Actual exam question from Google's Professional Machine Learning Engineer

Question #: 1

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are building an ML model to detect anomalies in real-time sensor data. You will use Pub/Sub to handle incoming requests. You want to store the results for analytics and visualization. How should you configure the pipeline?

A. 1 = Dataflow, 2 = AI Platform, 3 = BigQuery

B. 1 = DataProc, 2 = AutoML, 3 = Cloud Bigtable

C. 1 = BigQuery, 2 = AutoML, 3 = Cloud Functions

D. 1 = BigQuery, 2 = AI Platform, 3 = Cloud Storage

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 2

Topic #: 1

[All Professional Machine Learning Engineer Questions]

Your organization wants to make its internal shuttle service route more efficient. The shuttles currently stop at all pick-up points across the city every 30 minutes between 7 am and 10 am. The development team has already built an application on Google Kubernetes Engine that requires users to confirm their presence and shuttle station one day in advance. What approach should you take?

A. 1. Build a tree-based regression model that predicts how many passengers will be picked up at each shuttle station. 2. Dispatch an appropriately sized shuttle and provide the map with the required stops based on the prediction.

B. 1. Build a tree-based classification model that predicts whether the shuttle should pick up passengers at each shuttle station. 2. Dispatch an available shuttle and provide the map with the required stops based on the prediction.

C. 1. Define the optimal route as the shortest route that passes by all shuttle stations with confirmed attendance at the given time under capacity constraints. 2. Dispatch an appropriately sized shuttle and indicate the required stops on the map.

D. 1. Build a reinforcement learning model with tree-based classification models that predict the presence of passengers at shuttle stops as agents and a reward function around a distance-based metric. 2. Dispatch an appropriately sized shuttle and provide the map with the required stops based on the simulated outcome.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 3

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You were asked to investigate failures of a production line component based on sensor readings. After receiving the dataset, you discover that less than 1% of the readings are positive examples representing failure incidents. You have tried to train several classification models, but none of them converge. How should you resolve the class imbalance problem?

A. Use the class distribution to generate 10% positive examples.

B. Use a convolutional neural network with max pooling and softmax activation.

C. Downsample the data with upweighting to create a sample with 10% positive examples.

D. Remove negative examples until the numbers of positive and negative examples are equal.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 4

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You want to rebuild your ML pipeline for structured data on Google Cloud. You are using PySpark to conduct data transformations at scale, but your pipelines are taking over 12 hours to run. To speed up development and pipeline run time, you want to use a serverless tool and SQL syntax. You have already moved your raw data into Cloud Storage. How should you build the pipeline on Google Cloud while meeting the speed and processing requirements?

A. Use Data Fusion's GUI to build the transformation pipelines, and then write the data into BigQuery.

B. Convert your PySpark into SparkSQL queries to transform the data, and then run your pipeline on Dataproc to write the data into BigQuery.

C. Ingest your data into Cloud SQL, convert your PySpark commands into SQL queries to transform the data, and then use federated queries from BigQuery for machine learning.

D. Ingest your data into BigQuery using BigQuery Load, convert your PySpark commands into BigQuery SQL queries to transform the data, and then write the transformations to a new table.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 5

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You manage a team of data scientists who use a cloud-based backend system to submit training jobs. This system has become very difficult to administer, and you want to use a managed service instead. The data scientists you work with use many different frameworks, including Keras, PyTorch, theano, Scikit-learn, and custom libraries. What should you do?

A. Use the AI Platform custom containers feature to receive training jobs using any framework.

B. Configure Kubeflow to run on Google Kubernetes Engine and receive training jobs through TF Job.

C. Create a library of VM images on Compute Engine, and publish these images on a centralized repository.

D. Set up Slurm workload manager to receive jobs that can be scheduled to run on your cloud infrastructure.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 6

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You work for an online retail company that is creating a visual search engine. You have set up an end-to-end ML pipeline on Google Cloud to classify whether an image contains your company's product. Expecting the release of new products in the near future, you configured a retraining functionality in the pipeline so that new data can be fed into your ML models. You also want to use AI Platform's continuous evaluation service to ensure that the models have high accuracy on your test dataset. What should you do?

A. Keep the original test dataset unchanged even if newer products are incorporated into retraining.

B. Extend your test dataset with images of the newer products when they are introduced to retraining.

C. Replace your test dataset with images of the newer products when they are introduced to retraining.

D. Update your test dataset with images of the newer products when your evaluation metrics drop below a pre-decided threshold.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 7

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You need to build classification workflows over several structured datasets currently stored in BigQuery. Because you will be performing the classification several times, you want to complete the following steps without writing code: exploratory data analysis, feature selection, model building, training, and hyperparameter tuning and serving. What should you do?

A. Configure AutoML Tables to perform the classification task.

B. Run a BigQuery ML task to perform logistic regression for the classification.

C. Use AI Platform Notebooks to run the classification model with pandas library.

D. Use AI Platform to run the classification model job configured for hyperparameter tuning.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 8

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You work for a public transportation company and need to build a model to estimate delay times for multiple transportation routes. Predictions are served directly to users in an app in real time. Because different seasons and population increases impact the data relevance, you will retrain the model every month. You want to follow Google-recommended best practices. How should you configure the end-to-end architecture of the predictive model?

A. Configure Kubeflow Pipelines to schedule your multi-step workflow from training to deploying your model.

B. Use a model trained and deployed on BigQuery ML, and trigger retraining with the scheduled query feature in BigQuery.

C. Write a Cloud Functions script that launches a training and deploying job on AI Platform that is triggered by Cloud Scheduler.

D. Use Cloud Composer to programmatically schedule a Dataflow job that executes the workflow from training to deploying your model.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 9
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are developing ML models with AI Platform for image segmentation on CT scans. You frequently update your model architectures based on the newest available research papers, and have to rerun training on the same dataset to benchmark their performance. You want to minimize computation costs and manual intervention while having version control for your code. What should you do?

A. Use Cloud Functions to identify changes to your code in Cloud Storage and trigger a retraining job.

B. Use the gcloud command-line tool to submit training jobs on AI Platform when you update your code.

C. Use Cloud Build linked with Cloud Source Repositories to trigger retraining when new code is pushed to the repository.

D. Create an automated workflow in Cloud Composer that runs daily and looks for changes in code in Cloud Storage using a sensor.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 10

Topic #: 1

[All Professional Machine Learning Engineer Questions]

Your team needs to build a model that predicts whether images contain a driver's license, passport, or credit card. The data engineering team already built the pipeline and generated a dataset composed of 10,000 images with driver's licenses, 1,000 images with passports, and 1,000 images with credit cards. You now have to train a model with the following label map: [`˜drivers_license', `˜passport', `˜credit_card']. Which loss function should you use?

A. Categorical hinge

B. Binary cross-entropy

C. Categorical cross-entropy

D. Sparse categorical cross-entropy

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 11

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are designing an ML recommendation model for shoppers on your company's ecommerce website. You will use Recommendations AI to build, test, and deploy your system. How should you develop recommendations that increase revenue while following best practices?

A. Use the ג€Other Products You May Likeג€ recommendation type to increase the click-through rate.

B. Use the ג€Frequently Bought Togetherג€ recommendation type to increase the shopping cart size for each order.

C. Import your user events and then your product catalog to make sure you have the highest quality event stream.

D. Because it will take time to collect and record product data, use placeholder values for the product catalog to test the viability of the model.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

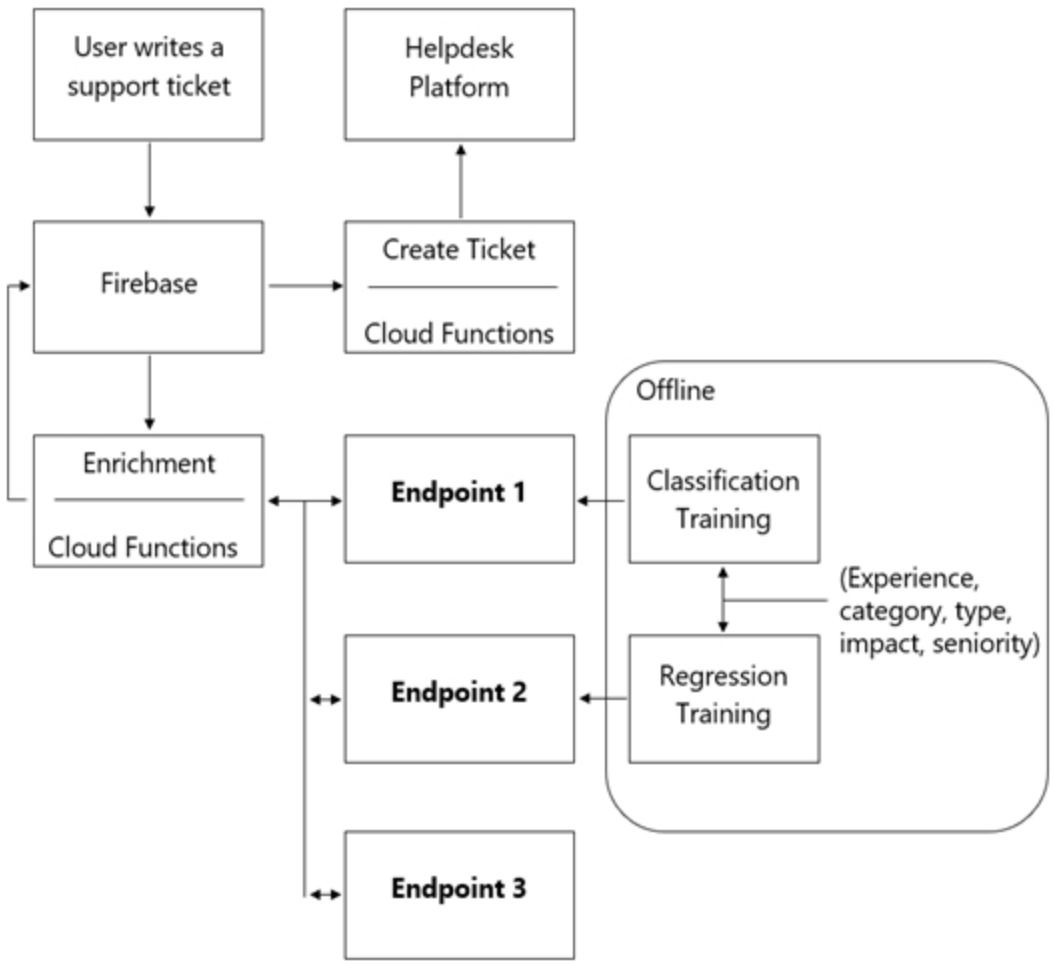
Question #: 12

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are designing an architecture with a serverless ML system to enrich customer support tickets with informative metadata before they are routed to a support agent. You need a set of models to predict ticket priority, predict ticket resolution time, and perform sentiment analysis to help agents make strategic decisions when they process support requests. Tickets are not expected to have any domain-specific terms or jargon.

