 bucket. By doing so, the data scientist can use IAM role-based access control to grant permissions to the notebook instance to access the S3 bucket without exposing any credentials or keys. The data scientist can also limit the scope of the permissions to only the necessary operations and resources, following the principle of least privilege.

upvoted 1 times

ccpmad 1 year, 5 months ago

Selected Answer: C

Option A suggests adding an S3 bucket policy, but it is not the recommended way to grant permissions to specific IAM roles associated with SageMaker notebook instances. Bucket policies are generally used for granting cross-account access or public access, not for specifying access for specific IAM roles.

upvoted 1 times

🗳️ 👤 **tigercorp** 1 year, 5 months ago

An IAM policy cannot attach to an ARN. An IAM policy can only attach to an IAM role or an IAM user. So the answer is C
upvoted 2 times

🗳️ 👤 **mirik** 1 year, 5 months ago

Selected Answer: A

A - we allow access to specific notebook. AIM role policy can be global and related to all user notebooks.
upvoted 1 times

🗳️ 👤 **mirik** 1 year, 5 months ago

On the other hand, in C they state "specific S3 bucket" and in the A - only "an S3 bucket". Maybe in A they add global policy to allow access to all S3 buckets?
upvoted 1 times

🗳️ 👤 **ZSun** 1 year, 8 months ago

AC are both correct answer, but A is better than C, mostly due to the limitation of IAM policy.
IAM policies: The maximum size of an IAM policy document is 6,144 characters. You can attach up to 10 policies to an IAM user, role, or group.
upvoted 1 times

🗳️ 👤 **Ajose0** 1 year, 10 months ago

Selected Answer: C

Option C ensures that the notebook instance is granted permission to access the S3 bucket without the need to provide credentials.

Option A is incorrect because it suggests adding a bucket policy that grants permission to a specific IAM principal, which is less secure than granting permission to an IAM role.
upvoted 1 times

🗳️ 👤 **ZSun** 1 year, 8 months ago

I don't agree with this. Restrict bucket access only to limited principal is much more secure than grant specific IAM principal. Restrict specific principal eliminates other visits, but grant specific IAM user permission does not exclude other visit.
upvoted 1 times

🗳️ 👤 **Shailendraa** 2 years, 3 months ago

12-sep exam
upvoted 4 times

🗳️ 👤 **[Removed]** 2 years, 6 months ago

C is correct
upvoted 4 times

🗳️ 👤 **edvardo** 2 years, 7 months ago

Selected Answer: A

Quoting the book "Data Science on AWS":
"Generally, we would use IAM identity-based policies if we need to define permissions for more than just S3, or if we have a number of S3 buckets, each with different permissions requirements. We might want to keep access control policies in the IAM environment.

We would use S3 bucket policies if we need a simple way to grant cross-account access to our S3 environment without using IAM roles, or if we reach the size limit for our IAM policy. We might want to keep access control policies in the S3 environment."

A would be the choice then.
upvoted 3 times

🗳️ 👤 **VinceCar** 2 years, 1 month ago

For A, only some operations are allowed, no specified users or roles have been granted this permission for these operations.
upvoted 1 times

🗳️ 👤 **dunhill** 2 years, 1 month ago

I am not sure but in question we don't have cross-account situation?
upvoted 1 times



🗳️ 👤 **colin1919** 2 years, 2 months ago


Based on this logic indeed A would be better.
upvoted 1 times

🗳️ 👤 **ayatkhrisat** 2 years, 7 months ago

Selected Answer: B

B is the answer
upvoted 1 times

  **VinceCar** 2 years, 1 month ago
Only "securely access" is required, not encryption.
upvoted 1 times

  **bluer1** 2 years, 7 months ago
A - for me
upvoted 2 times

A company is launching a new product and needs to build a mechanism to monitor comments about the company and its new product on social media. The company needs to be able to evaluate the sentiment expressed in social media posts, and visualize trends and configure alarms based on various thresholds.

The company needs to implement this solution quickly, and wants to minimize the infrastructure and data science resources needed to evaluate the messages.

The company already has a solution in place to collect posts and store them within an Amazon S3 bucket.

What services should the data science team use to deliver this solution?

A. Train a model in Amazon SageMaker by using the BlazingText algorithm to detect sentiment in the corpus of social media posts. Expose an endpoint that can be called by AWS Lambda. Trigger a Lambda function when posts are added to the S3 bucket to invoke the endpoint and record the sentiment in an Amazon DynamoDB table and in a custom Amazon CloudWatch metric. Use CloudWatch alarms to notify analysts of trends.

B. Train a model in Amazon SageMaker by using the semantic segmentation algorithm to model the semantic content in the corpus of social media posts. Expose an endpoint that can be called by AWS Lambda. Trigger a Lambda function when objects are added to the S3 bucket to invoke the endpoint and record the sentiment in an Amazon DynamoDB table. Schedule a second Lambda function to query recently added records and send an Amazon Simple Notification Service (Amazon SNS) notification to notify analysts of trends.

