Question #1 *Topic 1*

Case Study -
A company is building a web-based AI application by using Amazon SageMaker. The application will provide the following capabilities and features: ML experimentation, training, a central model registry, model deployment, and model monitoring.
The application must ensure secure and isolated use of training data during the ML lifecycle. The training data is stored in Amazon S3.
The company needs to use the central model registry to manage different versions of models in the application.
Which action will meet this requirement with the LEAST operational overhead?

A. Create a separate Amazon Elastic Container Registry (Amazon ECR) repository for each model.

B. Use Amazon Elastic Container Registry (Amazon ECR) and unique tags for each model version.

C. Use the SageMaker Model Registry and model groups to catalog the models.

D. Use the SageMaker Model Registry and unique tags for each model version.

Suggested Answer: *C*

Community vote distribution

C (100%)

caseb18249 Highly Voted 1 month ago

Selected Answer: C

The correct answer is C. Use the SageMaker Model Registry and model groups to catalog the models.

SageMaker Model Registry is specifically designed for managing ML models within the SageMaker ecosystem.
Big thanks to Passexamhub! Their MLA-C01 material was the key to my exam success.
It provides built-in versioning and cataloging capabilities for ML models.
Model groups in the Model Registry allow for logical organization of related models.
It integrates seamlessly with other SageMaker components like training, deployment, and monitoring.
This solution offers the least operational overhead as it's a native SageMaker feature designed for this exact purpose.
It provides features like model approval workflows, which are crucial for managing models in production environments.
upvoted 7 times

GiorgioGss Highly Voted 2 months, 1 week ago

Selected Answer: C

https://docs.aws.amazon.com/sagemaker/latest/dg/model-registry-models.html
"Each model package in a Model Group corresponds to a trained model. The version of each model package is a numerical value that starts at 1 and is incremented with each new model package added to a Model Group. For example, if 5 model packages are added to a Model Group, the model package versions will be 1, 2, 3, 4, and 5."
upvoted 5 times

khchan123 Most Recent 1 month ago

Selected Answer: C

The correct answer is C. Use the SageMaker Model Registry and model groups to catalog the models.

SageMaker Model Registry is specifically designed for managing ML models within the SageMaker ecosystem.
It provides built-in versioning and cataloging capabilities for ML models.
Model groups in the Model Registry allow for logical organization of related models.
It integrates seamlessly with other SageMaker components like training, deployment, and monitoring.
This solution offers the least operational overhead as it's a native SageMaker feature designed for this exact purpose.
It provides features like model approval workflows, which are crucial for managing models in production environments.
upvoted 1 times

prabirg 1 month ago

Selected Answer: C

Amazon SageMaker Model Registry creates Catalog models for production and Manage model versions.
upvoted 1 times

S_201996 1 month, 1 week ago

Selected Answer: C

Amazon SageMaker Model Registry is specifically designed to manage and catalog models in a centralized way, including versioning, approval workflows, and deployment history. It simplifies the process of managing different versions of models, which aligns with the company's requirement to use a central model registry.

upvoted 1 times

ninomfr64 1 month, 1 week ago

Selected Answer: C

A. No, ECR is used to store container images

B. No, ECR is used to store container images

C. Yes

D. No, Each model package in a Model Group corresponds to a trained model. The version of each model package is a numerical value that starts at 1 and is incremented with each new model package added to a Model Group - https://docs.aws.amazon.com/sagemaker/latest/dg/model-registry-models.html

upvoted 1 times

motk123 1 month, 3 weeks ago

Selected Answer: C

C: The SageMaker Model Registry organizes models into Model Package Groups, tracks versions as Model Packages, and optionally aggregates groups into Collections. This structure ensures robust versioning and manageability for trained models.
https://docs.aws.amazon.com/sagemaker/latest/dg/model-registry-models.html

upvoted 2 times

Neo_2022 2 months, 1 week ago

Selected Answer: C

https://aws.amazon.com/blogs/machine-learning/centralize-model-governance-with-sagemaker-model-registry-resource-access-manager-sharing/

upvoted 4 times

Question #2 *Topic 1*

Case Study -
A company is building a web-based AI application by using Amazon SageMaker. The application will provide the following capabilities and features: ML experimentation, training, a central model registry, model deployment, and model monitoring.
The application must ensure secure and isolated use of training data during the ML lifecycle. The training data is stored in Amazon S3.
The company is experimenting with consecutive training jobs.
How can the company MINIMIZE infrastructure startup times for these jobs?

A. Use Managed Spot Training.

B. Use SageMaker managed warm pools.

C. Use SageMaker Training Compiler.

D. Use the SageMaker distributed data parallelism (SMDDP) library.

Suggested Answer: *B*

Community vote distribution

B (100%)

S_201996 1 month, 1 week ago

Selected Answer: B

SageMaker managed warm pools allow instances to stay in a ready state between consecutive training jobs, which minimizes infrastructure startup times. This feature is ideal for scenarios with frequent or consecutive training jobs, as it avoids the time-consuming process of provisioning infrastructure for each job.

upvoted 1 times

ninomfr64 1 month, 1 week ago

Selected Answer: B

A. No, Managed Spot Training is used to reduce the compute cost for training
B. Yes, Warm Pool allow to retain and re-use provisioned infrastructure, also use a persistent cache to store data across training job and help reduce infrastructure startup time as well as cost - https://docs.aws.amazon.com/sagemaker/latest/dg/train-warm-pools.html
C. No, SageMaker Training Compiler is used to optimize your code for a specific target architecture
D. No, the SageMaker distributed data parallelism (SMDDP) library is used parallelize training by distributing data across multiple instances. This doesn't reduce infrastructure startu ptime

upvoted 1 times

andy_10 2 months ago

Selected Answer: B

https://docs.aws.amazon.com/sagemaker/latest/dg/train-warm-pools.html#train-warm-pools-how-it-works
SageMaker managed warm pools let you retain and reuse provisioned infrastructure after the completion of a training job to reduce latency for repetitive workloads, such as iterative experimentation or running many jobs consecutively.

upvoted 4 times

Neo_2022 2 months, 1 week ago

Selected Answer: B

https://aws.amazon.com/about-aws/whats-new/2022/09/reduce-ml-model-training-job-startup-time-8x-sagemaker-training-managed-warm-pools/

upvoted 2 times

tigrex73 2 months, 1 week ago

Selected Answer: B

SageMaker managed warm pools are designed to reduce infrastructure startup times by keeping the training environment (instances, containers, and environment setup) ready between consecutive training jobs.

upvoted 2 times

GiorgioGss 2 months, 1 week ago

Selected Answer: B

https://docs.aws.amazon.com/sagemaker/latest/dg/train-warm-pools.html
"which speeds up start times by reducing the time spent provisioning resources."

upvoted 2 times

Question #3 *Topic 1*

Case Study -
A company is building a web-based AI application by using Amazon SageMaker. The application will provide the following capabilities and features: ML experimentation, training, a central model registry, model deployment, and model monitoring.
The application must ensure secure and isolated use of training data during the ML lifecycle. The training data is stored in Amazon S3.
The company must implement a manual approval-based workflow to ensure that only approved models can be deployed to production endpoints.
Which solution will meet this requirement?

A. Use SageMaker Experiments to facilitate the approval process during model registration.

B. Use SageMaker ML Lineage Tracking on the central model registry. Create tracking entities for the approval process.

C. Use SageMaker Model Monitor to evaluate the performance of the model and to manage the approval.

D. Use SageMaker Pipelines. When a model version is registered, use the AWS SDK to change the approval status to "Approved."