The proposed architecture has the following flow:



Which endpoints should the Enrichment Cloud Functions call?

A. 1 = AI Platform, 2 = AI Platform, 3 = AutoML Vision

B. 1 = AI Platform, 2 = AI Platform, 3 = AutoML Natural Language

C. 1 = AI Platform, 2 = AI Platform, 3 = Cloud Natural Language API

D. 1 = Cloud Natural Language API, 2 = AI Platform, 3 = Cloud Vision API

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 13

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You have trained a deep neural network model on Google Cloud. The model has low loss on the training data, but is performing worse on the validation data. You want the model to be resilient to overfitting. Which strategy should you use when retraining the model?

A. Apply a dropout parameter of 0.2, and decrease the learning rate by a factor of 10.

B. Apply a L2 regularization parameter of 0.4, and decrease the learning rate by a factor of 10.

C. Run a hyperparameter tuning job on AI Platform to optimize for the L2 regularization and dropout parameters.

D. Run a hyperparameter tuning job on AI Platform to optimize for the learning rate, and increase the number of neurons by a factor of 2.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 14
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You built and manage a production system that is responsible for predicting sales numbers. Model accuracy is crucial, because the production model is required to keep up with market changes. Since being deployed to production, the model hasn't changed; however the accuracy of the model has steadily deteriorated.
What issue is most likely causing the steady decline in model accuracy?

A. Poor data quality

B. Lack of model retraining

C. Too few layers in the model for capturing information

D. Incorrect data split ratio during model training, evaluation, validation, and test

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 15

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You have been asked to develop an input pipeline for an ML training model that processes images from disparate sources at a low latency. You discover that your input data does not fit in memory. How should you create a dataset following Google-recommended best practices?

A. Create a tf.data.Dataset.prefetch transformation.

B. Convert the images to tf.Tensor objects, and then run Dataset.from_tensor_slices().

C. Convert the images to tf.Tensor objects, and then run tf.data.Dataset.from_tensors().

D. Convert the images into TFRecords, store the images in Cloud Storage, and then use the tf.data API to read the images for training.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 16

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are an ML engineer at a large grocery retailer with stores in multiple regions. You have been asked to create an inventory prediction model. Your model's features include region, location, historical demand, and seasonal popularity. You want the algorithm to learn from new inventory data on a daily basis. Which algorithms should you use to build the model?

A. Classification

B. Reinforcement Learning

C. Recurrent Neural Networks (RNN)

D. Convolutional Neural Networks (CNN)

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 17

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are building a real-time prediction engine that streams files which may contain Personally Identifiable Information (PII) to Google Cloud. You want to use the Cloud Data Loss Prevention (DLP) API to scan the files. How should you ensure that the PII is not accessible by unauthorized individuals?

A. Stream all files to Google Cloud, and then write the data to BigQuery. Periodically conduct a bulk scan of the table using the DLP API.

B. Stream all files to Google Cloud, and write batches of the data to BigQuery. While the data is being written to BigQuery, conduct a bulk scan of the data using the DLP API.

C. Create two buckets of data: Sensitive and Non-sensitive. Write all data to the Non-sensitive bucket. Periodically conduct a bulk scan of that bucket using the DLP API, and move the sensitive data to the Sensitive bucket.

D. Create three buckets of data: Quarantine, Sensitive, and Non-sensitive. Write all data to the Quarantine bucket. Periodically conduct a bulk scan of that bucket using the DLP API, and move the data to either the Sensitive or Non-Sensitive bucket.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 18
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for a large hotel chain and have been asked to assist the marketing team in gathering predictions for a targeted marketing strategy. You need to make predictions about user lifetime value (LTV) over the next 20 days so that marketing can be adjusted accordingly. The customer dataset is in BigQuery, and you are preparing the tabular data for training with AutoML Tables. This data has a time signal that is spread across multiple columns. How should you ensure that AutoML fits the best model to your data?

A. Manually combine all columns that contain a time signal into an array. Allow AutoML to interpret this array appropriately. Choose an automatic data split across the training, validation, and testing sets.

B. Submit the data for training without performing any manual transformations. Allow AutoML to handle the appropriate transformations. Choose an automatic data split across the training, validation, and testing sets.

C. Submit the data for training without performing any manual transformations, and indicate an appropriate column as the Time column. Allow AutoML to split your data based on the time signal provided, and reserve the more recent data for the validation and testing sets.

D. Submit the data for training without performing any manual transformations. Use the columns that have a time signal to manually split your data. Ensure that the data in your validation set is from 30 days after the data in your training set and that the data in your testing sets from 30 days after your validation set.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 19
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You have written unit tests for a Kubeflow Pipeline that require custom libraries. You want to automate the execution of unit tests with each new push to your development branch in Cloud Source Repositories. What should you do?

A. Write a script that sequentially performs the push to your development branch and executes the unit tests on Cloud Run.

B. Using Cloud Build, set an automated trigger to execute the unit tests when changes are pushed to your development branch.

C. Set up a Cloud Logging sink to a Pub/Sub topic that captures interactions with Cloud Source Repositories. Configure a Pub/Sub trigger for Cloud Run, and execute the unit tests on Cloud Run.

D. Set up a Cloud Logging sink to a Pub/Sub topic that captures interactions with Cloud Source Repositories. Execute the unit tests using a Cloud Function that is triggered when messages are sent to the Pub/Sub topic.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 20

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are training an LSTM-based model on AI Platform to summarize text using the following job submission script: gcloud ai-platform jobs submit training $JOB_NAME \
--package-path $TRAINER_PACKAGE_PATH \
--module-name $MAIN_TRAINER_MODULE \
--job-dir $JOB_DIR \
--region $REGION \
--scale-tier basic \
-- \
--epochs 20 \
--batch_size=32 \
--learning_rate=0.001 \
You want to ensure that training time is minimized without significantly compromising the accuracy of your model. What should you do?

A. Modify the 'epochs' parameter.

B. Modify the 'scale-tier' parameter.

C. Modify the 'batch size' parameter.

D. Modify the 'learning rate' parameter.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 21

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You have deployed multiple versions of an image classification model on AI Platform. You want to monitor the performance of the model versions over time. How should you perform this comparison?

A. Compare the loss performance for each model on a held-out dataset.

B. Compare the loss performance for each model on the validation data.

C. Compare the receiver operating characteristic (ROC) curve for each model using the What-If Tool.

D. Compare the mean average precision across the models using the Continuous Evaluation feature.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 22

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You trained a text classification model. You have the following SignatureDefs:

```
signature_def['serving_default']:
  The given SavedModel SignatureDef contains the following input(s):
    inputs['text'] tensor_info:
        dtype: DT_STRING
        shape: (-1, 2)
        name: serving_default_text: 0
  The given SavedModel SignatureDef contains the following output(s):
    outputs ['Softmax'] tensor_info:
        dtype: DT_FLOAT
        shape: (-1, 2)
        name: StatefulPartitionedCall:0
  Method name is: tensorflow/serving/predict
```

You started a TensorFlow-serving component server and tried to send an HTTP request to get a prediction using: headers = {"content-type": "application/json"}

json_response = requests.post('http: //localhost:8501/v1/models/text_model:predict', data=data, headers=headers)

What is the correct way to write the predict request?

A. data = json.dumps({ι€signature_nameι :€ι€seving_defaultι ,€ι€instancesι€ [['ab', 'bc', 'cd']]})

B. data = json.dumps({ι€signature_nameι :€ι€serving_defaultι ,€ι€instancesι€ [['a', 'b', 'c', 'd', 'e', 'f']]})

C. data = json.dumps({ι€signature_nameι :€ι€serving_defaultι ,€ι€instancesι€ [['a', 'b', 'c'], ['d', 'e', 'f']]})

D. data = json.dumps({ι€signature_nameι :€ι€serving_defaultι ,€ι€instancesι€ [['a', 'b'], ['c', 'd'], ['e', 'f']]})

Show Suggested Answer

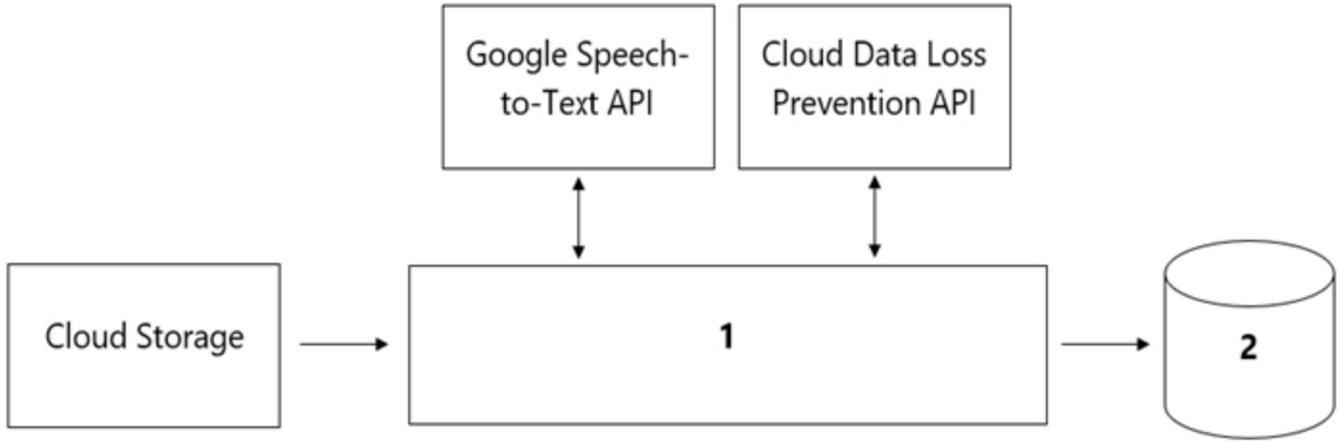
Actual exam question from Google's Professional Machine Learning Engineer