C. Trigger an AWS Lambda function when social media posts are added to the S3 bucket. Call Amazon Comprehend for each post to capture the sentiment in the message and record the sentiment in an Amazon DynamoDB table. Schedule a second Lambda function to query recently added records and send an Amazon Simple Notification Service (Amazon SNS) notification to notify analysts of trends.

D. Trigger an AWS Lambda function when social media posts are added to the S3 bucket. Call Amazon Comprehend for each post to capture the sentiment in the message and record the sentiment in a custom Amazon CloudWatch metric and in S3. Use CloudWatch alarms to notify analysts of trends.

Suggested Answer: A

Community vote distribution

D (100%)

 **cron0001** Highly Voted 3 years, 2 months ago

Selected Answer: D

D is the correct answer.

Following from the previous comment. The company wants to minimize the infrastructure and data science resources needed to evaluate the messages. Therefore any custom services would be eliminated (A and B). Similarly DynamoDB would add complexity to the infrastructure there C is eliminated. leaving D

upvoted 13 times

 **hk0308** Most Recent 6 months, 3 weeks ago

Selected Answer: C

Recording Sentiment in cloudwatch metric seems odd. DynamoDB seems more accurate.

upvoted 2 times

 **Sharath1783** 1 year, 10 months ago

Selected Answer: D

Option D is the right answer.


Following are the key terms in question to notice,

sentiment expressed in social media posts --> Comprehend

configure alarms based on various thresholds --> CloudWatch (can send alerts without SNS)

wants to minimize the infrastructure and data science resources --> AWS S3

upvoted 1 times

 **Mickey321** 1 year, 10 months ago

Selected Answer: D

The best services for the data science team to use to deliver this solution are option D, trigger an AWS Lambda function when social media posts are added to the S3 bucket, call Amazon Comprehend for each post to capture the sentiment in the message and record the sentiment in a custom Amazon CloudWatch metric and in S3, and use CloudWatch alarms to notify analysts of trends. By doing so, the data science team can use Amazon Comprehend, a natural language processing (NLP) service that uses machine learning to find insights and relationships in text, to evaluate the

sentiment expressed in social media posts. Amazon Comprehend can detect positive, negative, neutral, or mixed sentiment from text input. The data science team can also use AWS Lambda, a service that lets you run code without provisioning or managing servers, to trigger a function when posts are added to the S3 bucket and call Amazon Comprehend for each post.

upvoted 1 times

🗨️ **uninit** 2 years, 5 months ago

Selected Answer: D

Amazingly D is possible - <https://catalog.us-east-1.prod.workshops.aws/workshops/4faab440-8c3a-4527-bd11-0c88a6e6213c/en-US/30-build-the-application/400-send-sentiment-to-cloudwatch>

I was so sure of option C, because sending a sentiment to a custom CloudWatch metric just didn't make any sense. But you learn something new everyday.

upvoted 4 times

🗨️ **dolorez** 3 years, 1 month ago

This is a puzzling question, as both answers C and D miss essential steps:

C is missing DynamoDB Streams to capture new records

D is missing a notification mechanism like SNS, as CloudWatch Alarms alone can only be used as a trigger, but are not sufficient for notification

I agree that A and B should be eliminated for requiring data science development

upvoted 1 times

🗨️ **NILKK** 3 years, 2 months ago

I also do agree that D is correct answer. In A, why we are adding extra dependency of Dynamo DB.

upvoted 4 times

🗨️ **knightknt** 3 years, 2 months ago

D, blazing text is not for sentiment analysis.

The Amazon SageMaker BlazingText algorithm provides highly optimized implementations of the Word2vec and text classification algorithms. The Word2vec algorithm is useful for many downstream natural language processing (NLP) tasks, such as sentiment analysis, named entity recognition, machine translation, etc. Text classification is an important task for applications that perform web searches, information retrieval, ranking, and document classification.

upvoted 4 times

🗨️ **Maaayaaa** 2 years, 2 months ago

BlazingText can do sentiment analysis:

<https://docs.aws.amazon.com/sagemaker/latest/dg/blazingtext.html>

upvoted 1 times

A bank wants to launch a low-rate credit promotion. The bank is located in a town that recently experienced economic hardship. Only some of the bank's customers were affected by the crisis, so the bank's credit team must identify which customers to target with the promotion. However, the credit team wants to make sure that loyal customers' full credit history is considered when the decision is made.

The bank's data science team developed a model that classifies account transactions and understands credit eligibility. The data science team used the XGBoost algorithm to train the model. The team used 7 years of bank transaction historical data for training and hyperparameter tuning over the course of several days.

The accuracy of the model is sufficient, but the credit team is struggling to explain accurately why the model denies credit to some customers.

The credit team has almost no skill in data science.

What should the data science team do to address this issue in the MOST operationally efficient manner?