Suggested Answer: *D*

Community vote distribution

S_201996 1 month, 1 week ago

Selected Answer: D

SageMaker Pipelines is designed to orchestrate machine learning workflows, including manual approval steps for model registration. You can define a step in the pipeline where a manual approval process is required before the model's status is changed to "Approved" for deployment.

upvoted 1 times

ninomfr64 1 month, 1 week ago

Selected Answer: D

This tricked my as option D is not clearly worded:
A. No, SageMaker Experiments allows to track and organize your experiment but not for approving models
B. No, SageMaker ML Lineage Tracking allows to track model lineage but do not allow to approve a model
C. No, SageMaker Model Monitor allows to monitor data quality, model quality, bias and feature attribution
D. Yes, After you create a model version, you typically evaluate its performance and then update the approval status of the model version. You can update the approval status of a model version by using the SDK, SageMaker Studio console or with a condition step in a SageMaker AI pipeline

upvoted 1 times

tigrex73 2 months, 1 week ago

Selected Answer: D

The SageMaker Model Registry within the pipeline provides functionality to manually or programmatically approve models for production deployment.

upvoted 3 times

GiorgioGss 2 months, 1 week ago

Selected Answer: D

https://docs.aws.amazon.com/en_us/sagemaker/latest/dg/model-registry-approve.html
"You can update the approval status of a model version by using the AWS SDK "

upvoted 3 times

Question #4 *Topic 1*

Case Study -
A company is building a web-based AI application by using Amazon SageMaker. The application will provide the following capabilities and features: ML experimentation, training, a central model registry, model deployment, and model monitoring.
The application must ensure secure and isolated use of training data during the ML lifecycle. The training data is stored in Amazon S3.
The company needs to run an on-demand workflow to monitor bias drift for models that are deployed to real-time endpoints from the application.
Which action will meet this requirement?

A. Configure the application to invoke an AWS Lambda function that runs a SageMaker Clarify job.

B. Invoke an AWS Lambda function to pull the sagemaker-model-monitor-analyzer built-in SageMaker image.

C. Use AWS Glue Data Quality to monitor bias.

D. Use SageMaker notebooks to compare the bias.

Suggested Answer: *A*

Community vote distribution

A (100%)

S_201996 1 month, 1 week ago

Selected Answer: A

SageMaker Clarify can be used to analyze bias drift in models. By integrating this with a Lambda function, the workflow can be triggered on-demand whenever the application requires bias monitoring.

upvoted 1 times

ninomfr64 1 month, 1 week ago

Selected Answer: A

A. Yes, Clarify allows to get bias - https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-configure-processing-jobs.html
B. No, the built-in image sagemaker-model-monitor-analyzer provides a range of model monitoring capabilities (constraint suggestion, statistics generation, constraint validation against a baseline, and emitting Amazon CloudWatch metrics) but you need Clarify for bias
C. No, Glue Data Quality doesn't analyze bias
D. No, well from a Notebook you can execute pretty much everything including a Clarify Job, however notebooks are for experiments and models development not for enabling real-time application features

upvoted 3 times

tigrex73 2 months, 1 week ago

Selected Answer: A

SageMaker Clarify is a tool designed to detect and monitor bias in datasets and models. It provides built-in capabilities for bias analysis, both pre-training (data bias) and post-training (model bias). Using AWS Lambda to invoke the job ensures automation and on-demand execution, reducing operational complexity while meeting the requirement for monitoring bias drift.

upvoted 2 times

GiorgioGss 2 months, 1 week ago

Selected Answer: A

https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-measure-data-bias.html

upvoted 2 times

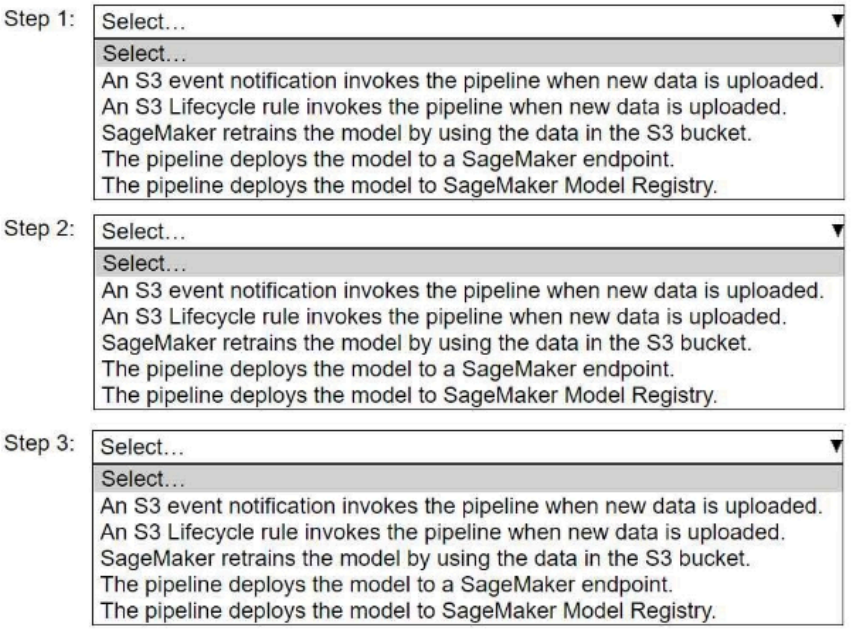
Question #7 *Topic 1*

HOTSPOT -

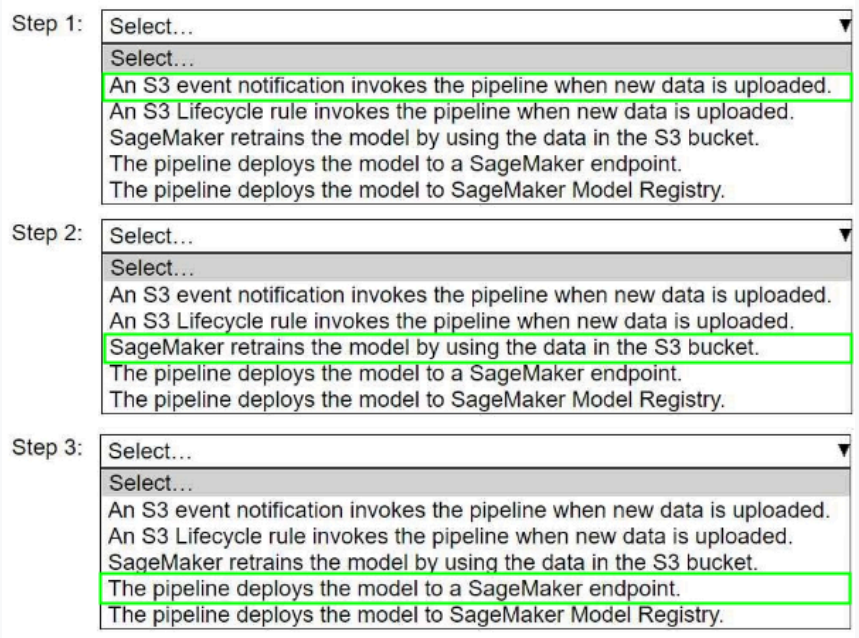
A company wants to host an ML model on Amazon SageMaker. An ML engineer is configuring a continuous integration and continuous delivery (CI/CD) pipeline in AWS CodePipeline to deploy the model. The pipeline must run automatically when new training data for the model is uploaded to an Amazon S3 bucket.

Select and order the pipeline's correct steps from the following list. Each step should be selected one time or not at all. (Select and order three.)

- An S3 event notification invokes the pipeline when new data is uploaded.
- S3 Lifecycle rule invokes the pipeline when new data is uploaded.
- SageMaker retrains the model by using the data in the S3 bucket.
- The pipeline deploys the model to a SageMaker endpoint.
- The pipeline deploys the model to SageMaker Model Registry.



Suggested Answer:



0c2d840 2 weeks, 4 days ago

First two steps are obvious. For the last (third) step, there are two choices.

1. The pipeline deploys the model to a SageMaker endpoint.
2. The pipeline deploys the model to SageMaker Model Registry.