Question #: 23

Topic #: 1

[All Professional Machine Learning Engineer Questions]

Your organization's call center has asked you to develop a model that analyzes customer sentiments in each call. The call center receives over one million calls daily, and data is stored in Cloud Storage. The data collected must not leave the region in which the call originated, and no Personally Identifiable Information (PII) can be stored or analyzed. The data science team has a third-party tool for visualization and access which requires a SQL ANSI-2011 compliant interface. You need to select components for data processing and for analytics. How should the data pipeline be designed?



A. 1= Dataflow, 2= BigQuery

B. 1 = Pub/Sub, 2= Datastore

C. 1 = Dataflow, 2 = Cloud SQL

D. 1 = Cloud Function, 2= Cloud SQL

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 24

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are an ML engineer at a global shoe store. You manage the ML models for the company's website. You are asked to build a model that will recommend new products to the user based on their purchase behavior and similarity with other users. What should you do?

A. Build a classification model

B. Build a knowledge-based filtering model

C. Build a collaborative-based filtering model

D. Build a regression model using the features as predictors

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 25

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You work for a social media company. You need to detect whether posted images contain cars. Each training example is a member of exactly one class. You have trained an object detection neural network and deployed the model version to AI Platform Prediction for evaluation. Before deployment, you created an evaluation job and attached it to the AI Platform Prediction model version. You notice that the precision is lower than your business requirements allow. How should you adjust the model's final layer softmax threshold to increase precision?

A. Increase the recall.

B. Decrease the recall.

C. Increase the number of false positives.

D. Decrease the number of false negatives.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 26
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are responsible for building a unified analytics environment across a variety of on-premises data marts. Your company is experiencing data quality and security challenges when integrating data across the servers, caused by the use of a wide range of disconnected tools and temporary solutions. You need a fully managed, cloud-native data integration service that will lower the total cost of work and reduce repetitive work. Some members on your team prefer a codeless interface for building Extract, Transform, Load (ETL) process. Which service should you use?

A. Dataflow

B. Dataprep

C. Apache Flink

D. Cloud Data Fusion

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 27

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are an ML engineer at a regulated insurance company. You are asked to develop an insurance approval model that accepts or rejects insurance applications from potential customers. What factors should you consider before building the model?

A. Redaction, reproducibility, and explainability

B. Traceability, reproducibility, and explainability

C. Federated learning, reproducibility, and explainability

D. Differential privacy, federated learning, and explainability

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 28

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are training a Resnet model on AI Platform using TPUs to visually categorize types of defects in automobile engines. You capture the training profile using the Cloud TPU profiler plugin and observe that it is highly input-bound. You want to reduce the bottleneck and speed up your model training process. Which modifications should you make to the tf.data dataset? (Choose two.)

A. Use the interleave option for reading data.

B. Reduce the value of the repeat parameter.

C. Increase the buffer size for the shuttle option.

D. Set the prefetch option equal to the training batch size.

E. Decrease the batch size argument in your transformation.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 29

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You have trained a model on a dataset that required computationally expensive preprocessing operations. You need to execute the same preprocessing at prediction time. You deployed the model on AI Platform for high-throughput online prediction. Which architecture should you use?

A. Validate the accuracy of the model that you trained on preprocessed data. Create a new model that uses the raw data and is available in real time. Deploy the new model onto AI Platform for online prediction.

B. Send incoming prediction requests to a Pub/Sub topic. Transform the incoming data using a Dataflow job. Submit a prediction request to AI Platform using the transformed data. Write the predictions to an outbound Pub/Sub queue.

C. Stream incoming prediction request data into Cloud Spanner. Create a view to abstract your preprocessing logic. Query the view every second for new records. Submit a prediction request to AI Platform using the transformed data. Write the predictions to an outbound Pub/Sub queue.

D. Send incoming prediction requests to a Pub/Sub topic. Set up a Cloud Function that is triggered when messages are published to the Pub/Sub topic. Implement your preprocessing logic in the Cloud Function. Submit a prediction request to AI Platform using the transformed data. Write the predictions to an outbound Pub/Sub queue.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 30

Topic #: 1

[All Professional Machine Learning Engineer Questions]

Your team trained and tested a DNN regression model with good results. Six months after deployment, the model is performing poorly due to a change in the distribution of the input data. How should you address the input differences in production?

A. Create alerts to monitor for skew, and retrain the model.

B. Perform feature selection on the model, and retrain the model with fewer features.

C. Retrain the model, and select an L2 regularization parameter with a hyperparameter tuning service.

D. Perform feature selection on the model, and retrain the model on a monthly basis with fewer features.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 31

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You need to train a computer vision model that predicts the type of government ID present in a given image using a GPU-powered virtual machine on Compute Engine. You use the following parameters:

- ✑ Optimizer: SGD
- ✑ Image shape = 224—224
- ✑ Batch size = 64
- ✑ Epochs = 10
- ✑ Verbose =2

During training you encounter the following error: ResourceExhaustedError: Out Of Memory (OOM) when allocating tensor. What should you do?

A. Change the optimizer.

B. Reduce the batch size.

C. Change the learning rate.

D. Reduce the image shape.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 32
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You developed an ML model with AI Platform, and you want to move it to production. You serve a few thousand queries per second and are experiencing latency issues. Incoming requests are served by a load balancer that distributes them across multiple Kubeflow CPU-only pods running on Google Kubernetes Engine (GKE). Your goal is to improve the serving latency without changing the underlying infrastructure. What should you do?

A. Significantly increase the max_batch_size TensorFlow Serving parameter.

B. Switch to the tensorflow-model-server-universal version of TensorFlow Serving.

C. Significantly increase the max_enqueued_batches TensorFlow Serving parameter.

D. Recompile TensorFlow Serving using the source to support CPU-specific optimizations. Instruct GKE to choose an appropriate baseline minimum CPU platform for serving nodes.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 33

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You have a demand forecasting pipeline in production that uses Dataflow to preprocess raw data prior to model training and prediction. During preprocessing, you employ Z-score normalization on data stored in BigQuery and write it back to BigQuery. New training data is added every week. You want to make the process more efficient by minimizing computation time and manual intervention. What should you do?

A. Normalize the data using Google Kubernetes Engine.

B. Translate the normalization algorithm into SQL for use with BigQuery.

C. Use the normalizer_fn argument in TensorFlow's Feature Column API.

D. Normalize the data with Apache Spark using the Dataproc connector for BigQuery.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 34

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You need to design a customized deep neural network in Keras that will predict customer purchases based on their purchase history. You want to explore model performance using multiple model architectures, store training data, and be able to compare the evaluation metrics in the same dashboard. What should you do?

A. Create multiple models using AutoML Tables.

B. Automate multiple training runs using Cloud Composer.

C. Run multiple training jobs on AI Platform with similar job names.

D. Create an experiment in Kubeflow Pipelines to organize multiple runs.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 35

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are developing a Kubeflow pipeline on Google Kubernetes Engine. The first step in the pipeline is to issue a query against BigQuery. You plan to use the results of that query as the input to the next step in your pipeline. You want to achieve this in the easiest way possible. What should you do?

A. Use the BigQuery console to execute your query, and then save the query results into a new BigQuery table.

B. Write a Python script that uses the BigQuery API to execute queries against BigQuery. Execute this script as the first step in your Kubeflow pipeline.

C. Use the Kubeflow Pipelines domain-specific language to create a custom component that uses the Python BigQuery client library to execute queries.

D. Locate the Kubeflow Pipelines repository on GitHub. Find the BigQuery Query Component, copy that component's URL, and use it to load the component into your pipeline. Use the component to execute queries against BigQuery.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 36

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are building a model to predict daily temperatures. You split the data randomly and then transformed the training and test datasets. Temperature data for model training is uploaded hourly. During testing, your model performed with 97% accuracy; however, after deploying to production, the model's accuracy dropped to 66%. How can you make your production model more accurate?

A. Normalize the data for the training, and test datasets as two separate steps.

B. Split the training and test data based on time rather than a random split to avoid leakage.

C. Add more data to your test set to ensure that you have a fair distribution and sample for testing.

D. Apply data transformations before splitting, and cross-validate to make sure that the transformations are applied to both the training and test sets.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 37
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are developing models to classify customer support emails. You created models with TensorFlow Estimators using small datasets on your on-premises system, but you now need to train the models using large datasets to ensure high performance. You will port your models to Google Cloud and want to minimize code refactoring and infrastructure overhead for easier migration from on-prem to cloud. What should you do?

A. Use AI Platform for distributed training.

B. Create a cluster on Dataproc for training.

C. Create a Managed Instance Group with autoscaling.

D. Use Kubeflow Pipelines to train on a Google Kubernetes Engine cluster.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 38

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You have trained a text classification model in TensorFlow using AI Platform. You want to use the trained model for batch predictions on text data stored in BigQuery while minimizing computational overhead. What should you do?