A. Use Amazon SageMaker Studio to rebuild the model. Create a notebook that uses the XGBoost training container to perform model training. Deploy the model at an endpoint. Enable Amazon SageMaker Model Monitor to store inferences. Use the inferences to create Shapley values that help explain model behavior. Create a chart that shows features and SHapley Additive exPlanations (SHAP) values to explain to the credit team how the features affect the model outcomes.

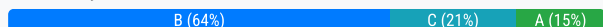
B. Use Amazon SageMaker Studio to rebuild the model. Create a notebook that uses the XGBoost training container to perform model training. Activate Amazon SageMaker Debugger, and configure it to calculate and collect Shapley values. Create a chart that shows features and SHapley Additive exPlanations (SHAP) values to explain to the credit team how the features affect the model outcomes.

C. Create an Amazon SageMaker notebook instance. Use the notebook instance and the XGBoost library to locally retrain the model. Use the `plot_importance()` method in the Python XGBoost interface to create a feature importance chart. Use that chart to explain to the credit team how the features affect the model outcomes.

D. Use Amazon SageMaker Studio to rebuild the model. Create a notebook that uses the XGBoost training container to perform model training. Deploy the model at an endpoint. Use Amazon SageMaker Processing to post-analyze the model and create a feature importance explainability chart automatically for the credit team.

Suggested Answer: C

Community vote distribution



spaceexplorer Highly Voted 3 years, 2 months ago

Selected Answer: B

B, SageMaker Model Debugger is used to generate SHAP values
upvoted 15 times

siju13 3 years, 1 month ago

<https://aws.amazon.com/blogs/machine-learning/ml-explainability-with-amazon-sagemaker-debugger/>
upvoted 7 times

V_B_ Highly Voted 2 years, 10 months ago

Selected Answer: C

I believe C is the right answer, it is simpler and more accurate than B.
upvoted 6 times

DimLam 1 year, 8 months ago

It will show only importance of features not their contribution to the final score
upvoted 2 times

2eb8df0 Most Recent 3 months, 3 weeks ago

Selected Answer: A

The problem is at inference time, not training time. So its A
upvoted 1 times

amlgeek 8 months, 3 weeks ago

I hesitate between A and B...

In the question, the credit team wants to understand the reason why the model denies credit at inference time, not at training time...

Sagemaker Model Monitor compute SHAP values at inference time while Sagemaker Debugger compute SHAP values at training time...

I'm leading more for A as an answer.

upvoted 2 times

🗳️ 👤 **kyuhuck** 1 year, 4 months ago

Selected Answer: A

The best option is to use Amazon SageMaker Studio to rebuild the model and deploy it at an endpoint. Then, use Amazon SageMaker Model Monitor to store inferences and use the inferences to create Shapley values that help explain model behavior. Shapley values are a way of attributing the contribution of each feature to the model output. They can help the credit team understand why the model makes certain decisions and how the features affect the model outcomes. A chart that shows features and SHapley Additive exPlanations (SHAP) values can be created using the SHAP library in Python. This option is the most operationally efficient because it leverages the existing XGBoost training container and the built-in capabilities of Amazon SageMaker Model Monitor and SHAP library.

upvoted 4 times

🗳️ 👤 **loict** 1 year, 9 months ago

Selected Answer: B

- A. NO - too complicated to compute SHAP
- B. YES - Debugger supports built-in SHAP
- C. NO - too complicated to compute SHAP
- D. NO - too complicated to compute SHAP

upvoted 2 times

🗳️ 👤 **ccpmad** 1 year, 11 months ago

Selected Answer: B

Option B utilizes Amazon SageMaker Studio to build and train the model, and it also activates Amazon SageMaker Debugger, which allows calculating and collecting Shapley values. These Shapley values will help explain accurately why the model denies credit to certain customers. Generating a chart that displays the features and their SHAP values will provide a visual and clear explanation of the impact of each feature on the model's decisions, making it easier for the credit team with limited data science skills to understand.

upvoted 1 times

🗳️ 👤 **Mickey321** 1 year, 11 months ago

Either A or B

Sage Maker Monitor require no experience so A is preferred while B can provide more details but depend if require knowledge to use it.

upvoted 2 times

🗳️ 👤 **Mickey321** 1 year, 11 months ago

More towards B

upvoted 1 times

🗳️ 👤 **Ahmedhadi_** 2 years, 2 months ago

Selected Answer: A

SageMaker Model Monitor is a tool that helps monitor the quality of model predictions over time by analyzing data inputs and outputs during inference. It can detect and alert when data drift or concept drift occurs, and can identify features that are most responsible for the changes in model behavior. Model Monitor can be used to continuously monitor and improve model performance, and can be integrated with SageMaker endpoints or SageMaker Pipelines.

SageMaker Debugger is a tool that helps debug machine learning models during training by analyzing the internal states of the model, such as weights and gradients, as well as the data inputs and outputs during training. It can detect and alert when common training issues occur, such as overfitting or underfitting, and can identify the root causes of these issues. Debugger can be used to improve model accuracy and convergence, and can be integrated with SageMaker training jobs.

upvoted 2 times

🗳️ 👤 **Ahmedhadi_** 2 years, 2 months ago

After reconsideration, it is actually B.