Since the question says deploy the model, 1st option is correct. If we add the model to Model Registry, it will be just there in the catalog, but won't get deployed. It needs to be explicitly deployed to the endpoint. So 2 is the correct third step.

upvoted 2 times

Question #9 *Topic 1*

HOTSPOT -

An ML engineer is working on an ML model to predict the prices of similarly sized homes. The model will base predictions on several features The ML engineer will use the following feature engineering techniques to estimate the prices of the homes:

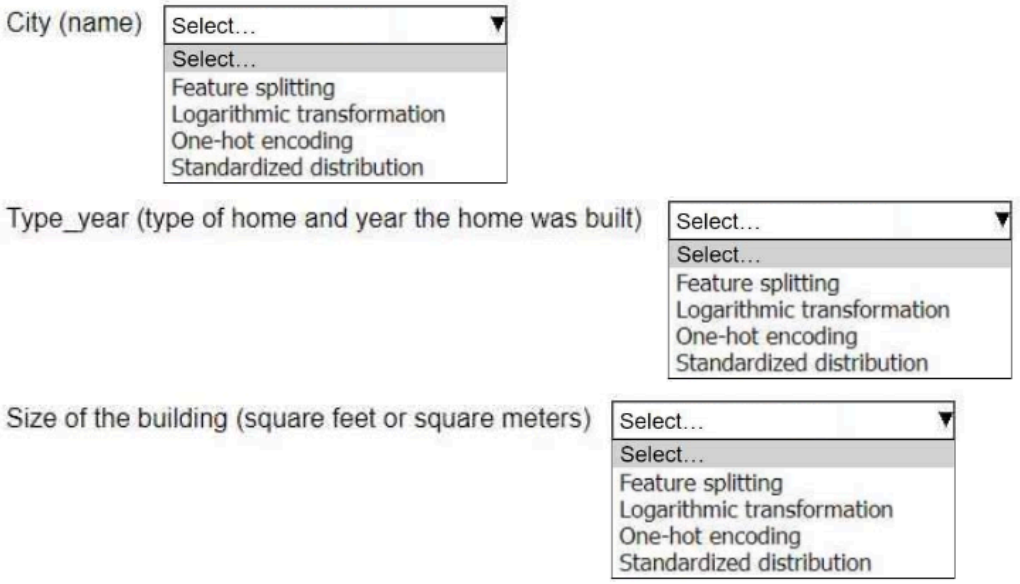
• Feature splitting

• Logarithmic transformation

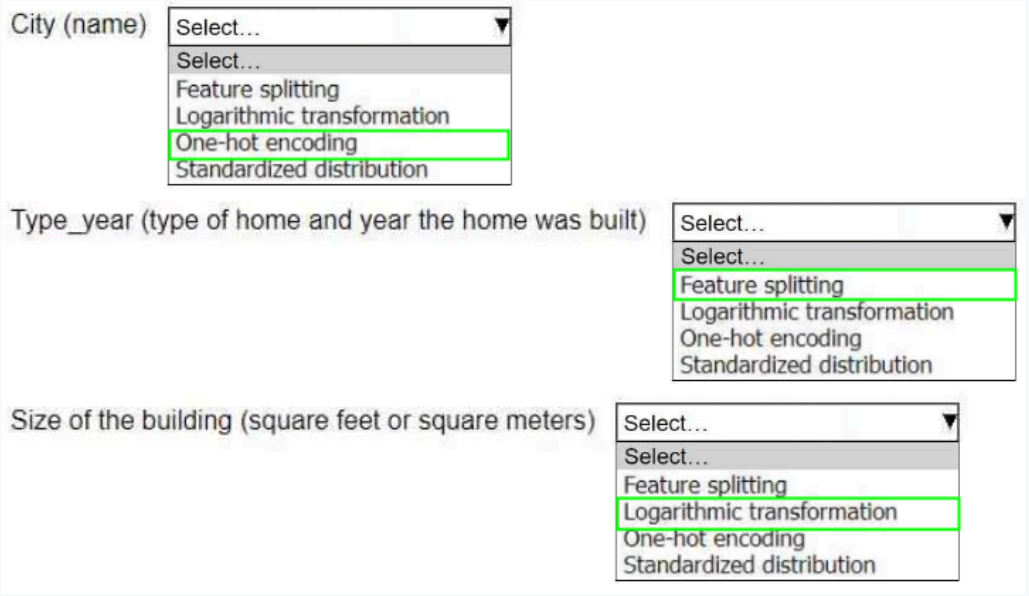
• One-hot encoding

• Standardized distribution

Select the correct feature engineering techniques for the following list of features. Each feature engineering technique should be selected one time or not at all (Select three.)



Suggested Answer:



abrarjahin 5 days, 15 hours ago

Size of the building is standard distribution

upvoted 1 times

Question #10 *Topic 1*

Case study -
An ML engineer is developing a fraud detection model on AWS. The training dataset includes transaction logs, customer profiles, and tables from an on-premises MySQL database. The transaction logs and customer profiles are stored in Amazon S3.
The dataset has a class imbalance that affects the learning of the model's algorithm. Additionally, many of the features have interdependencies. The algorithm is not capturing all the desired underlying patterns in the data.
Which AWS service or feature can aggregate the data from the various data sources?

A. Amazon EMR Spark jobs

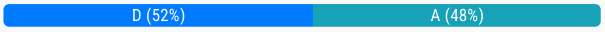
B. Amazon Kinesis Data Streams

C. Amazon DynamoDB

D. AWS Lake Formation

Suggested Answer: *D*

Community vote distribution



tigrex73 Highly Voted 2 months, 1 week ago

Selected Answer: A

Amazon EMR with Spark is an excellent choice for aggregating, processing, and transforming large datasets from multiple sources (e.g., Amazon S3 and on-premises MySQL database). Spark jobs can handle both structured and unstructured. While Lake Formation is great for managing data lakes, it doesn't provide the ETL and data processing capabilities required to aggregate and transform datasets from multiple sources.

upvoted 9 times

dbcert87 Most Recent 1 week, 1 day ago

Selected Answer: D

AWS Lake Formation is a service designed to aggregate, catalog, and manage data from multiple data sources, including on-premises databases and Amazon S3, making it an ideal choice for this scenario.

While Amazon EMR with Apache Spark is powerful for processing and analyzing large datasets, it focuses more on data processing than on data aggregation and cataloging. It doesn't inherently manage interdependencies or schema enforcement

upvoted 1 times

dbcert87 1 week, 1 day ago

Selected Answer: A

Amazon EMR is correct answer for aggregation

upvoted 1 times

fnuuu 1 week, 4 days ago

Selected Answer: D

Data Lake is used for data discovery

upvoted 1 times

xukun 3 weeks ago

Selected Answer: D

Once you specify where your existing databases are and provide your access credentials, Lake Formation reads the data and its metadata (schema) to understand the contents of the data source. It then imports the data to your new data lake and records the metadata in a central catalog. With Lake Formation, you can import data from MySQL, PostgreSQL, SQL Server, MariaDB, and Oracle databases running in Amazon RDS or hosted in Amazon EC2. Both bulk and incremental data loading are supported.
https://docs.aws.amazon.com/lake-formation/latest/dg/what-is-lake-formation.html

upvoted 1 times

Makendran 3 weeks, 6 days ago

Selected Answer: A

While AWS Lake Formation could potentially be used in conjunction with other services for data lake management, Amazon EMR with Spark jobs is the most direct and powerful solution for aggregating and processing data from the various sources mentioned in this scenario. It provides the

necessary tools to handle the data integration, address the class imbalance, and perform the complex feature engineering that may be required for the fraud detection model.

upvoted 1 times

CloudHandsOn 1 month ago

Selected Answer: D

My first choice was Lake Formation

upvoted 1 times

ninomfr64 1 month ago

Selected Answer: D

Yet another poorly worded AWS certification question. Here is my reasoning, the question is about "aggregate the data from S3 and on-premise mysql" and I do intend "aggregate" as put in the same place, therefore:

A. No, while EMR spark job can connect to S3 and MySQL (spark can connect to mysql database), but it is a better tool to process data and then sore them in S3

B. No, KDS it is for delivering streaming data sources to specific destinations (S3, OpenSearch ...)

C. No, DynamoDB is a nosql db that is not a great fit here

D. Yes, Lake Formation "combine different types of structured and unstructured data into a centralized repository" https://docs.aws.amazon.com/lake-formation/latest/dg/what-is-lake-formation.html and "with Lake Formation, you can import your data using workflows" and as it is based on AWS Glue it supports both S3 and mysql

upvoted 2 times

Asankalshara 1 month, 1 week ago

Selected Answer: D

Question is which AWS service or feature can aggregate the data from the various data sources? So lake formation

upvoted 1 times

breathingcloud 1 month, 1 week ago

Selected Answer: A

I think it is A, it is more aligned with machine learning model

upvoted 1 times

AbhayD 1 month, 2 weeks ago

Selected Answer: A

Lake formation can catalog data from various sources, it doesn't provide the data processing capabilities needed for this scenario. EMR is more appropriate in this case.

upvoted 1 times

TonyKean888 1 month, 3 weeks ago

Selected Answer: D

Data Aggregation: Lake Formation is designed to create a data lake, a centralized repository that stores and manages data from various sources, including S3, relational databases (like MySQL), and other data sources.