A. Export the model to BigQuery ML.

B. Deploy and version the model on AI Platform.

C. Use Dataflow with the SavedModel to read the data from BigQuery.

D. Submit a batch prediction job on AI Platform that points to the model location in Cloud Storage.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 39

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You work with a data engineering team that has developed a pipeline to clean your dataset and save it in a Cloud Storage bucket. You have created an ML model and want to use the data to refresh your model as soon as new data is available. As part of your CI/CD workflow, you want to automatically run a Kubeflow Pipelines training job on Google Kubernetes Engine (GKE). How should you architect this workflow?

A. Configure your pipeline with Dataflow, which saves the files in Cloud Storage. After the file is saved, start the training job on a GKE cluster.

B. Use App Engine to create a lightweight python client that continuously polls Cloud Storage for new files. As soon as a file arrives, initiate the training job.

C. Configure a Cloud Storage trigger to send a message to a Pub/Sub topic when a new file is available in a storage bucket. Use a Pub/Sub-triggered Cloud Function to start the training job on a GKE cluster.

D. Use Cloud Scheduler to schedule jobs at a regular interval. For the first step of the job, check the timestamp of objects in your Cloud Storage bucket. If there are no new files since the last run, abort the job.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 40
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You have a functioning end-to-end ML pipeline that involves tuning the hyperparameters of your ML model using AI Platform, and then using the best-tuned parameters for training. Hypertuning is taking longer than expected and is delaying the downstream processes. You want to speed up the tuning job without significantly compromising its effectiveness. Which actions should you take? (Choose two.)

A. Decrease the number of parallel trials.

B. Decrease the range of floating-point values.

C. Set the early stopping parameter to TRUE.

D. Change the search algorithm from Bayesian search to random search.

E. Decrease the maximum number of trials during subsequent training phases.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 41
Topic #: 1
[All Professional Machine Learning Engineer Questions]

Your team is building an application for a global bank that will be used by millions of customers. You built a forecasting model that predicts customers' account balances 3 days in the future. Your team will use the results in a new feature that will notify users when their account balance is likely to drop below $25. How should you serve your predictions?

A. 1. Create a Pub/Sub topic for each user. 2. Deploy a Cloud Function that sends a notification when your model predicts that a user's account balance will drop below the $25 threshold.

B. 1. Create a Pub/Sub topic for each user. 2. Deploy an application on the App Engine standard environment that sends a notification when your model predicts that a user's account balance will drop below the $25 threshold.

C. 1. Build a notification system on Firebase. 2. Register each user with a user ID on the Firebase Cloud Messaging server, which sends a notification when the average of all account balance predictions drops below the $25 threshold.

D. 1. Build a notification system on Firebase. 2. Register each user with a user ID on the Firebase Cloud Messaging server, which sends a notification when your model predicts that a user's account balance will drop below the $25 threshold.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 42
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for an advertising company and want to understand the effectiveness of your company's latest advertising campaign. You have streamed 500 MB of campaign data into BigQuery. You want to query the table, and then manipulate the results of that query with a pandas dataframe in an AI Platform notebook. What should you do?

A. Use AI Platform Notebooks' BigQuery cell magic to query the data, and ingest the results as a pandas dataframe.

B. Export your table as a CSV file from BigQuery to Google Drive, and use the Google Drive API to ingest the file into your notebook instance.

C. Download your table from BigQuery as a local CSV file, and upload it to your AI Platform notebook instance. Use pandas.read_csv to ingest he file as a pandas dataframe.

D. From a bash cell in your AI Platform notebook, use the bq extract command to export the table as a CSV file to Cloud Storage, and then use gsutil cp to copy the data into the notebook. Use pandas.read_csv to ingest the file as a pandas dataframe.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 43

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are an ML engineer at a global car manufacture. You need to build an ML model to predict car sales in different cities around the world. Which features or feature crosses should you use to train city-specific relationships between car type and number of sales?

A. Thee individual features: binned latitude, binned longitude, and one-hot encoded car type.

B. One feature obtained as an element-wise product between latitude, longitude, and car type.

C. One feature obtained as an element-wise product between binned latitude, binned longitude, and one-hot encoded car type.

D. Two feature crosses as an element-wise product: the first between binned latitude and one-hot encoded car type, and the second between binned longitude and one-hot encoded car type.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 44
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for a large technology company that wants to modernize their contact center. You have been asked to develop a solution to classify incoming calls by product so that requests can be more quickly routed to the correct support team. You have already transcribed the calls using the Speech-to-Text API. You want to minimize data preprocessing and development time. How should you build the model?

A. Use the AI Platform Training built-in algorithms to create a custom model.

B. Use AutoMIL Natural Language to extract custom entities for classification.

C. Use the Cloud Natural Language API to extract custom entities for classification.

D. Build a custom model to identify the product keywords from the transcribed calls, and then run the keywords through a classification algorithm.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 45

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are training a TensorFlow model on a structured dataset with 100 billion records stored in several CSV files. You need to improve the input/output execution performance. What should you do?

A. Load the data into BigQuery, and read the data from BigQuery.

B. Load the data into Cloud Bigtable, and read the data from Bigtable.

C. Convert the CSV files into shards of TFRecords, and store the data in Cloud Storage.

D. Convert the CSV files into shards of TFRecords, and store the data in the Hadoop Distributed File System (HDFS).

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 46
Topic #: 1
[All Professional Machine Learning Engineer Questions]

As the lead ML Engineer for your company, you are responsible for building ML models to digitize scanned customer forms. You have developed a TensorFlow model that converts the scanned images into text and stores them in Cloud Storage. You need to use your ML model on the aggregated data collected at the end of each day with minimal manual intervention. What should you do?

A. Use the batch prediction functionality of AI Platform.

B. Create a serving pipeline in Compute Engine for prediction.

C. Use Cloud Functions for prediction each time a new data point is ingested.

D. Deploy the model on AI Platform and create a version of it for online inference.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 47
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You recently joined an enterprise-scale company that has thousands of datasets. You know that there are accurate descriptions for each table in BigQuery, and you are searching for the proper BigQuery table to use for a model you are building on AI Platform. How should you find the data that you need?

A. Use Data Catalog to search the BigQuery datasets by using keywords in the table description.

B. Tag each of your model and version resources on AI Platform with the name of the BigQuery table that was used for training.

C. Maintain a lookup table in BigQuery that maps the table descriptions to the table ID. Query the lookup table to find the correct table ID for the data that you need.

D. Execute a query in BigQuery to retrieve all the existing table names in your project using the INFORMATION_SCHEMA metadata tables that are native to BigQuery. Use the result o find the table that you need.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 48
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You started working on a classification problem with time series data and achieved an area under the receiver operating characteristic curve (AUC ROC) value of 99% for training data after just a few experiments. You haven't explored using any sophisticated algorithms or spent any time on hyperparameter tuning. What should your next step be to identify and fix the problem?

A. Address the model overfitting by using a less complex algorithm.

B. Address data leakage by applying nested cross-validation during model training.

C. Address data leakage by removing features highly correlated with the target value.

D. Address the model overfitting by tuning the hyperparameters to reduce the AUC ROC value.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 49
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for an online travel agency that also sells advertising placements on its website to other companies. You have been asked to predict the most relevant web banner that a user should see next. Security is important to your company. The model latency requirements are 300ms@p99, the inventory is thousands of web banners, and your exploratory analysis has shown that navigation context is a good predictor. You want to Implement the simplest solution. How should you configure the prediction pipeline?

A. Embed the client on the website, and then deploy the model on AI Platform Prediction.

B. Embed the client on the website, deploy the gateway on App Engine, and then deploy the model on AI Platform Prediction.

C. Embed the client on the website, deploy the gateway on App Engine, deploy the database on Cloud Bigtable for writing and for reading the user's navigation context, and then deploy the model on AI Platform Prediction.

D. Embed the client on the website, deploy the gateway on App Engine, deploy the database on Memorystore for writing and for reading the user's navigation context, and then deploy the model on Google Kubernetes Engine.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 50

Topic #: 1

[All Professional Machine Learning Engineer Questions]

Your team is building a convolutional neural network (CNN)-based architecture from scratch. The preliminary experiments running on your on-premises CPU-only infrastructure were encouraging, but have slow convergence. You have been asked to speed up model training to reduce time-to-market. You want to experiment with virtual machines (VMs) on Google Cloud to leverage more powerful hardware. Your code does not include any manual device placement and has not been wrapped in Estimator model-level abstraction. Which environment should you train your model on?

A. AVM on Compute Engine and 1 TPU with all dependencies installed manually.

B. AVM on Compute Engine and 8 GPUs with all dependencies installed manually.

C. A Deep Learning VM with an n1-standard-2 machine and 1 GPU with all libraries pre-installed.

D. A Deep Learning VM with more powerful CPU e2-highcpu-16 machines with all libraries pre-installed.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 51
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work on a growing team of more than 50 data scientists who all use AI Platform. You are designing a strategy to organize your jobs, models, and versions in a clean and scalable way. Which strategy should you choose?

A. Set up restrictive IAM permissions on the AI Platform notebooks so that only a single user or group can access a given instance.

B. Separate each data scientist's work into a different project to ensure that the jobs, models, and versions created by each data scientist are accessible only to that user.

C. Use labels to organize resources into descriptive categories. Apply a label to each created resource so that users can filter the results by label when viewing or monitoring the resources.

D. Set up a BigQuery sink for Cloud Logging logs that is appropriately filtered to capture information about AI Platform resource usage. In BigQuery, create a SQL view that maps users to the resources they are using

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 52

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are training a deep learning model for semantic image segmentation with reduced training time. While using a Deep Learning VM Image, you receive the following error: The resource 'projects/deeplearning-platform/zones/europe-west4-c/acceleratorTypes/nvidia-tesla-k80' was not found. What should you do?