<https://aws.amazon.com/blogs/machine-learning/ml-explainability-with-amazon-sagemaker-debugger/>

upvoted 2 times

🗳️ 👤 **MIib** 2 years, 2 months ago

Selected Answer: B

Debugger because we are in the context of "training data"

upvoted 1 times

🗳️ 👤 **ZSun** 2 years, 1 month ago

There are so many explanations, but most of them are just superficial, focusing on what service is related to SHAP. This is the only one really answer the difference between A and C.

1. Both SageMaker Model Monitor and Debugger can explain model, can generate SHAP. so it should be either A or C



2. Monitor is about inference. After deploy the model, we may find some attributes start to contribute more to the model, contradict to the training dataset. This case we use SageMaker Model Monitor.

But our problem is not about deploying, is still in training stage. We only want to figure out why some customer with specific characteristics are more likely to get loan, in other words, certain feature contribute more to the prediction.

It is C !!!!

If you don't fully understand the question, stop explaining !!!

upvoted 2 times

  **ZSun** 2 years, 1 month ago

not comparing between A and C, should be A and B

upvoted 2 times

  **Amit11011996** 2 years, 3 months ago

Selected Answer: C

C is the straight forward and simpler.

upvoted 1 times

  **Amit11011996** 2 years, 3 months ago

Why not C?

'C' is the most easiest way to find out.!



upvoted 1 times

  **Chelseajcole** 2 years, 3 months ago

Selected Answer: B

This is debugger's work



upvoted 2 times

  **Ajose0** 2 years, 4 months ago

Selected Answer: A

Option A suggests using Amazon SageMaker Model Monitor to store inferences and create Shapley values that can help explain the model's behavior. This option can be more operationally efficient because it doesn't require the credit team to understand the complexities of Shapley values, and it doesn't necessarily slow down the model's inference time.

upvoted 1 times

  **Ajose0** 2 years, 4 months ago

After a review, I go with option B

upvoted 1 times

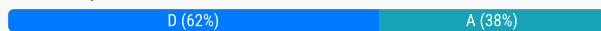
A data science team is planning to build a natural language processing (NLP) application. The application's text preprocessing stage will include part-of-speech tagging and key phrase extraction. The preprocessed text will be input to a custom classification algorithm that the data science team has already written and trained using Apache MXNet.

Which solution can the team build MOST quickly to meet these requirements?

- A. Use Amazon Comprehend for the part-of-speech tagging, key phrase extraction, and classification tasks.
- B. Use an NLP library in Amazon SageMaker for the part-of-speech tagging. Use Amazon Comprehend for the key phrase extraction. Use AWS Deep Learning Containers with Amazon SageMaker to build the custom classifier.
- C. Use Amazon Comprehend for the part-of-speech tagging and key phrase extraction tasks. Use Amazon SageMaker built-in Latent Dirichlet Allocation (LDA) algorithm to build the custom classifier.
- D. Use Amazon Comprehend for the part-of-speech tagging and key phrase extraction tasks. Use AWS Deep Learning Containers with Amazon SageMaker to build the custom classifier.

Suggested Answer: B

Community vote distribution



exam_prep Highly Voted 3 years, 1 month ago

I will go with A. Refer to link : <https://aws.amazon.com/comprehend/features/>
upvoted 18 times

ZSun 2 years, 1 month ago

whoever select A misunderstand "Custom classification", it is model for custom classificaiton, not submitting your own script!!!!
and for the above reply with document, read document first.
upvoted 1 times

tgaos 3 years, 1 month ago

Agree. A is my answer.

1. part of speech tagging : https://docs.aws.amazon.com/comprehend/latest/dg/API_PartOfSpeechTag.html

2. Key phas extraction

<https://docs.aws.amazon.com/comprehend/latest/dg/how-key-phrases.html>

3. custum classification algorithm

<https://docs.aws.amazon.com/comprehend/latest/dg/how-document-classification.html>

upvoted 10 times

ovokpus Highly Voted 3 years ago

Selected Answer: D

D is the answer. Using Apache MXNet rules out Comprehend from making the classification task

upvoted 14 times

VinceCar 2 years, 7 months ago

any reference?

upvoted 1 times

VinceCar 2 years, 7 months ago

"Automatically improve performance with optimized model training for popular frameworks like TensorFlow, PyTorch, and Apache MXNet."

<https://aws.amazon.com/cn/machine-learning/containers/>

upvoted 2 times

2bc8f6c Most Recent 5 months, 2 weeks ago

Selected Answer: D

Preprocessing using Comprehend. Then use preprocessed text as input to custom classifier.

upvoted 1 times

MJSY 9 months ago

Selected Answer: D

Amazon Comprehend can't bring your own model, the feature of "custom classification" is meaning that you can train a classifier on the service with your own data, not bring your own model on.