Data Transformation: It can transform and clean data, making it suitable for analysis and machine learning. This includes handling class imbalance and feature interdependencies.

Data Access: It provides a unified interface to access data, simplifying the process of integrating data from different sources into the ML model.

While other options like Amazon EMR Spark jobs and Amazon Kinesis Data Streams could be used for data processing and streaming, they are not the most efficient and straightforward solutions for this specific use case. Amazon DynamoDB is a NoSQL database, not designed for batch data processing and aggregation.

Therefore, AWS Lake Formation is the best choice to aggregate and prepare the data for the ML model.
ref:https://docs.aws.amazon.com/lake-formation/latest/dg/what-is-lake-formation.html

upvoted 2 times

LR2023 1 month, 4 weeks ago

Selected Answer: D

Lake formation would be a better choice over EMR as it involes complexity setting up . For data aggregation and ETL processes, especially involving multiple data sources and ensuring data quality and security, AWS Lake Formation or Amazon Glue are more specialized and suitable option

upvoted 2 times

GiorgioGss 2 months, 1 week ago

Selected Answer: A

I would go with EMR Spark jobs just because I think Lake Formation is not designed for feature engineering. Spark is.

upvoted 1 times

a4002bd 2 months, 1 week ago

Selected Answer: D

Is it D? AWS Lake Formation ? EMR Spark jobs is more manual.

upvoted 4 times

Question #11 *Topic 1*

Case study -
An ML engineer is developing a fraud detection model on AWS. The training dataset includes transaction logs, customer profiles, and tables from an on-premises MySQL database. The transaction logs and customer profiles are stored in Amazon S3.
The dataset has a class imbalance that affects the learning of the model's algorithm. Additionally, many of the features have interdependencies. The algorithm is not capturing all the desired underlying patterns in the data.
After the data is aggregated, the ML engineer must implement a solution to automatically detect anomalies in the data and to visualize the result.
Which solution will meet these requirements?

A. Use Amazon Athena to automatically detect the anomalies and to visualize the result.

B. Use Amazon Redshift Spectrum to automatically detect the anomalies. Use Amazon QuickSight to visualize the result.

C. Use Amazon SageMaker Data Wrangler to automatically detect the anomalies and to visualize the result.

D. Use AWS Batch to automatically detect the anomalies. Use Amazon QuickSight to visualize the result.

Suggested Answer: *C*

Community vote distribution

C (100%)

ninomfr64 1 month ago

Selected Answer: C

SageMaker Data Wangler identify anomalies as part of the Data Quality and Insights Report (https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler-data-insights.html) and provides various options for data visualization - https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler-analyses.html

upvoted 1 times

motk123 1 month, 3 weeks ago

Selected Answer: C

Why Transform Categorical Data into Numerical Data?
Machine learning algorithms generally require categorical data to be converted into numerical representations (e.g., one-hot encoding or embeddings) for training. Transforming numerical data into categorical data is unnecessary unless the problem explicitly requires it (e.g., binning for some specific applications).

Why Use SageMaker Data Wrangler?
Minimal Operational Overhead: Amazon SageMaker Data Wrangler provides a user-friendly interface to clean, preprocess, and transform data without needing to write custom code.
Comprehensive Data Handling: Supports data sources like S3 and on-premises databases, and can handle both categorical and numerical data transformations efficiently.

Why Not AWS Glue?
AWS Glue is more suitable for large-scale ETL (Extract, Transform, Load) operations, such as schema inference or combining large datasets. It has higher operational overhead for specific ML data preprocessing tasks compared to SageMaker Data Wrangler.

upvoted 1 times

GiorgioGss 2 months, 1 week ago

Selected Answer: C

https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler-analyses.html
"Amazon SageMaker Data Wrangler includes built-in analyses that help you generate visualizations and data analyses in a few clicks. "
This question is tricky because it makes you think you need Quicksight for the "visualization' part.

upvoted 3 times

Question #12 *Topic 1*

Case study -
An ML engineer is developing a fraud detection model on AWS. The training dataset includes transaction logs, customer profiles, and tables from an on-premises MySQL database. The transaction logs and customer profiles are stored in Amazon S3.
The dataset has a class imbalance that affects the learning of the model's algorithm. Additionally, many of the features have interdependencies. The algorithm is not capturing all the desired underlying patterns in the data.
The training dataset includes categorical data and numerical data. The ML engineer must prepare the training dataset to maximize the accuracy of the model.
Which action will meet this requirement with the LEAST operational overhead?

A. Use AWS Glue to transform the categorical data into numerical data.

B. Use AWS Glue to transform the numerical data into categorical data.

C. Use Amazon SageMaker Data Wrangler to transform the categorical data into numerical data.

D. Use Amazon SageMaker Data Wrangler to transform the numerical data into categorical data.

Suggested Answer: *C*

Community vote distribution

C (100%)

ninomfr64 1 month ago

Selected Answer: C

You need to transform category to numeric as ML model works with numbers, thus it is either A or C. Data Wrangler provides a builtin transformation to encode categorical data - https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler-transform.html#data-wrangler-transform-cat-encode while Glue doesn't provide a managed transformation for encoding data - https://docs.aws.amazon.com/glue/latest/dg/edit-jobs-transforms.html

upvoted 1 times

Pofmagic 1 month, 1 week ago

Selected Answer: C

Data Wrangler can be used for encoding categorical data, i.e. the process of creating a numerical representation for categories. Categorical encoding encodes categorical data that is in string format into arrays of integers. Data Wrangler supports ordinal and a one-hot encoding, also similarity encoding (more advanced).
https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler-transform.html#data-wrangler-transform-cat-encode

AWS Glue also has Data science recipe steps for One Hot Encoding and Categorical Mapping.
https://docs.aws.amazon.com/databrew/latest/dg/recipe-actions.data-science.html

However Data Wrangler is more user-friendly with visual and natural language interfaces for less operational overhead

upvoted 1 times

GiorgioGss 2 months, 1 week ago

Selected Answer: C

https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler-transform.html

upvoted 3 times

Question #13 *Topic 1*

Case study -
An ML engineer is developing a fraud detection model on AWS. The training dataset includes transaction logs, customer profiles, and tables from an on-premises MySQL database. The transaction logs and customer profiles are stored in Amazon S3.
The dataset has a class imbalance that affects the learning of the model's algorithm. Additionally, many of the features have interdependencies. The algorithm is not capturing all the desired underlying patterns in the data.
Before the ML engineer trains the model, the ML engineer must resolve the issue of the imbalanced data.
Which solution will meet this requirement with the LEAST operational effort?

A. Use Amazon Athena to identify patterns that contribute to the imbalance. Adjust the dataset accordingly.

B. Use Amazon SageMaker Studio Classic built-in algorithms to process the imbalanced dataset.

C. Use AWS Glue DataBrew built-in features to oversample the minority class.

D. Use the Amazon SageMaker Data Wrangler balance data operation to oversample the minority class.

Suggested Answer: *D*

Community vote distribution

ninomfr64 1 month ago

Selected Answer: D

Both Glue DataBrew and Data Wrangler allows data preparation for ML with no-code/low-code (aka low ops effort). However, Data Wrangler provides built-in transformation for balancing dataset (random oversampling, random undersampling and smote)
https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler-transform.html#data-wrangler-transform-balance-data while DataBrew doesn't provide built-in recipe step for balancing dataset, actually it provides a smaller set of data science recipe steps limited to binarization, bucketization, categorical mapping, one-hot encoding, scaling, skewness and tokenization
https://docs.aws.amazon.com/databrew/latest/dg/recipe-actions.data-science.html

upvoted 1 times

GiorgioGss 2 months, 1 week ago

Selected Answer: D

LEAST effort
https://aws.amazon.com/blogs/machine-learning/balance-your-data-for-machine-learning-with-amazon-sagemaker-data-wrangler/

upvoted 4 times

Question #14 *Topic 1*

Case study -
An ML engineer is developing a fraud detection model on AWS. The training dataset includes transaction logs, customer profiles, and tables from an on-premises MySQL database. The transaction logs and customer profiles are stored in Amazon S3.
The dataset has a class imbalance that affects the learning of the model's algorithm. Additionally, many of the features have interdependencies. The algorithm is not capturing all the desired underlying patterns in the data.
The ML engineer needs to use an Amazon SageMaker built-in algorithm to train the model.
Which algorithm should the ML engineer use to meet this requirement?