A. Ensure that you have GPU quota in the selected region.

B. Ensure that the required GPU is available in the selected region.

C. Ensure that you have preemptible GPU quota in the selected region.

D. Ensure that the selected GPU has enough GPU memory for the workload.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 53

Topic #: 1

[All Professional Machine Learning Engineer Questions]

Your team is working on an NLP research project to predict political affiliation of authors based on articles they have written. You have a large training dataset that is structured like this:

```
AuthorA:Political Party A
    TextA1: [SentenceA11, SentenceA12, SentenceA13, ...]
    TextA2: [SentenceA21, SentenceA22, SentenceA23, ...]
    ...
AuthorB:Political Party B
    TextB1: [SentenceB11, SentenceB12, SentenceB13, ...]
    TextB2: [SentenceB21, SentenceB22, SentenceB23, ...]
    ...
AuthorC:Political Party B
    TextC1: [SentenceC11, SentenceC12, SentenceC13, ...]
    TextC2: [SentenceC21, SentenceC22, SentenceC23, ...]
    ...
AuthorD:Political Party A
    TextD1: [SentenceD11, SentenceD12, SentenceD13, ...]
    TextD2: [SentenceD21, SentenceD22, SentenceD23, ...]
    ...
...
```

You followed the standard 80%-10%-10% data distribution across the training, testing, and evaluation subsets. How should you distribute the training examples across the train-test-eval subsets while maintaining the 80-10-10 proportion?

A. Distribute texts randomly across the train-test-eval subsets: Train set: [TextA1, TextB2, ...] Test set: [TextA2, TextC1, TextD2, ...] Eval set: [TextB1, TextC2, TextD1, ...]

B. Distribute authors randomly across the train-test-eval subsets: (*) Train set: [TextA1, TextA2, TextD1, TextD2, ...] Test set: [TextB1, TextB2, ...] Eval set: [TexC1,TextC2 ...]

C. Distribute sentences randomly across the train-test-eval subsets: Train set: [SentenceA11, SentenceA21, SentenceB11, SentenceB21, SentenceC11, SentenceD21 ...] Test set: [SentenceA12, SentenceA22, SentenceB12, SentenceC22, SentenceC12, SentenceD22 ...] Eval set: [SentenceA13, SentenceA23, SentenceB13, SentenceC23, SentenceC13, SentenceD31 ...]

D. Distribute paragraphs of texts (i.e., chunks of consecutive sentences) across the train-test-eval subsets: Train set: [SentenceA11, SentenceA12, SentenceD11, SentenceD12 ...] Test set: [SentenceA13, SentenceB13, SentenceB21, SentenceD23, SentenceC12, SentenceD13 ...] Eval set: [SentenceA11, SentenceA22, SentenceB13, SentenceD22, SentenceC23, SentenceD11 ...]

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 54
Topic #: 1
[All Professional Machine Learning Engineer Questions]

Your team has been tasked with creating an ML solution in Google Cloud to classify support requests for one of your platforms. You analyzed the requirements and decided to use TensorFlow to build the classifier so that you have full control of the model's code, serving, and deployment. You will use Kubeflow pipelines for the ML platform. To save time, you want to build on existing resources and use managed services instead of building a completely new model. How should you build the classifier?

A. Use the Natural Language API to classify support requests.

B. Use AutoML Natural Language to build the support requests classifier.

C. Use an established text classification model on AI Platform to perform transfer learning.

D. Use an established text classification model on AI Platform as-is to classify support requests.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 55
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You recently joined a machine learning team that will soon release a new project. As a lead on the project, you are asked to determine the production readiness of the ML components. The team has already tested features and data, model development, and infrastructure. Which additional readiness check should you recommend to the team?

A. Ensure that training is reproducible.

B. Ensure that all hyperparameters are tuned.

C. Ensure that model performance is monitored.

D. Ensure that feature expectations are captured in the schema.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 56

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You work for a credit card company and have been asked to create a custom fraud detection model based on historical data using AutoML Tables. You need to prioritize detection of fraudulent transactions while minimizing false positives. Which optimization objective should you use when training the model?

A. An optimization objective that minimizes Log loss

B. An optimization objective that maximizes the Precision at a Recall value of 0.50

C. An optimization objective that maximizes the area under the precision-recall curve (AUC PR) value

D. An optimization objective that maximizes the area under the receiver operating characteristic curve (AUC ROC) value

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 57
Topic #: 1
[All Professional Machine Learning Engineer Questions]

Your company manages a video sharing website where users can watch and upload videos. You need to create an ML model to predict which newly uploaded videos will be the most popular so that those videos can be prioritized on your company's website. Which result should you use to determine whether the model is successful?

A. The model predicts videos as popular if the user who uploads them has over 10,000 likes.

B. The model predicts 97.5% of the most popular clickbait videos measured by number of clicks.

C. The model predicts 95% of the most popular videos measured by watch time within 30 days of being uploaded.

D. The Pearson correlation coefficient between the log-transformed number of views after 7 days and 30 days after publication is equal to 0.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 58

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are working on a Neural Network-based project. The dataset provided to you has columns with different ranges. While preparing the data for model training, you discover that gradient optimization is having difficulty moving weights to a good solution. What should you do?

A. Use feature construction to combine the strongest features.

B. Use the representation transformation (normalization) technique.

C. Improve the data cleaning step by removing features with missing values.

D. Change the partitioning step to reduce the dimension of the test set and have a larger training set.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 59

Topic #: 1

[All Professional Machine Learning Engineer Questions]

Your data science team needs to rapidly experiment with various features, model architectures, and hyperparameters. They need to track the accuracy metrics for various experiments and use an API to query the metrics over time. What should they use to track and report their experiments while minimizing manual effort?

A. Use Kubeflow Pipelines to execute the experiments. Export the metrics file, and query the results using the Kubeflow Pipelines API.

B. Use AI Platform Training to execute the experiments. Write the accuracy metrics to BigQuery, and query the results using the BigQuery API.

C. Use AI Platform Training to execute the experiments. Write the accuracy metrics to Cloud Monitoring, and query the results using the Monitoring API.

D. Use AI Platform Notebooks to execute the experiments. Collect the results in a shared Google Sheets file, and query the results using the Google Sheets API.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 60

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You work for a bank and are building a random forest model for fraud detection. You have a dataset that includes transactions, of which 1% are identified as fraudulent. Which data transformation strategy would likely improve the performance of your classifier?

A. Write your data in TFRecords.

B. Z-normalize all the numeric features.

C. Oversample the fraudulent transaction 10 times.

D. Use one-hot encoding on all categorical features.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 61

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are using transfer learning to train an image classifier based on a pre-trained EfficientNet model. Your training dataset has 20,000 images. You plan to retrain the model once per day. You need to minimize the cost of infrastructure. What platform components and configuration environment should you use?

A. A Deep Learning VM with 4 V100 GPUs and local storage.

B. A Deep Learning VM with 4 V100 GPUs and Cloud Storage.

C. A Google Kubernetes Engine cluster with a V100 GPU Node Pool and an NFS Server

D. An AI Platform Training job using a custom scale tier with 4 V100 GPUs and Cloud Storage

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 62

Topic #: 1

[All Professional Machine Learning Engineer Questions]

While conducting an exploratory analysis of a dataset, you discover that categorical feature A has substantial predictive power, but it is sometimes missing. What should you do?

A. Drop feature A if more than 15% of values are missing. Otherwise, use feature A as-is.

B. Compute the mode of feature A and then use it to replace the missing values in feature A.

C. Replace the missing values with the values of the feature with the highest Pearson correlation with feature A.

D. Add an additional class to categorical feature A for missing values. Create a new binary feature that indicates whether feature A is missing.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 63

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You work for a large retailer and have been asked to segment your customers by their purchasing habits. The purchase history of all customers has been uploaded to BigQuery. You suspect that there may be several distinct customer segments, however you are unsure of how many, and you don't yet understand the commonalities in their behavior. You want to find the most efficient solution. What should you do?

A. Create a k-means clustering model using BigQuery ML. Allow BigQuery to automatically optimize the number of clusters.

B. Create a new dataset in Dataprep that references your BigQuery table. Use Dataprep to identify similarities within each column.

C. Use the Data Labeling Service to label each customer record in BigQuery. Train a model on your labeled data using AutoML Tables. Review the evaluation metrics to understand whether there is an underlying pattern in the data.

D. Get a list of the customer segments from your company's Marketing team. Use the Data Labeling Service to label each customer record in BigQuery according to the list. Analyze the distribution of labels in your dataset using Data Studio.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 64
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You recently designed and built a custom neural network that uses critical dependencies specific to your organization's framework. You need to train the model using a managed training service on Google Cloud. However, the ML framework and related dependencies are not supported by AI Platform Training. Also, both your model and your data are too large to fit in memory on a single machine. Your ML framework of choice uses the scheduler, workers, and servers distribution structure. What should you do?

A. Use a built-in model available on AI Platform Training.

B. Build your custom container to run jobs on AI Platform Training.

C. Build your custom containers to run distributed training jobs on AI Platform Training.

D. Reconfigure your code to a ML framework with dependencies that are supported by AI Platform Training.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 65

Topic #: 1

[All Professional Machine Learning Engineer Questions]

While monitoring your model training's GPU utilization, you discover that you have a native synchronous implementation. The training data is split into multiple files. You want to reduce the execution time of your input pipeline. What should you do?

A. Increase the CPU load

B. Add caching to the pipeline

C. Increase the network bandwidth

D. Add parallel interleave to the pipeline

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 66
Topic #: 1
[All Professional Machine Learning Engineer Questions]

Your data science team is training a PyTorch model for image classification based on a pre-trained RestNet model. You need to perform hyperparameter tuning to optimize for several parameters. What should you do?