So the answer is definitely D.

upvoted 1 times

🗳️ 👤 **kyuhuck** 1 year, 4 months ago

Selected Answer: D

Amazon Comprehend is a natural language processing (NLP) service that can perform part-of-speech tagging and key phrase extraction tasks. AWS Deep Learning Containers are Docker images that are pre-installed with popular deep learning frameworks such as Apache MXNet. Amazon SageMaker is a fully managed service that can help build, train, and deploy machine learning models. Using Amazon Comprehend for the text preprocessing tasks and AWS Deep Learning Containers with Amazon SageMaker to build the custom classifier is the solution that can be built most quickly to meet the requirements.

References:

Amazon Comprehend

AWS Deep Learning Container

upvoted 1 times

🗳️ 👤 **rav009** 1 year, 5 months ago

Selected Answer: D

The Custom classification in AWS Comprehend cannot choose algorithm, you cannot use your own algorithm in it. You only feed dataset to it.

So A is wrong.

The data science team want to use their own MXNET model, so D.

upvoted 1 times

🗳️ 👤 **backbencher2022** 1 year, 8 months ago

Selected Answer: A

Will go with A

upvoted 1 times

🗳️ 👤 **ixdb** 1 year, 9 months ago

Selected Answer: A

A is the most quickly solution.

upvoted 1 times

🗳️ 👤 **kaike_reis** 1 year, 10 months ago

Selected Answer: D

We have to solve two NLP problems: part-of-speech tagging and key phrase extraction. Note that the custom classifier already exists and has been trained! The question asks that it be done as quickly as possible, so the idea is to use a ready-made service. Letter A is wrong, as it uses another service compared to the already created model to classify. Letter B requires development and therefore would not be the fastest solution. Letter C is wrong for the same reason as Letter A, in addition it proposes an unsupervised service (LDA) for a supervised problem. Letter D is correct.

upvoted 3 times

🗳️ 👤 **Mickey321** 1 year, 11 months ago

Selected Answer: D

Therefore, option D is the most efficient solution for building a NLP application that meets the requirements of the data science team.

upvoted 1 times

🗳️ 👤 **ADVIT** 1 year, 11 months ago

Selected Answer: A

Quickest A

upvoted 4 times

🗳️ 👤 **kukreti18** 1 year, 11 months ago

Latest is A.

upvoted 1 times

🗳️ 👤 **MIlb** 2 years, 2 months ago

Selected Answer: D

The other mxnet model is the key

upvoted 3 times

🗨️ 👤 **Ajose0** 2 years, 4 months ago

Selected Answer: D

option D is the most appropriate answer, given that the team has already written and trained a custom classification algorithm using Apache MXNet.

Option D allows the team to use Amazon Comprehend for part-of-speech tagging and key phrase extraction, while also using AWS Deep Learning Containers with Amazon SageMaker to build and deploy the custom classifier.

upvoted 3 times

🗨️ 👤 **wolfsong** 2 years, 5 months ago

D for me. Question says "The preprocessed text WILL be input to a custom classification algorithm that the data science team has already written and trained using Apache MXNet". So for some reason they want to use MXNet to do the classification, not Amazon Comprehend. So using MXNet for classification is a part of their requirement. How do we meet these requirements quickly? Well, use Amazon Comprehend for part-of-speech and key phrase tasks; and use container for the MXNet stuff.

upvoted 5 times

🗨️ 👤 **drcok87** 2 years, 4 months ago

I had selected "A" in my first go, thanks for understanding the question. Although, comprehend does all three, since they have already built custom classification, we only need to provide solution for first two.

D for me too.

upvoted 2 times

🗨️ 👤 **hamimelon** 2 years, 5 months ago

The question did not make it clear whether the new solution has to use the custom model that the team built or not.

upvoted 2 times

🗨️ 👤 **tsangckl** 2 years, 7 months ago

Selected Answer: A

A for me

upvoted 3 times

🗨️ 👤 **ystotest** 2 years, 7 months ago

Selected Answer: A

Agreed with A, Comprehend 3 functions

upvoted 3 times

A machine learning (ML) specialist must develop a classification model for a financial services company. A domain expert provides the dataset, which is tabular with 10,000 rows and 1,020 features. During exploratory data analysis, the specialist finds no missing values and a small percentage of duplicate rows. There are correlation scores of > 0.9 for 200 feature pairs. The mean value of each feature is similar to its 50th percentile.

Which feature engineering strategy should the ML specialist use with Amazon SageMaker?

- A. Apply dimensionality reduction by using the principal component analysis (PCA) algorithm.
- B. Drop the features with low correlation scores by using a Jupyter notebook.
- C. Apply anomaly detection by using the Random Cut Forest (RCF) algorithm.
- D. Concatenate the features with high correlation scores by using a Jupyter notebook.