A. LightGBM

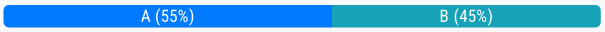
B. Linear learner

C. K-means clustering

D. Neural Topic Model (NTM)

Suggested Answer: *A*

Community vote distribution



aragon_saa Highly Voted 2 months, 1 week ago

Selected Answer: B

Answer is B

upvoted 6 times

Leo2023aws Highly Voted 2 months, 1 week ago

Selected Answer: A

https://docs.aws.amazon.com/en_kr/sagemaker/latest/dg/lightgbm.html

upvoted 5 times

Jacobog3 Most Recent 5 days, 6 hours ago

Selected Answer: A

Is supported by Sagemaker

upvoted 1 times

abrarjahin 1 week, 6 days ago

Selected Answer: B

Linear Learner is a built-in algorithm provided by SageMaker for supervised learning tasks like regression and classification. LightGBM is not a built-in algorithm in Amazon SageMaker. While it is a strong gradient-boosting algorithm, it would need to be implemented as a custom script in SageMaker, which increases operational overhead.

upvoted 1 times

ninomfr64 5 days, 18 hours ago

LightGBM is built-in https://docs.aws.amazon.com/sagemaker/latest/dg/lightgbm.html

upvoted 1 times

xukun 3 weeks ago

Selected Answer: A

https://docs.aws.amazon.com/en_kr/sagemaker/latest/dg/lightgbm.html

upvoted 1 times

Makendran 3 weeks, 6 days ago

Selected Answer: B

In an ideal scenario, for a problem with these characteristics (fraud detection, class imbalance, feature interdependencies, complex patterns), a tree-based ensemble method like XGBoost (which is a SageMaker built-in algorithm) would be more suitable. XGBoost can handle non-linear relationships, is robust to class imbalance with proper tuning, and can capture complex patterns in the data.

However, given the options provided and the requirement to use a SageMaker built-in algorithm, the Linear learner is the best available choice among these options for this specific fraud detection task.

upvoted 1 times

ninomfr64 5 days, 18 hours ago

LightGBM is better for this use case https://docs.aws.amazon.com/sagemaker/latest/dg/lightgbm.html

upvoted 1 times

gulf1324 1 month ago

Selected Answer: B

A. Light BGM : It's suitable model, but not built-in model for SageMaker.

Answer B. Linear learner : suitable model, built-in model for SageMaker.

C. K-means clustering : groups similar data points, not suitable for classification problems, and it's unsupervised learning algorithm so doesn't fit in this case(fraud detection).

D. Neural Topic Model: used for topic modeling and document classification, not suitable for fraud detection

upvoted 2 times

minhhnh 3 weeks, 4 days ago

Light BGM is built-in model for SageMaker

https://docs.aws.amazon.com/sagemaker/latest/dg/lightgbm.html

upvoted 1 times

khchan123 1 month ago

Selected Answer: A

Here's why LightGBM is the most suitable algorithm for this fraud detection task:

Handling Class Imbalance: LightGBM is particularly effective at handling imbalanced datasets, which is a key issue mentioned in the problem statement. It has built-in mechanisms to deal with class imbalance.

Feature Interdependencies: LightGBM can capture complex feature interactions through its tree-based structure, addressing the issue of feature interdependencies mentioned in the problem.

Capturing Underlying Patterns: As an advanced gradient boosting framework, LightGBM is excellent at capturing complex patterns in data, which the current algorithm is struggling with.

Suitable for Fraud Detection: LightGBM is widely used in fraud detection tasks due to its high performance and ability to handle large datasets efficiently.

Handling Various Data Types: It can work well with the mix of data types likely present in transaction logs, customer profiles, and database tables.

upvoted 1 times

ninomfr64 1 month ago

Selected Answer: A

We have an unbalanced dataset, this means we have labelled dataset thus we are going to use a supervised model training. This reduce options to A and B (K-means and NTM are unsupervised). Both LightGBM and Linear Learner provides hyperparameter to manage unbalanced datasets, respectively "scale-pos_weight" and "positive_example_weight_mult". I would go for LightGBM as this algorithm is more suited to handle complex relationship among features, while Linear Learner learns a linear function, or, for classification problems, a linear threshold function, and maps a vector x to an approximation of the label y.

upvoted 1 times

Ell89 1 month ago

Selected Answer: B

Linear Learner. LightGBM is NOT a built in algorithm which the question asks for.

upvoted 1 times

michaelcloud 1 month ago

Selected Answer: A

This is a binary classification problem so LightGBM so be used. Other algorithms are not for binary classification.

upvoted 1 times

bakju0 1 month ago

Selected Answer: B

Is LightGBM Built-in?

Technically, no, LightGBM is not a “built-in algorithm” in the same category as SageMaker’s core algorithms (like Linear Learner or XGBoost). However, it is supported through prebuilt containers, which makes it easy to use in SageMaker.

upvoted 2 times

Saransundar 2 months ago

Selected Answer: A

A. LightGBM: Handles class imbalance; captures feature interdependencies; models complex patterns.

B. Linear Learner: Limited with interdependent features; struggles with complex patterns; suitable for linear relationships.

C. K-means Clustering: Unsupervised algorithm; not suitable for classification; can't handle class imbalance.

D. Neural Topic Model (NTM): Designed for topic modeling; unsuitable for fraud detection; doesn't address class imbalance.

upvoted 2 times

Linux_master 2 months ago

Selected Answer: A

This is a binary classification problem so LightGBM so be used. Other algorithms are not for binary classification.

upvoted 3 times

GiorgioGss 2 months ago

Selected Answer: A

Light Gradient Boosting Machine is effective for handling class imbalances and feature interdependencies.

upvoted 1 times

Question #15 *Topic 1*

A company has deployed an XGBoost prediction model in production to predict if a customer is likely to cancel a subscription. The company uses Amazon SageMaker Model Monitor to detect deviations in the F1 score.
During a baseline analysis of model quality, the company recorded a threshold for the F1 score. After several months of no change, the model's F1 score decreases significantly.
What could be the reason for the reduced F1 score?

A. Concept drift occurred in the underlying customer data that was used for predictions.

B. The model was not sufficiently complex to capture all the patterns in the original baseline data.

C. The original baseline data had a data quality issue of missing values.

D. Incorrect ground truth labels were provided to Model Monitor during the calculation of the baseline.

Suggested Answer: *A*

Community vote distribution

A (100%)

ninomfr64 1 month ago

Selected Answer: A

A. Yes, concept drift is an evolution of data that invalidates the data model. It happens when the statistical properties of the target variable, which the model is trying to predict, change over time in unforeseen ways. This causes problems because the predictions become less accurate as time passes.
B. No, if it was the case the F1 would have been low since the begin and this is not justifying a change after months
C. No, same as B
D. No, incorrect labels in the baseline calculations would undermine F1 baseline value, but this is not explain a significant drop after months
upvoted 1 times

motk123 1 month, 3 weeks ago

Selected Answer: A

Concept Drift: Occurs when the statistical properties of the data used for predictions change over time, causing the model to underperform on current data.

Why Not the Other Options?

B. If the model complexity was insufficient, the issue would have been detected during the initial evaluation or baseline analysis, not after months of stable performance.
C. A data quality issue would have impacted the model's performance immediately after deployment, not months later.
D. Incorrect labels during baseline calculation could result in an inaccurate baseline F1 score, but it wouldn't explain a significant drop after stable performance over months.
upvoted 3 times

Saransundar 2 months ago

Selected Answer: A

Concept Drift: Refers to the change in the statistical properties of the underlying data distribution over time --> Decrease F1 score --> perform poorly on new data
upvoted 3 times

GiorgioGss 2 months, 1 week ago

Selected Answer: A

Option A could be the only one possible reason for drifting "after several months".
upvoted 2 times

Question #16 *Topic 1*

A company has a team of data scientists who use Amazon SageMaker notebook instances to test ML models. When the data scientists need new permissions, the company attaches the permissions to each individual role that was created during the creation of the SageMaker notebook instance.
The company needs to centralize management of the team's permissions.
Which solution will meet this requirement?