A. Convert the model to a Keras model, and run a Keras Tuner job.

B. Run a hyperparameter tuning job on AI Platform using custom containers.

C. Create a Kuberflow Pipelines instance, and run a hyperparameter tuning job on Katib.

D. Convert the model to a TensorFlow model, and run a hyperparameter tuning job on AI Platform.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 67

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You have a large corpus of written support cases that can be classified into 3 separate categories: Technical Support, Billing Support, or Other Issues. You need to quickly build, test, and deploy a service that will automatically classify future written requests into one of the categories. How should you configure the pipeline?

A. Use the Cloud Natural Language API to obtain metadata to classify the incoming cases.

B. Use AutoML Natural Language to build and test a classifier. Deploy the model as a REST API.

C. Use BigQuery ML to build and test a logistic regression model to classify incoming requests. Use BigQuery ML to perform inference.

D. Create a TensorFlow model using Google's BERT pre-trained model. Build and test a classifier, and deploy the model using Vertex AI.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 68

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You need to quickly build and train a model to predict the sentiment of customer reviews with custom categories without writing code. You do not have enough data to train a model from scratch. The resulting model should have high predictive performance. Which service should you use?

A. AutoML Natural Language

B. Cloud Natural Language API

C. AI Hub pre-made Jupyter Notebooks

D. AI Platform Training built-in algorithms

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 69

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You need to build an ML model for a social media application to predict whether a user's submitted profile photo meets the requirements. The application will inform the user if the picture meets the requirements. How should you build a model to ensure that the application does not falsely accept a non-compliant picture?

A. Use AutoML to optimize the model's recall in order to minimize false negatives.

B. Use AutoML to optimize the model's F1 score in order to balance the accuracy of false positives and false negatives.

C. Use Vertex AI Workbench user-managed notebooks to build a custom model that has three times as many examples of pictures that meet the profile photo requirements.

D. Use Vertex AI Workbench user-managed notebooks to build a custom model that has three times as many examples of pictures that do not meet the profile photo requirements.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 70

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You lead a data science team at a large international corporation. Most of the models your team trains are large-scale models using high-level TensorFlow APIs on AI Platform with GPUs. Your team usually takes a few weeks or months to iterate on a new version of a model. You were recently asked to review your team's spending. How should you reduce your Google Cloud compute costs without impacting the model's performance?

A. Use AI Platform to run distributed training jobs with checkpoints.

B. Use AI Platform to run distributed training jobs without checkpoints.

C. Migrate to training with Kuberflow on Google Kubernetes Engine, and use preemptible VMs with checkpoints.

D. Migrate to training with Kuberflow on Google Kubernetes Engine, and use preemptible VMs without checkpoints.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 71

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You need to train a regression model based on a dataset containing 50,000 records that is stored in BigQuery. The data includes a total of 20 categorical and numerical features with a target variable that can include negative values. You need to minimize effort and training time while maximizing model performance. What approach should you take to train this regression model?

A. Create a custom TensorFlow DNN model

B. Use BQML XGBoost regression to train the model.

C. Use AutoML Tables to train the model without early stopping.

D. Use AutoML Tables to train the model with RMSLE as the optimization objective.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 72

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are building a linear model with over 100 input features, all with values between –1 and 1. You suspect that many features are non-informative. You want to remove the non-informative features from your model while keeping the informative ones in their original form. Which technique should you use?

A. Use principal component analysis (PCA) to eliminate the least informative features.

B. Use L1 regularization to reduce the coefficients of uninformative features to 0.

C. After building your model, use Shapley values to determine which features are the most informative.

D. Use an iterative dropout technique to identify which features do not degrade the model when removed.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 73

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You work for a global footwear retailer and need to predict when an item will be out of stock based on historical inventory data Customer behavior is highly dynamic since footwear demand is influenced by many different factors. You want to serve models that are trained on all available data, but track your performance on specific subsets of data before pushing to production. What is the most streamlined and reliable way to perform this validation?

A. Use then TFX ModelValidator tools to specify performance metrics for production readiness.

B. Use k-fold cross-validation as a validation strategy to ensure that your model is ready for production.

C. Use the last relevant week of data as a validation set to ensure that your model is performing accurately on current data.

D. Use the entire dataset and treat the area under the receiver operating characteristics curve (AUC ROC) as the main metric.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 74

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You have deployed a model on Vertex AI for real-time inference. During an online prediction request, you get an "Out of Memory" error. What should you do?

A. Use batch prediction mode instead of online mode.

B. Send the request again with a smaller batch of instances.

C. Use base64 to encode your data before using it for prediction.

D. Apply for a quota increase for the number of prediction requests.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 75
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work at a subscription-based company. You have trained an ensemble of trees and neural networks to predict customer churn, which is the likelihood that customers will not renew their yearly subscription. The average prediction is a 15% churn rate, but for a particular customer the model predicts that they are 70% likely to churn. The customer has a product usage history of 30%, is located in New York City, and became a customer in 1997. You need to explain the difference between the actual prediction, a 70% churn rate, and the average prediction. You want to use Vertex Explainable AI. What should you do?

A. Train local surrogate models to explain individual predictions.

B. Configure sampled Shapley explanations on Vertex Explainable AI.

C. Configure integrated gradients explanations on Vertex Explainable AI.

D. Measure the effect of each feature as the weight of the feature multiplied by the feature value.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 76

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are working on a classification problem with time series data. After conducting just a few experiments using random cross-validation, you achieved an Area Under the Receiver Operating Characteristic Curve (AUC ROC) value of 99% on the training data. You haven't explored using any sophisticated algorithms or spent any time on hyperparameter tuning. What should your next step be to identify and fix the problem?

A. Address the model overfitting by using a less complex algorithm and use k-fold cross-validation.

B. Address data leakage by applying nested cross-validation during model training.

C. Address data leakage by removing features highly correlated with the target value.

D. Address the model overfitting by tuning the hyperparameters to reduce the AUC ROC value.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 77
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You need to execute a batch prediction on 100 million records in a BigQuery table with a custom TensorFlow DNN regressor model, and then store the predicted results in a BigQuery table. You want to minimize the effort required to build this inference pipeline. What should you do?

A. Import the TensorFlow model with BigQuery ML, and run the ml.predict function.

B. Use the TensorFlow BigQuery reader to load the data, and use the BigQuery API to write the results to BigQuery.

C. Create a Dataflow pipeline to convert the data in BigQuery to TFRecords. Run a batch inference on Vertex AI Prediction, and write the results to BigQuery.

D. Load the TensorFlow SavedModel in a Dataflow pipeline. Use the BigQuery I/O connector with a custom function to perform the inference within the pipeline, and write the results to BigQuery.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 78

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are creating a deep neural network classification model using a dataset with categorical input values. Certain columns have a cardinality greater than 10,000 unique values. How should you encode these categorical values as input into the model?

A. Convert each categorical value into an integer value.

B. Convert the categorical string data to one-hot hash buckets.

C. Map the categorical variables into a vector of boolean values.

D. Convert each categorical value into a run-length encoded string.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 79

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You need to train a natural language model to perform text classification on product descriptions that contain millions of examples and 100,000 unique words. You want to preprocess the words individually so that they can be fed into a recurrent neural network. What should you do?

A. Create a hot-encoding of words, and feed the encodings into your model.

B. Identify word embeddings from a pre-trained model, and use the embeddings in your model.

C. Sort the words by frequency of occurrence, and use the frequencies as the encodings in your model.

D. Assign a numerical value to each word from 1 to 100,000 and feed the values as inputs in your model.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 80
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for an online travel agency that also sells advertising placements on its website to other companies. You have been asked to predict the most relevant web banner that a user should see next. Security is important to your company. The model latency requirements are 300ms@p99, the inventory is thousands of web banners, and your exploratory analysis has shown that navigation context is a good predictor. You want to Implement the simplest solution. How should you configure the prediction pipeline?

A. Embed the client on the website, and then deploy the model on AI Platform Prediction.

B. Embed the client on the website, deploy the gateway on App Engine, deploy the database on Firestore for writing and for reading the user's navigation context, and then deploy the model on AI Platform Prediction.

C. Embed the client on the website, deploy the gateway on App Engine, deploy the database on Cloud Bigtable for writing and for reading the user's navigation context, and then deploy the model on AI Platform Prediction.

D. Embed the client on the website, deploy the gateway on App Engine, deploy the database on Memorystore for writing and for reading the user's navigation context, and then deploy the model on Google Kubernetes Engine.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 81
Topic #: 1
[All Professional Machine Learning Engineer Questions]

Your data science team has requested a system that supports scheduled model retraining, Docker containers, and a service that supports autoscaling and monitoring for online prediction requests. Which platform components should you choose for this system?

A. Vertex AI Pipelines and App Engine

B. Vertex AI Pipelines, Vertex AI Prediction, and Vertex AI Model Monitoring

C. Cloud Composer, BigQuery ML, and Vertex AI Prediction

D. Cloud Composer, Vertex AI Training with custom containers, and App Engine

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 82

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are profiling the performance of your TensorFlow model training time and notice a performance issue caused by inefficiencies in the input data pipeline for a single 5 terabyte CSV file dataset on Cloud Storage. You need to optimize the input pipeline performance. Which action should you try first to increase the efficiency of your pipeline?

A. Preprocess the input CSV file into a TFRecord file.

B. Randomly select a 10 gigabyte subset of the data to train your model.

C. Split into multiple CSV files and use a parallel interleave transformation.

D. Set the reshuffle_each_iteration parameter to true in the tf.data.Dataset.shuffle method.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 83

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You need to design an architecture that serves asynchronous predictions to determine whether a particular mission-critical machine part will fail. Your system collects data from multiple sensors from the machine. You want to build a model that will predict a failure in the next N minutes, given the average of each sensor's data from the past 12 hours. How should you design the architecture?

A. 1. HTTP requests are sent by the sensors to your ML model, which is deployed as a microservice and exposes a REST API for prediction
2. Your application queries a Vertex AI endpoint where you deployed your model.
3. Responses are received by the caller application as soon as the model produces the prediction.