Suggested Answer: C



Community vote distribution

A (100%)

  **ovokpus** Highly Voted 2 years ago

Selected Answer: A

Dimensions are too high. Use PCA
upvoted 10 times

  **LydiaGom** Highly Voted 2 years, 1 month ago

A should be the answer to avoid the curse of dimensionality
upvoted 7 times


  **chet100** Most Recent 10 months, 1 week ago

Easy choice. Always choose PCA for dim reduction
upvoted 2 times

  **Mickey321** 11 months ago



Selected Answer: A

the best feature engineering strategy for the ML specialist to use with Amazon SageMaker is to apply dimensionality reduction by using the PCA algorithm.
upvoted 3 times

  **Gaby999** 1 year, 2 months ago

Selected Answer: A

Given that the dataset has 1,020 features and 200 of them are highly correlated, it is likely that the dataset suffers from multicollinearity. In such cases, dimensionality reduction techniques like principal component analysis (PCA) can be used to transform the data into a lower dimensional space without losing much information. Therefore, option A, "Apply dimensionality reduction by using the principal component analysis (PCA) algorithm" is the most appropriate feature engineering strategy for the ML specialist to use with Amazon SageMaker. This would help reduce the computational complexity of the model, improve model performance, and help to avoid overfitting.
upvoted 1 times

  **Ajose0** 1 year, 4 months ago

Selected Answer: A

A. Apply dimensionality reduction by using the principal component analysis (PCA) algorithm.

Since the dataset has many features, and a significant number of them have high correlation scores, the model may suffer from the curse of dimensionality. To reduce the dimensionality of the dataset, the specialist can use a technique like PCA, which reduces the number of features while still retaining the maximum amount of information. PCA can help remove redundant features and improve the model's performance by reducing the chances of overfitting. Additionally, since there are no missing values and a small percentage of duplicate rows, no data cleaning techniques like anomaly detection or dropping the features are required. Concatenating features with high correlation scores is not an appropriate strategy since it may lead to collinearity issues.

upvoted 1 times

  **drcok87** 1 year, 4 months ago

A PCA: PCA is a linear dimensionality reduction technique (algorithm) that transforms a set of correlated variables (p) into a smaller k ($k < p$) number of uncorrelated variables called principal components while retaining as much of the variation in the original dataset as possible


upvoted 1 times

  **Peeking** 1 year, 6 months ago

Selected Answer: A

Choosing C is answer by ExamTopics is completely laughable.

upvoted 1 times

  **DJiang** 2 years, 1 month ago

Selected Answer: A

I think it's A.

upvoted 4 times

A manufacturing company asks its machine learning specialist to develop a model that classifies defective parts into one of eight defect types. The company has provided roughly 100,000 images per defect type for training. During the initial training of the image classification model, the specialist notices that the validation accuracy is 80%, while the training accuracy is 90%. It is known that human-level performance for this type of image classification is around 90%.

What should the specialist consider to fix this issue?

- A. A longer training time
- B. Making the network larger
- C. Using a different optimizer
- D. Using some form of regularization

Suggested Answer: D

Reference:

<https://acloud.guru/forums/aws-certified-machine-learning-specialty/discussion/-MGdBUKmq02zC3uQ4VL/AWS%20Exam%20Machine%20Learning>

Community vote distribution

D (90%)

10%

 **bluer1** Highly Voted 2 years, 8 months ago

D - over fitting problem.
upvoted 16 times

 **Ajose0** Highly Voted 1 year, 10 months ago

Selected Answer: D

The specialist should consider using some form of regularization to fix this issue. Regularization techniques such as dropout or L2 regularization can help prevent overfitting, which can occur when the model performs well on the training data but poorly on the validation data.

Option A, a longer training time, might not necessarily fix the issue and could lead to overfitting if the model is already performing well on the training data.

Option B, making the network larger, could also lead to overfitting and may not be necessary if the current network architecture is sufficient to perform the classification task.

Option C, using a different optimizer, might not necessarily fix the issue and could lead to slower convergence or worse performance.

Therefore, option D, using some form of regularization, is the most appropriate solution to consider in this situation.

upvoted 6 times

 **vkajoria** Most Recent 9 months ago

Selected Answer: D

some form of regularization
upvoted 1 times

 **giustino98** 1 year, 1 month ago

Selected Answer: B

I wouldn't go with D since it doesn't seem an overfitting problem considering training accuracy is not so high. So the main problem here is to get an higher accuracy even on training set. I would go with A or B
upvoted 1 times

 **ArturoZapatero** 1 year, 4 months ago

A - IMO it's an underfitting problem, as training accuracy is not better than baseline error (human accuracy). Would consider B as well, but it may actually decrease accuracy.
upvoted 1 times

 **Mickey321** 1 year, 5 months ago

Selected Answer: D

typical overfitting problem



upvoted 1 times

  **ystotest** 2 years, 1 month ago

Selected Answer: D



typical overfitting problem

upvoted 1 times

  **DD4** 2 years, 3 months ago



C - It is not a overfitting problem as the training accuracy stands at 90%, which is at same level of human performance. That means the algorithm used is not optimized for this problem. So, some other algorithm should applied for this problem.