A. Create a single IAM role that has the necessary permissions. Attach the role to each notebook instance that the team uses.

B. Create a single IAM group. Add the data scientists to the group. Associate the group with each notebook instance that the team uses.

C. Create a single IAM user. Attach the AdministratorAccess AWS managed IAM policy to the user. Configure each notebook instance to use the IAM user.

D. Create a single IAM group. Add the data scientists to the group. Create an IAM role. Attach the AdministratorAccess AWS managed IAM policy to the role. Associate the role with the group. Associate the group with each notebook instance that the team uses.

Suggested Answer: *A*

Community vote distribution

ninomfr64 1 month ago

Selected Answer: A

Yet another unclear question from AWS ... anyway, I am basically picking A as all the other options are not applicable or are unclear.

A. Yes, this make sense
B. No, you cannot assign (aka associate) group to notebook instances
C. No, for two reason: AdministratorAccess policy is overly broad (violate least privilege principle) and you cannot assign IAM user to notebook instance
D. No, for many reasons: AdministratorAccess policy is overly broad, not clear what associating a role to a group means (maybe a group has permissions to assume a role ...) and you cannot assign a group to a notebook

upvoted 1 times

motk123 1 month, 3 weeks ago

Selected Answer: A

Creating a single IAM role allows centralized management of permissions for all SageMaker notebook instances. When permissions need to be updated, the changes are applied to the role, and all notebook instances automatically inherit the updated permissions.
Why IAM Roles?
IAM roles are the recommended way to provide permissions to AWS services like SageMaker because they securely delegate permissions without requiring long-term credentials.

Why Not the Other Options?
B. IAM groups manage permissions for users, not for AWS services or resources like SageMaker notebook instances. Groups cannot be attached directly to notebook instances.
C. Using an IAM user with AdministratorAccess violates the principle of least privilege, granting unnecessary permissions. Additionally, IAM users are not intended to be attached to resources like notebook instances.
D. This option combines unnecessary complexity (group and role association) and grants excessive permissions (AdministratorAccess), which is not secure or efficient.

upvoted 3 times

GiorgioGss 2 months, 1 week ago

Selected Answer: A

Actually this is a best practice when working with notebooks in SageMaker.
https://docs.aws.amazon.com/sagemaker/latest/dg/gs-setup-working-env.html

upvoted 3 times

Question #17 *Topic 1*

An ML engineer needs to use an ML model to predict the price of apartments in a specific location.
Which metric should the ML engineer use to evaluate the model's performance?

A. Accuracy

B. Area Under the ROC Curve (AUC)

C. F1 score

D. Mean absolute error (MAE)

Suggested Answer: *D*

Community vote distribution

D (100%)

GiorgioGss Highly Voted 2 months, 1 week ago

Selected Answer: D

The only one for regression is D. Other 3 are for classification.

upvoted 5 times

ninomfr64 Most Recent 3 weeks, 3 days ago

Selected Answer: D

This is a regression problem, thus MAE is the right answer. Accuracy, AUC-ROC and F1 are for classification.

upvoted 1 times

Question #18 *Topic 1*

An ML engineer has trained a neural network by using stochastic gradient descent (SGD). The neural network performs poorly on the test set. The values for training loss and validation loss remain high and show an oscillating pattern. The values decrease for a few epochs and then increase for a few epochs before repeating the same cycle.
What should the ML engineer do to improve the training process?

A. Introduce early stopping.

B. Increase the size of the test set.

C. Increase the learning rate.

D. Decrease the learning rate.

Suggested Answer: *D*

Community vote distribution

D (100%)

ninomfr64 3 weeks, 3 days ago

Selected Answer: D

A. No, early stopping is for preventing overfitting
B. No, increasing test will not help with oscillating loss
C. No, increasing learning rate will make things worsening
D. Oscillating loss in training is a sign that the training is not converging, this can happen when learning rate is too high. Reducing learning rate will help here

upvoted 3 times

gulf1324 1 month ago

Selected Answer: D

Oscillating patterns in a train/validation loss shows it's not converging to a minima. Low learning rate will make it converge.

upvoted 1 times

feelgoodfactor 1 month, 2 weeks ago

Selected Answer: D

The oscillating loss during training is a clear sign that the learning rate is too high. Reducing the learning rate will stabilize the optimization process, allowing the model to converge smoothly.

upvoted 2 times

motk123 1 month, 3 weeks ago

Selected Answer: D

The oscillating pattern of training and validation loss indicates that the learning rate is too high. A high learning rate causes the model to overshoot the optimal point in the loss landscape, leading to oscillations instead of convergence. Reducing the learning rate allows the model to make smaller, more precise updates to the weights, improving convergence.

A. Early stopping prevents overfitting by halting training when validation performance stops improving. However, it does not address the root cause of oscillating loss.
B. The size of the test set does not affect the training dynamics or loss patterns.
C. Increasing the learning rate would worsen the oscillations and prevent the model from converging.

upvoted 4 times

GiorgioGss 2 months ago

Selected Answer: D

oscillating = decrease the learning rate

upvoted 2 times

Question #19 *Topic 1*

An ML engineer needs to process thousands of existing CSV objects and new CSV objects that are uploaded. The CSV objects are stored in a central Amazon S3 bucket and have the same number of columns. One of the columns is a transaction date. The ML engineer must query the data based on the transaction date.
Which solution will meet these requirements with the LEAST operational overhead?

A. Use an Amazon Athena CREATE TABLE AS SELECT (CTAS) statement to create a table based on the transaction date from data in the central S3 bucket. Query the objects from the table.

B. Create a new S3 bucket for processed data. Set up S3 replication from the central S3 bucket to the new S3 bucket. Use S3 Object Lambda to query the objects based on transaction date.

C. Create a new S3 bucket for processed data. Use AWS Glue for Apache Spark to create a job to query the CSV objects based on transaction date. Configure the job to store the results in the new S3 bucket. Query the objects from the new S3 bucket.

D. Create a new S3 bucket for processed data. Use Amazon Data Firehose to transfer the data from the central S3 bucket to the new S3 bucket. Configure Firehose to run an AWS Lambda function to query the data based on transaction date.

Suggested Answer: *A*

Community vote distribution

A (100%)

ninomfr64 3 weeks, 3 days ago

Selected Answer: A

A. Yes, Athena is the right service to query data in S3.
B. No, maybe this might also work, but it is quite cumbersome
C. No, SparkSQL can be used to query files on data, but it is more work than Athena and creating a new S3 bucket is not needed
D. No, Data Firehose cannot consume from S3 directly

upvoted 1 times

feelgoodfactor 1 month, 2 weeks ago

Selected Answer: A

Using Amazon Athena with a CREATE TABLE AS SELECT (CTAS) statement is the simplest and most efficient way to query the CSV objects based on the transaction date, while requiring minimal operational effort.

upvoted 1 times

motk123 1 month, 3 weeks ago

Selected Answer: A

Athena allows direct querying of data stored in Amazon S3 using SQL without requiring data movement or transformation. CTAS (CREATE TABLE AS SELECT): Creates a new table based on a filtered or transformed dataset, such as transaction dates, and stores the results in S3.
Why Not the Other Options?
B. S3 Object Lambda is designed for on-the-fly data transformation, not querying data efficiently. Adding replication increases complexity without addressing the querying requirement directly.
C. Glue is suited for complex ETL workflows, but it introduces significant operational overhead for a task that Athena can handle more easily.
D. Firehose is designed for streaming data, not processing large existing datasets.

upvoted 2 times

GiorgioGss 2 months ago

Selected Answer: A

Base usage of CTAS

upvoted 2 times

Question #20 *Topic 1*

A company has a large, unstructured dataset. The dataset includes many duplicate records across several key attributes.
Which solution on AWS will detect duplicates in the dataset with the LEAST code development?

A. Use Amazon Mechanical Turk jobs to detect duplicates.

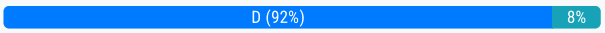
B. Use Amazon QuickSight ML Insights to build a custom deduplication model.