B. 1. Events are sent by the sensors to Pub/Sub, consumed in real time, and processed by a Dataflow stream processing pipeline.
2. The pipeline invokes the model for prediction and sends the predictions to another Pub/Sub topic.
3. Pub/Sub messages containing predictions are then consumed by a downstream system for monitoring.

C. 1. Export your data to Cloud Storage using Dataflow.
2. Submit a Vertex AI batch prediction job that uses your trained model in Cloud Storage to perform scoring on the preprocessed data.
3. Export the batch prediction job outputs from Cloud Storage and import them into Cloud SQL.

D. 1. Export the data to Cloud Storage using the BigQuery command-line tool
2. Submit a Vertex AI batch prediction job that uses your trained model in Cloud Storage to perform scoring on the preprocessed data.
3. Export the batch prediction job outputs from Cloud Storage and import them into BigQuery.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 84

Topic #: 1

[All Professional Machine Learning Engineer Questions]

Your company manages an application that aggregates news articles from many different online sources and sends them to users. You need to build a recommendation model that will suggest articles to readers that are similar to the articles they are currently reading. Which approach should you use?

A. Create a collaborative filtering system that recommends articles to a user based on the user's past behavior.

B. Encode all articles into vectors using word2vec, and build a model that returns articles based on vector similarity.

C. Build a logistic regression model for each user that predicts whether an article should be recommended to a user.

D. Manually label a few hundred articles, and then train an SVM classifier based on the manually classified articles that categorizes additional articles into their respective categories.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 85
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for a large social network service provider whose users post articles and discuss news. Millions of comments are posted online each day, and more than 200 human moderators constantly review comments and flag those that are inappropriate. Your team is building an ML model to help human moderators check content on the platform. The model scores each comment and flags suspicious comments to be reviewed by a human. Which metric(s) should you use to monitor the model's performance?

A. Number of messages flagged by the model per minute

B. Number of messages flagged by the model per minute confirmed as being inappropriate by humans.

C. Precision and recall estimates based on a random sample of 0.1% of raw messages each minute sent to a human for review

D. Precision and recall estimates based on a sample of messages flagged by the model as potentially inappropriate each minute

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 86

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are a lead ML engineer at a retail company. You want to track and manage ML metadata in a centralized way so that your team can have reproducible experiments by generating artifacts. Which management solution should you recommend to your team?

A. Store your tf.logging data in BigQuery.

B. Manage all relational entities in the Hive Metastore.

C. Store all ML metadata in Google Cloud's operations suite.

D. Manage your ML workflows with Vertex ML Metadata.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 87
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You have been given a dataset with sales predictions based on your company's marketing activities. The data is structured and stored in BigQuery, and has been carefully managed by a team of data analysts. You need to prepare a report providing insights into the predictive capabilities of the data. You were asked to run several ML models with different levels of sophistication, including simple models and multilayered neural networks. You only have a few hours to gather the results of your experiments. Which Google Cloud tools should you use to complete this task in the most efficient and self-serviced way?

A. Use BigQuery ML to run several regression models, and analyze their performance.

B. Read the data from BigQuery using Dataproc, and run several models using SparkML.

C. Use Vertex AI Workbench user-managed notebooks with scikit-learn code for a variety of ML algorithms and performance metrics.

D. Train a custom TensorFlow model with Vertex AI, reading the data from BigQuery featuring a variety of ML algorithms.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 88

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are an ML engineer at a bank. You have developed a binary classification model using AutoML Tables to predict whether a customer will make loan payments on time. The output is used to approve or reject loan requests. One customer's loan request has been rejected by your model, and the bank's risks department is asking you to provide the reasons that contributed to the model's decision. What should you do?

A. Use local feature importance from the predictions.

B. Use the correlation with target values in the data summary page.

C. Use the feature importance percentages in the model evaluation page.

D. Vary features independently to identify the threshold per feature that changes the classification.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 89
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for a magazine distributor and need to build a model that predicts which customers will renew their subscriptions for the upcoming year. Using your company's historical data as your training set, you created a TensorFlow model and deployed it to AI Platform. You need to determine which customer attribute has the most predictive power for each prediction served by the model. What should you do?

A. Use AI Platform notebooks to perform a Lasso regression analysis on your model, which will eliminate features that do not provide a strong signal.

B. Stream prediction results to BigQuery. Use BigQuery's CORR(X1, X2) function to calculate the Pearson correlation coefficient between each feature and the target variable.

C. Use the AI Explanations feature on AI Platform. Submit each prediction request with the 'explain' keyword to retrieve feature attributions using the sampled Shapley method.

D. Use the What-If tool in Google Cloud to determine how your model will perform when individual features are excluded. Rank the feature importance in order of those that caused the most significant performance drop when removed from the model.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 90

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are working on a binary classification ML algorithm that detects whether an image of a classified scanned document contains a company's logo. In the dataset, 96% of examples don't have the logo, so the dataset is very skewed. Which metrics would give you the most confidence in your model?

A. F-score where recall is weighed more than precision

B. RMSE

C. F1 score

D. F-score where precision is weighed more than recall

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 91

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You work on the data science team for a multinational beverage company. You need to develop an ML model to predict the company's profitability for a new line of naturally flavored bottled waters in different locations. You are provided with historical data that includes product types, product sales volumes, expenses, and profits for all regions. What should you use as the input and output for your model?

A. Use latitude, longitude, and product type as features. Use profit as model output.

B. Use latitude, longitude, and product type as features. Use revenue and expenses as model outputs.

C. Use product type and the feature cross of latitude with longitude, followed by binning, as features. Use profit as model output.

D. Use product type and the feature cross of latitude with longitude, followed by binning, as features. Use revenue and expenses as model outputs.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 92
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work as an ML engineer at a social media company, and you are developing a visual filter for users' profile photos. This requires you to train an ML model to detect bounding boxes around human faces. You want to use this filter in your company's iOS-based mobile phone application. You want to minimize code development and want the model to be optimized for inference on mobile phones. What should you do?

A. Train a model using AutoML Vision and use the "export for Core ML" option.

B. Train a model using AutoML Vision and use the "export for Coral" option.

C. Train a model using AutoML Vision and use the "export for TensorFlow.js" option.

D. Train a custom TensorFlow model and convert it to TensorFlow Lite (TFLite).

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 93

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You have been asked to build a model using a dataset that is stored in a medium-sized (~10 GB) BigQuery table. You need to quickly determine whether this data is suitable for model development. You want to create a one-time report that includes both informative visualizations of data distributions and more sophisticated statistical analyses to share with other ML engineers on your team. You require maximum flexibility to create your report. What should you do?

A. Use Vertex AI Workbench user-managed notebooks to generate the report.

B. Use the Google Data Studio to create the report.

C. Use the output from TensorFlow Data Validation on Dataflow to generate the report.

D. Use Dataprep to create the report.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 94
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work on an operations team at an international company that manages a large fleet of on-premises servers located in few data centers around the world. Your team collects monitoring data from the servers, including CPU/memory consumption. When an incident occurs on a server, your team is responsible for fixing it. Incident data has not been properly labeled yet. Your management team wants you to build a predictive maintenance solution that uses monitoring data from the VMs to detect potential failures and then alerts the service desk team. What should you do first?

A. Train a time-series model to predict the machines' performance values. Configure an alert if a machine's actual performance values significantly differ from the predicted performance values.

B. Implement a simple heuristic (e.g., based on z-score) to label the machines' historical performance data. Train a model to predict anomalies based on this labeled dataset.

C. Develop a simple heuristic (e.g., based on z-score) to label the machines' historical performance data. Test this heuristic in a production environment.

D. Hire a team of qualified analysts to review and label the machines' historical performance data. Train a model based on this manually labeled dataset.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 95
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are developing an ML model that uses sliced frames from video feed and creates bounding boxes around specific objects. You want to automate the following steps in your training pipeline: ingestion and preprocessing of data in Cloud Storage, followed by training and hyperparameter tuning of the object model using Vertex AI jobs, and finally deploying the model to an endpoint. You want to orchestrate the entire pipeline with minimal cluster management. What approach should you use?

A. Use Kubeflow Pipelines on Google Kubernetes Engine.

B. Use Vertex AI Pipelines with TensorFlow Extended (TFX) SDK.

C. Use Vertex AI Pipelines with Kubeflow Pipelines SDK.

D. Use Cloud Composer for the orchestration.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 96

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are training an object detection machine learning model on a dataset that consists of three million X-ray images, each roughly 2 GB in size. You are using Vertex AI Training to run a custom training application on a Compute Engine instance with 32-cores, 128 GB of RAM, and 1 NVIDIA P100 GPU. You notice that model training is taking a very long time. You want to decrease training time without sacrificing model performance. What should you do?

A. Increase the instance memory to 512 GB and increase the batch size.

B. Replace the NVIDIA P100 GPU with a v3-32 TPU in the training job.

C. Enable early stopping in your Vertex AI Training job.

D. Use the tf.distribute.Strategy API and run a distributed training job.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 97
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are a data scientist at an industrial equipment manufacturing company. You are developing a regression model to estimate the power consumption in the company's manufacturing plants based on sensor data collected from all of the plants. The sensors collect tens of millions of records every day. You need to schedule daily training runs for your model that use all the data collected up to the current date. You want your model to scale smoothly and require minimal development work. What should you do?

A. Train a regression model using AutoML Tables.

B. Develop a custom TensorFlow regression model, and optimize it using Vertex AI Training.

C. Develop a custom scikit-learn regression model, and optimize it using Vertex AI Training.

D. Develop a regression model using BigQuery ML.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 98

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You built a custom ML model using scikit-learn. Training time is taking longer than expected. You decide to migrate your model to Vertex AI Training, and you want to improve the model's training time. What should you try out first?