upvoted 2 times

  **KlaudYu** 2 years, 5 months ago


I'd go A. Regularization could not guarantee higher validation accuracy.

upvoted 2 times

  **rhuanca** 2 years, 6 months ago

I believe answer is B , because clearly it is a overfitting problem , if we reduce complexity the accurate will reduce close to 80% ... But human works can reach up to 90% .

upvoted 1 times

  **rhuanca** 2 years, 6 months ago

I mean looks like a overfitting problem....

upvoted 2 times

A machine learning specialist needs to analyze comments on a news website with users across the globe. The specialist must find the most discussed topics in the comments that are in either English or Spanish.
What steps could be used to accomplish this task? (Choose two.)

- A. Use an Amazon SageMaker BlazingText algorithm to find the topics independently from language. Proceed with the analysis.
- B. Use an Amazon SageMaker seq2seq algorithm to translate from Spanish to English, if necessary. Use a SageMaker Latent Dirichlet Allocation (LDA) algorithm to find the topics.
- C. Use Amazon Translate to translate from Spanish to English, if necessary. Use Amazon Comprehend topic modeling to find the topics.
- D. Use Amazon Translate to translate from Spanish to English, if necessary. Use Amazon Lex to extract topics from the content.
- E. Use Amazon Translate to translate from Spanish to English, if necessary. Use Amazon SageMaker Neural Topic Model (NTM) to find the topics.

Suggested Answer: B

Reference:

<https://docs.aws.amazon.com/sagemaker/latest/dg/lda.html>

Community vote distribution



LydiaGom Highly Voted 3 years, 1 month ago

C and E

B needs to build custom model

upvoted 28 times

tgaos 3 years, 1 month ago

The SageMaker seq2seq algorithm is a supervised learning algorithm. And it needs to train then translate. translate can directly use to translate from Spanish to English

upvoted 7 times

hamimelon 2 years, 6 months ago

The question did not say you cannot build a custom model. They have a ML specialist, so building a custom model shouldn't be a problem.

upvoted 3 times

NILKK Highly Voted 3 years, 2 months ago

It asked 2 answers, but I can see only one answer. Please advise. Thanks!

upvoted 9 times

Carpediem78 Most Recent 3 months ago

Selected Answer: E

(Choose two.)

C,E

upvoted 1 times

MultiCloudIronMan 8 months, 1 week ago

Selected Answer: C

C E - Amazon Translate can handle the translation from Spanish to English, ensuring all comments are in a single language.

Amazon Comprehend provides robust topic modeling capabilities to identify the most discussed topics in the translated comments.

Use Amazon Translate to translate from Spanish to English, if necessary. Use Amazon SageMaker Neural Topic Model (NTM) to find the topics.

Amazon SageMaker Neural Topic Model (NTM) is an unsupervised learning algorithm designed for topic modeling, which can effectively identify topics in the translated comments

upvoted 1 times

amlgeek 8 months, 3 weeks ago

CE are the answers.

C use Amazon Comprehend for topic modeling - use LDA (<https://docs.aws.amazon.com/comprehend/latest/dg/topic-modeling.html>)

E is using NTM (<https://docs.aws.amazon.com/sagemaker/latest/dg/ntm.html>)

For ease of use, I will start with Amazon Comprehend and go to NTM if the success criteria isn't met.

upvoted 1 times

🗳️ 👤 **MJSY** 9 months ago

Selected Answer: E

C, E is the correct answer.

upvoted 1 times

🗳️ 👤 **Chiquitabandita** 1 year ago

Selected Answer: E

C & E based on comments, but you are not allowed to select multiple choices.

upvoted 1 times

🗳️ 👤 **AIWave** 1 year, 4 months ago

C and E

- Use translate so that text is in common language

- In options with translate only Comprehend and NTM allow for topic modeling (C & E)

Other options Blazingtext is for text classification, not topic modelling, LDA is requires user specified topics and Lex is for conversational interfaces

upvoted 1 times

🗳️ 👤 **CloudHandsOn** 1 year, 5 months ago

Selected Answer: C

C & E

Option C (Amazon Translate and Amazon Comprehend): This is a strong combination. Amazon Translate can be used to translate Spanish comments into English, and then Amazon Comprehend, which supports topic modeling, can be used to identify the most discussed topics.

Option E (Amazon Translate and Amazon SageMaker Neural Topic Model): This is also a viable combination. Amazon Translate would handle the translation of Spanish comments, and the Neural Topic Model (NTM) in Amazon SageMaker can then be used for topic modeling. NTM uses neural networks for topic discovery and is well-suited for analyzing large sets of text data.

upvoted 2 times

🗳️ 👤 **geoan13** 1 year, 7 months ago

B and E

I dont think amazon comprehend can do topic modelling.