C. Use Amazon SageMaker Data Wrangler to pre-process and detect duplicates.

D. Use the AWS Glue FindMatches transform to detect duplicates.

Suggested Answer: *D*

Community vote distribution



Saransundar Highly Voted 2 months ago

Selected Answer: D

AWS Glue FindMatches is specifically designed to identify duplicate or matching records in datasets without requiring labeled training data. It uses machine learning to find fuzzy matches and allows customization to fine-tune the matching process, making it ideal for this scenario.

upvoted 5 times

feelgoodfactor Most Recent 1 month, 2 weeks ago

Selected Answer: D

The AWS Glue FindMatches transform is the most appropriate solution because it is specifically designed to detect duplicates, requires minimal development effort, and scales efficiently for large datasets.

upvoted 3 times

nakidal495 2 months ago

Selected Answer: A

I'm not sure but I think this is the correct answer.

upvoted 1 times

GiorgioGss 2 months, 1 week ago

Selected Answer: D

https://aws.amazon.com/about-aws/whats-new/2021/11/aws-glue-findmatches-new-data-existing-dataset/
"allows you to identify duplicate or matching records in your dataset"

upvoted 4 times

Question #21 *Topic 1*

A company needs to run a batch data-processing job on Amazon EC2 instances. The job will run during the weekend and will take 90 minutes to finish running. The processing can handle interruptions. The company will run the job every weekend for the next 6 months.
Which EC2 instance purchasing option will meet these requirements MOST cost-effectively?

A. Spot Instances

B. Reserved Instances

C. On-Demand Instances

D. Dedicated Instances

Suggested Answer: *A*

Community vote distribution

A (100%)

Saransundar 2 months ago

Selected Answer: A

Cost effective + Interruptions + Short duration 90mins = Spot instance

upvoted 2 times

GiorgioGss 2 months, 1 week ago

Selected Answer: A

key: "The processing can handle interruptions"

upvoted 2 times

Question #22 *Topic 1*

An ML engineer has an Amazon Comprehend custom model in Account A in the us-east-1 Region. The ML engineer needs to copy the model to Account B in the same Region.
Which solution will meet this requirement with the LEAST development effort?

A. Use Amazon S3 to make a copy of the model. Transfer the copy to Account B.

B. Create a resource-based IAM policy. Use the Amazon Comprehend ImportModel API operation to copy the model to Account B.

C. Use AWS DataSync to replicate the model from Account A to Account B.

D. Create an AWS Site-to-Site VPN connection between Account A and Account B to transfer the model.

Suggested Answer: *B*

Community vote distribution

B (100%)

Saransundar 2 months ago

Selected Answer: B

Amazon Comprehend - ImportModel API to facilitate the transfer of custom models between AWS accounts. STEPS:
1. Exporting the model from Account A.
2. Creating a resource-based IAM policy in Account A to grant access to Account B.
3. Using the ImportModel API in Account B to import the model.

upvoted 2 times

GiorgioGss 2 months, 1 week ago

Selected Answer: B

https://docs.aws.amazon.com/comprehend/latest/APIReference/API_ImportModel.html
"The source model can be in your AWS account or another one."

upvoted 2 times

Question #23 *Topic 1*

An ML engineer is training a simple neural network model. The ML engineer tracks the performance of the model over time on a validation dataset. The model's performance improves substantially at first and then degrades after a specific number of epochs.
Which solutions will mitigate this problem? (Choose two.)

A. Enable early stopping on the model.

B. Increase dropout in the layers.

C. Increase the number of layers.

D. Increase the number of neurons.

E. Investigate and reduce the sources of model bias.

Suggested Answer: *AB*

Community vote distribution

AB (100%)

Saransundar 2 months ago

Selected Answer: AB

The issue is overfitting. Soln:-
A. Early stopping:- Stops training when validation performance declines
B. Increase dropout:- reduces overfitting by randomly disabling neurons
upvoted 1 times

GiorgioGss 2 months ago

Selected Answer: AB

"improves substantially at first and then degrades after a specific number of epochs."
Clear sign to stop it early and to drop
upvoted 1 times

Question #24 *Topic 1*

A company has a Retrieval Augmented Generation (RAG) application that uses a vector database to store embeddings of documents. The company must migrate the application to AWS and must implement a solution that provides semantic search of text files. The company has already migrated the text repository to an Amazon S3 bucket.
Which solution will meet these requirements?

A. Use an AWS Batch job to process the files and generate embeddings. Use AWS Glue to store the embeddings. Use SQL queries to perform the semantic searches.

B. Use a custom Amazon SageMaker notebook to run a custom script to generate embeddings. Use SageMaker Feature Store to store the embeddings. Use SQL queries to perform the semantic searches.

C. Use the Amazon Kendra S3 connector to ingest the documents from the S3 bucket into Amazon Kendra. Query Amazon Kendra to perform the semantic searches.

D. Use an Amazon Textract asynchronous job to ingest the documents from the S3 bucket. Query Amazon Textract to perform the semantic searches.

Suggested Answer: *C*

Community vote distribution

C (100%)

emupsx1 1 month, 2 weeks ago

Selected Answer: C

https://docs.aws.amazon.com/kendra/latest/dg/data-source-s3.html

upvoted 1 times

Question #26 *Topic 1*

A company wants to predict the success of advertising campaigns by considering the color scheme of each advertisement. An ML engineer is preparing data for a neural network model. The dataset includes color information as categorical data.
Which technique for feature engineering should the ML engineer use for the model?

A. Apply label encoding to the color categories. Automatically assign each color a unique integer.

B. Implement padding to ensure that all color feature vectors have the same length.

C. Perform dimensionality reduction on the color categories.

D. One-hot encode the color categories to transform the color scheme feature into a binary matrix.

Suggested Answer: *D*

Community vote distribution

D (100%)

Saransundar 2 months ago

Selected Answer: D

1. Label Encoding: Ordinal relationship
2. Padding: Sequence data
3. Dimensionality Reduction: High-dimensional data
4. One-Hot Encoding: Categorical data (Right)

upvoted 3 times

GiorgioGss 2 months, 1 week ago

Selected Answer: D

One-hot encoding creates a new binary feature for each unique category (color in this case). For example, if there are three colors (red, blue, green), one-hot encoding would create three binary columns like this:
Red: [1, 0, 0]
Blue: [0, 1, 0]
Green: [0, 0, 1]
this way, the model can work with the color feature without assuming ordinal relationships between colors.

upvoted 1 times

Question #27 *Topic 1*

A company uses a hybrid cloud environment. A model that is deployed on premises uses data in Amazon 53 to provide customers with a live conversational engine.
The model is using sensitive data. An ML engineer needs to implement a solution to identify and remove the sensitive data.
Which solution will meet these requirements with the LEAST operational overhead?

A. Deploy the model on Amazon SageMaker. Create a set of AWS Lambda functions to identify and remove the sensitive data.

B. Deploy the model on an Amazon Elastic Container Service (Amazon ECS) cluster that uses AWS Fargate. Create an AWS Batch job to identify and remove the sensitive data.

C. Use Amazon Macie to identify the sensitive data. Create a set of AWS Lambda functions to remove the sensitive data.

D. Use Amazon Comprehend to identify the sensitive data. Launch Amazon EC2 instances to remove the sensitive data.

Suggested Answer: *C*

Community vote distribution

C (100%)

Saransundar 2 months ago

Selected Answer: C

Macie - Identify sensitive data

upvoted 1 times

GiorgioGss 2 months, 1 week ago

Selected Answer: C

This is the purpose of Macie. Comprehend will work also for PII but that EC2...

upvoted 1 times

Question #28 *Topic 1*

An ML engineer needs to create data ingestion pipelines and ML model deployment pipelines on AWS. All the raw data is stored in Amazon S3 buckets.
Which solution will meet these requirements?

A. Use Amazon Data Firehose to create the data ingestion pipelines. Use Amazon SageMaker Studio Classic to create the model deployment pipelines.

B. Use AWS Glue to create the data ingestion pipelines. Use Amazon SageMaker Studio Classic to create the model deployment pipelines.

C. Use Amazon Redshift ML to create the data ingestion pipelines. Use Amazon SageMaker Studio Classic to create the model deployment pipelines.

D. Use Amazon Athena to create the data ingestion pipelines. Use an Amazon SageMaker notebook to create the model deployment pipelines.