LDA is used for topic modelling

upvoted 1 times

🗳️ 👤 **ixdb** 1 year, 9 months ago

BCE are all right. https://docs.amazonaws.cn/en_us/sagemaker/latest/dg/algos.html

LDA and NTM are all topic modeling tools.

upvoted 1 times

🗳️ 👤 **loict** 1 year, 9 months ago

Selected Answer: E

A. NO - BlazingText is word2vec, will not do topic modeling alone

B. NO - Translate better than custom seq2seq

C. NO - NTM better than LDA used by Comprehend

D. NO - Lex is for chatbots

E. YES

upvoted 1 times

🗳️ 👤 **teka112233** 1 year, 9 months ago

Selected Answer: E

The right answers are C & E

The other steps are not suitable because:

A. The BlazingText algorithm is for word embeddings and text classification, not topic modeling.

B. The LDA algorithm is an unsupervised learning algorithm that requires a user-specified number of topics.

D. Amazon Lex is for building conversational interfaces, not extracting topics from content

upvoted 1 times

🗳️ 👤 **Sharath1783** 1 year, 9 months ago

Selected Answer: B

Correct answer is BE

upvoted 3 times

🗨️ 👤 **chet100** 1 year, 10 months ago

It has to be B + C .. for spanish to English use Translate. For Topics it has to be LDA.

upvoted 1 times

🗨️ 👤 **chet100** 1 year, 10 months ago

Sorry and NTM.. in that case, C is a winner for translation.. then pick E to be consistent.. final answer B + E.

upvoted 1 times

🗨️ 👤 **kaike_reis** 1 year, 10 months ago

Selected Answer: E

For me: B - C - E are correct: it's solved translation + topic modelling.

The question is not well construct from my POV.

upvoted 3 times

🗨️ 👤 **Mickey321** 1 year, 11 months ago

Selected Answer: C

C and E

upvoted 1 times

A machine learning (ML) specialist is administering a production Amazon SageMaker endpoint with model monitoring configured. Amazon SageMaker Model

Monitor detects violations on the SageMaker endpoint, so the ML specialist retrains the model with the latest dataset. This dataset is statistically representative of the current production traffic. The ML specialist notices that even after deploying the new SageMaker model and running the first monitoring job, the SageMaker endpoint still has violations.

What should the ML specialist do to resolve the violations?

- A. Manually trigger the monitoring job to re-evaluate the SageMaker endpoint traffic sample.
- B. Run the Model Monitor baseline job again on the new training set. Configure Model Monitor to use the new baseline.
- C. Delete the endpoint and recreate it with the original configuration.
- D. Retrain the model again by using a combination of the original training set and the new training set.

Suggested Answer: B

Community vote distribution

B (89%)

11%

  **edvardo** Highly Voted 2 years, 7 months ago

I would go with B:

<https://docs.aws.amazon.com/sagemaker/latest/dg/model-monitor-create-baseline.html>

upvoted 6 times

  **tgaos** 2 years, 7 months ago

Agree, the answer is B.

From the document, the violation file contains several checks and "The violations file is generated as the output of a MonitoringExecution".

<https://docs.aws.amazon.com/sagemaker/latest/dg/model-monitor-interpreting-violations.html>.

upvoted 3 times

  **AIWave** Most Recent 10 months, 1 week ago

Selected Answer: B

The baseline job computes baseline statistics and constraints for the new training set. By using this updated baseline, Model Monitor can better detect any drift or violations in the production traffic.

upvoted 2 times



  **CloudHandsOn** 11 months, 3 weeks ago

Selected Answer: B

B. Run the Model Monitor baseline job again on the new training set: This is a key step after retraining the model. Since the model has been retrained with a new dataset, the baseline against which its predictions are compared should also be updated. Running the baseline job again on the new training set and configuring Model Monitor to use this new baseline will ensure that the monitoring is relevant to the current state of the model and the data it's processing.

D. Retrain the model again with a combination of the original and new training sets: While retraining the model can be a good approach in some scenarios, there's no indication in this case that the issue lies with the model's performance itself. The issue seems to be with the Model Monitor's baseline not aligning with the current model.

upvoted 3 times

  **sukye** 1 year, 1 month ago

Selected Answer: D

<https://docs.aws.amazon.com/sagemaker/latest/dg/model-monitor-interpreting-violations.html>



upvoted 1 times

  **Mickey321** 1 year, 5 months ago

Selected Answer: B

running the Model Monitor baseline job again on the new training set and configuring Model Monitor to use the new baseline, is the most appropriate step to resolve the violations and ensure the SageMaker endpoint's performance is in line with expectations.

upvoted 1 times

  **Ajose0** 1 year, 10 months ago

Selected Answer: B

Running the Model Monitor baseline job again with the new training set and configuring Model Monitor to use the new baseline is a valid option to resolve the violations.

By running the baseline job with the new training set, a new baseline is created, which can be used to compare with the new data to detect any drifts in the data distribution. Then, the updated baseline can be set as the new baseline for monitoring the endpoint.

So, option B is also a valid solution to resolve the violations.

upvoted 3 times