Suggested Answer: *B*

Community vote distribution

B (100%)

Saransundar 2 months ago

Selected Answer: B

Data ingestion - Glue ; Model deployment pipeline - sagemaker studio classic

upvoted 1 times

GiorgioGss 2 months, 1 week ago

Selected Answer: B

This is the main use-case for Glue.

upvoted 1 times

Question #29 *Topic 1*

A company that has hundreds of data scientists is using Amazon SageMaker to create ML models. The models are in model groups in the SageMaker Model Registry.
The data scientists are grouped into three categories: computer vision, natural language processing (NLP), and speech recognition. An ML engineer needs to implement a solution to organize the existing models into these groups to improve model discoverability at scale. The solution must not affect the integrity of the model artifacts and their existing groupings.
Which solution will meet these requirements?

A. Create a custom tag for each of the three categories. Add the tags to the model packages in the SageMaker Model Registry.

B. Create a model group for each category. Move the existing models into these category model groups.

C. Use SageMaker ML Lineage Tracking to automatically identify and tag which model groups should contain the models.

D. Create a Model Registry collection for each of the three categories. Move the existing model groups into the collections.

Suggested Answer: *D*

Community vote distribution

D (100%)

abrarjahin 1 week, 5 days ago

Selected Answer: D

Because according to the documentation -
"Any operation you perform on your Collections does not affect the integrity of the individual Model Groups they contain—the underlying Model Group artifacts in Amazon S3 and Amazon ECR are not modified."

upvoted 1 times

Saransundar 2 months ago

Selected Answer: D

Option-D Any operation you perform on your Collections does not affect the integrity of the individual Model Groups they contain—the underlying Model Group artifacts in Amazon S3 and Amazon ECR are not modified.

upvoted 1 times

Linux_master 2 months ago

Selected Answer: D

https://docs.aws.amazon.com/sagemaker/latest/dg/modelcollections.html

upvoted 2 times

GiorgioGss 2 months, 1 week ago

Selected Answer: D

A could also be a valid option but in here we see exactly this:
https://docs.aws.amazon.com/sagemaker/latest/dg/modelcollections.html
"Any operation you perform on your Collections does not affect the integrity of the individual Model Groups they contain—the underlying Model Group artifacts in Amazon S3 and Amazon ECR are not modified."

upvoted 1 times

a4002bd 2 months, 1 week ago

Selected Answer: D

I pick D. Creating custom tags for each of the three categories and adding them to the model packages in the SageMaker Model Registry (Option A) is a valid approach. However, it might not be as effective for organizing models at scale compared to using Model Registry collections.

upvoted 1 times

helloworldabc 2 months ago

just A

upvoted 1 times

Question #30 *Topic 1*

A company runs an Amazon SageMaker domain in a public subnet of a newly created VPC. The network is configured properly, and ML engineers can access the SageMaker domain.
Recently, the company discovered suspicious traffic to the domain from a specific IP address. The company needs to block traffic from the specific IP address.
Which update to the network configuration will meet this requirement?

A. Create a security group inbound rule to deny traffic from the specific IP address. Assign the security group to the domain.

B. Create a network ACL inbound rule to deny traffic from the specific IP address. Assign the rule to the default network Ad for the subnet where the domain is located.

C. Create a shadow variant for the domain. Configure SageMaker Inference Recommender to send traffic from the specific IP address to the shadow endpoint.

D. Create a VPC route table to deny inbound traffic from the specific IP address. Assign the route table to the domain.

Suggested Answer: *B*

Community vote distribution

Saransundar 2 months ago

Selected Answer: B

Protection at subnet level: Network ACL. Specific IP addresses can be denied at inbound connection level

upvoted 2 times

GiorgioGss 2 months, 1 week ago

Selected Answer: B

That's basic network topic.

upvoted 1 times

Question #31 *Topic 1*

A company is gathering audio, video, and text data in various languages. The company needs to use a large language model (LLM) to summarize the gathered data that is in Spanish.
Which solution will meet these requirements in the LEAST amount of time?

A. Train and deploy a model in Amazon SageMaker to convert the data into English text. Train and deploy an LLM in SageMaker to summarize the text.

B. Use Amazon Transcribe and Amazon Translate to convert the data into English text. Use Amazon Bedrock with the Jurassic model to summarize the text.

C. Use Amazon Rekognition and Amazon Translate to convert the data into English text. Use Amazon Bedrock with the Anthropic Claude model to summarize the text.

D. Use Amazon Comprehend and Amazon Translate to convert the data into English text. Use Amazon Bedrock with the Stable Diffusion model to summarize the text.

Suggested Answer: *B*

Community vote distribution

B (100%)

Saransundar 2 months ago

Selected Answer: B

For Spanish : https://docs.ai21.com/docs/jurassic-2-models

upvoted 2 times

GiorgioGss 2 months, 1 week ago

Selected Answer: B

LEAST amount of time -> A is out

C is out because Claude does NOT fit for summarization

D is out because that's for image generation

upvoted 2 times

Question #32 *Topic 1*

A financial company receives a high volume of real-time market data streams from an external provider. The streams consist of thousands of JSON records every second.
The company needs to implement a scalable solution on AWS to identify anomalous data points.
Which solution will meet these requirements with the LEAST operational overhead?

A. Ingest real-time data into Amazon Kinesis data streams. Use the built-in RANDOM_CUT_FOREST function in Amazon Managed Service for Apache Flink to process the data streams and to detect data anomalies.

B. Ingest real-time data into Amazon Kinesis data streams. Deploy an Amazon SageMaker endpoint for real-time outlier detection. Create an AWS Lambda function to detect anomalies. Use the data streams to invoke the Lambda function.

C. Ingest real-time data into Apache Kafka on Amazon EC2 instances. Deploy an Amazon SageMaker endpoint for real-time outlier detection. Create an AWS Lambda function to detect anomalies. Use the data streams to invoke the Lambda function.

D. Send real-time data to an Amazon Simple Queue Service (Amazon SQS) FIFO queue. Create an AWS Lambda function to consume the queue messages. Program the Lambda function to start an AWS Glue extract, transform, and load (ETL) job for batch processing and anomaly detection.

Suggested Answer: *A*

Community vote distribution

Saransundar 2 months ago

Selected Answer: A

Option A
High-volume real-time: Kinesis Data Streams
Scalable: Managed Apache Flink
Anomaly detection: RANDOM_CUT_FOREST
Low overhead: Fully managed services
upvoted 2 times

GiorgioGss 2 months, 1 week ago

Selected Answer: A

https://docs.aws.amazon.com/kinesisanalytics/latest/sqlref/sqlrf-random-cut-forest.html
"Detects anomalies in your data stream."
upvoted 2 times

Question #33 *Topic 1*

A company has a large collection of chat recordings from customer interactions after a product release. An ML engineer needs to create an ML model to analyze the chat data. The ML engineer needs to determine the success of the product by reviewing customer sentiments about the product.
Which action should the ML engineer take to complete the evaluation in the LEAST amount of time?

A. Use Amazon Rekognition to analyze sentiments of the chat conversations.

B. Train a Naive Bayes classifier to analyze sentiments of the chat conversations.

C. Use Amazon Comprehend to analyze sentiments of the chat conversations.

D. Use random forests to classify sentiments of the chat conversations.

Suggested Answer: *C*

Community vote distribution

C (100%)

Saransundar 2 months ago

Selected Answer: C

Prebuilt sentiment analysis + Fast setup + NLP --Comprehend

upvoted 2 times

GiorgioGss 2 months, 1 week ago

Selected Answer: C

https://docs.aws.amazon.com/comprehend/latest/dg/what-is.html

upvoted 2 times

Question #34 *Topic 1*

A company has a conversational AI assistant that sends requests through Amazon Bedrock to an Anthropic Claude large language model (LLM). Users report that when they ask similar questions multiple times, they sometimes receive different answers. An ML engineer needs to improve the responses to be more consistent and less random.
Which solution will meet these requirements?

A. Increase the temperature parameter and the top_k parameter.

B. Increase the temperature parameter. Decrease the top_k parameter.

C. Decrease the temperature parameter. Increase the top_k parameter.

D. Decrease the temperature parameter and the top_k parameter.

Suggested Answer: *D*

Community vote distribution

D (100%)

Saransundar 2 months ago

Selected Answer: D

Lower temperature: High probable output
Lower Top k : Focus on likely output
upvoted 3 times

GiorgioGss 2 months, 1 week ago

Selected Answer: D

https://docs.aws.amazon.com/bedrock/latest/userguide/inference-parameters.html
upvoted 2 times