A. Migrate your model to TensorFlow, and train it using Vertex AI Training.

B. Train your model in a distributed mode using multiple Compute Engine VMs.

C. Train your model with DLVM images on Vertex AI, and ensure that your code utilizes NumPy and SciPy internal methods whenever possible.

D. Train your model using Vertex AI Training with GPUs.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 99

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are an ML engineer at a travel company. You have been researching customers' travel behavior for many years, and you have deployed models that predict customers' vacation patterns. You have observed that customers' vacation destinations vary based on seasonality and holidays; however, these seasonal variations are similar across years. You want to quickly and easily store and compare the model versions and performance statistics across years. What should you do?

A. Store the performance statistics in Cloud SQL. Query that database to compare the performance statistics across the model versions.

B. Create versions of your models for each season per year in Vertex AI. Compare the performance statistics across the models in the Evaluate tab of the Vertex AI UI.

C. Store the performance statistics of each pipeline run in Kubeflow under an experiment for each season per year. Compare the results across the experiments in the Kubeflow UI.

D. Store the performance statistics of each version of your models using seasons and years as events in Vertex ML Metadata. Compare the results across the slices.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 100
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are an ML engineer at a manufacturing company. You need to build a model that identifies defects in products based on images of the product taken at the end of the assembly line. You want your model to preprocess the images with lower computation to quickly extract features of defects in products. Which approach should you use to build the model?

A. Reinforcement learning

B. Recommender system

C. Recurrent Neural Networks (RNN)

D. Convolutional Neural Networks (CNN)

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 101
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are developing an ML model intended to classify whether X-ray images indicate bone fracture risk. You have trained a ResNet architecture on Vertex AI using a TPU as an accelerator, however you are unsatisfied with the training time and memory usage. You want to quickly iterate your training code but make minimal changes to the code. You also want to minimize impact on the model's accuracy. What should you do?

A. Reduce the number of layers in the model architecture.

B. Reduce the global batch size from 1024 to 256.

C. Reduce the dimensions of the images used in the model.

D. Configure your model to use bfloat16 instead of float32.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 102
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You have successfully deployed to production a large and complex TensorFlow model trained on tabular data. You want to predict the lifetime value (LTV) field for each subscription stored in the BigQuery table named subscription. subscriptionPurchase in the project named my-fortune500-company-project.

You have organized all your training code, from preprocessing data from the BigQuery table up to deploying the validated model to the Vertex AI endpoint, into a TensorFlow Extended (TFX) pipeline. You want to prevent prediction drift, i.e., a situation when a feature data distribution in production changes significantly over time. What should you do?

A. Implement continuous retraining of the model daily using Vertex AI Pipelines.

B. Add a model monitoring job where 10% of incoming predictions are sampled 24 hours.

C. Add a model monitoring job where 90% of incoming predictions are sampled 24 hours.

D. Add a model monitoring job where 10% of incoming predictions are sampled every hour.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 103

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You recently developed a deep learning model using Keras, and now you are experimenting with different training strategies. First, you trained the model using a single GPU, but the training process was too slow. Next, you distributed the training across 4 GPUs using tf.distribute.MirroredStrategy (with no other changes), but you did not observe a decrease in training time. What should you do?

A. Distribute the dataset with tf.distribute.Strategy.experimental_distribute_dataset

B. Create a custom training loop.

C. Use a TPU with tf.distribute.TPUStrategy.

D. Increase the batch size.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 104
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for a gaming company that has millions of customers around the world. All games offer a chat feature that allows players to communicate with each other in real time. Messages can be typed in more than 20 languages and are translated in real time using the Cloud Translation API. You have been asked to build an ML system to moderate the chat in real time while assuring that the performance is uniform across the various languages and without changing the serving infrastructure.

You trained your first model using an in-house word2vec model for embedding the chat messages translated by the Cloud Translation API. However, the model has significant differences in performance across the different languages. How should you improve it?

A. Add a regularization term such as the Min-Diff algorithm to the loss function.

B. Train a classifier using the chat messages in their original language.

C. Replace the in-house word2vec with GPT-3 or T5.

D. Remove moderation for languages for which the false positive rate is too high.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 105
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for a gaming company that develops massively multiplayer online (MMO) games. You built a TensorFlow model that predicts whether players will make in-app purchases of more than $10 in the next two weeks. The model's predictions will be used to adapt each user's game experience. User data is stored in BigQuery. How should you serve your model while optimizing cost, user experience, and ease of management?

A. Import the model into BigQuery ML. Make predictions using batch reading data from BigQuery, and push the data to Cloud SQL

B. Deploy the model to Vertex AI Prediction. Make predictions using batch reading data from Cloud Bigtable, and push the data to Cloud SQL.

C. Embed the model in the mobile application. Make predictions after every in-app purchase event is published in Pub/Sub, and push the data to Cloud SQL.

D. Embed the model in the streaming Dataflow pipeline. Make predictions after every in-app purchase event is published in Pub/Sub, and push the data to Cloud SQL.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 106
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are building a linear regression model on BigQuery ML to predict a customer's likelihood of purchasing your company's products. Your model uses a city name variable as a key predictive component. In order to train and serve the model, your data must be organized in columns. You want to prepare your data using the least amount of coding while maintaining the predictable variables. What should you do?

A. Use TensorFlow to create a categorical variable with a vocabulary list. Create the vocabulary file, and upload it as part of your model to BigQuery ML.

B. Create a new view with BigQuery that does not include a column with city information

C. Use Cloud Data Fusion to assign each city to a region labeled as 1, 2, 3, 4, or 5, and then use that number to represent the city in the model.

D. Use Dataprep to transform the state column using a one-hot encoding method, and make each city a column with binary values.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 107
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are an ML engineer at a bank that has a mobile application. Management has asked you to build an ML-based biometric authentication for the app that verifies a customer's identity based on their fingerprint. Fingerprints are considered highly sensitive personal information and cannot be downloaded and stored into the bank databases. Which learning strategy should you recommend to train and deploy this ML mode?

A. Data Loss Prevention API

B. Federated learning

C. MD5 to encrypt data

D. Differential privacy

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 108
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are experimenting with a built-in distributed XGBoost model in Vertex AI Workbench user-managed notebooks. You use BigQuery to split your data into training and validation sets using the following queries:

```
CREATE OR REPLACE TABLE 'myproject.mydataset.training' AS
(SELECT * FROM 'myproject.mydataset.mytable' WHERE RAND() <= 0.8);

CREATE OR REPLACE TABLE 'myproject.mydataset.validation' AS
(SELECT * FROM 'myproject.mydataset.mytable' WHERE RAND() <= 0.2);
```

After training the model, you achieve an area under the receiver operating characteristic curve (AUC ROC) value of 0.8, but after deploying the model to production, you notice that your model performance has dropped to an AUC ROC value of 0.65. What problem is most likely occurring?

A. There is training-serving skew in your production environment.

B. There is not a sufficient amount of training data.

C. The tables that you created to hold your training and validation records share some records, and you may not be using all the data in your initial table.

D. The RAND() function generated a number that is less than 0.2 in both instances, so every record in the validation table will also be in the training table.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 109

Topic #: 1

[All Professional Machine Learning Engineer Questions]

During batch training of a neural network, you notice that there is an oscillation in the loss. How should you adjust your model to ensure that it converges?

A. Decrease the size of the training batch.

B. Decrease the learning rate hyperparameter.

C. Increase the learning rate hyperparameter.

D. Increase the size of the training batch.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 110
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for a toy manufacturer that has been experiencing a large increase in demand. You need to build an ML model to reduce the amount of time spent by quality control inspectors checking for product defects. Faster defect detection is a priority. The factory does not have reliable Wi-Fi. Your company wants to implement the new ML model as soon as possible. Which model should you use?

A. AutoML Vision Edge mobile-high-accuracy-1 model

B. AutoML Vision Edge mobile-low-latency-1 model

C. AutoML Vision model

D. AutoML Vision Edge mobile-versatile-1 model

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 111
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You need to build classification workflows over several structured datasets currently stored in BigQuery. Because you will be performing the classification several times, you want to complete the following steps without writing code: exploratory data analysis, feature selection, model building, training, and hyperparameter tuning and serving. What should you do?

A. Train a TensorFlow model on Vertex AI.

B. Train a classification Vertex AutoML model.

C. Run a logistic regression job on BigQuery ML.

D. Use scikit-learn in Notebooks with pandas library.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 112

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are an ML engineer in the contact center of a large enterprise. You need to build a sentiment analysis tool that predicts customer sentiment from recorded phone conversations. You need to identify the best approach to building a model while ensuring that the gender, age, and cultural differences of the customers who called the contact center do not impact any stage of the model development pipeline and results. What should you do?

A. Convert the speech to text and extract sentiments based on the sentences.

B. Convert the speech to text and build a model based on the words.

C. Extract sentiment directly from the voice recordings.

D. Convert the speech to text and extract sentiment using syntactical analysis.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 113

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You need to analyze user activity data from your company's mobile applications. Your team will use BigQuery for data analysis, transformation, and experimentation with ML algorithms. You need to ensure real-time ingestion of the user activity data into BigQuery. What should you do?

A. Configure Pub/Sub to stream the data into BigQuery.

B. Run an Apache Spark streaming job on Dataproc to ingest the data into BigQuery.

C. Run a Dataflow streaming job to ingest the data into BigQuery.

D. Configure Pub/Sub and a Dataflow streaming job to ingest the data into BigQuery,

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 114

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You work for a gaming company that manages a popular online multiplayer game where teams with 6 players play against each other in 5-minute battles. There are many new players every day. You need to build a model that automatically assigns available players to teams in real time. User research indicates that the game is more enjoyable when battles have players with similar skill levels. Which business metrics should you track to measure your model's performance?

A. Average time players wait before being assigned to a team

B. Precision and recall of assigning players to teams based on their predicted versus actual ability

C. User engagement as measured by the number of battles played daily per user

D. Rate of return as measured by additional revenue generated minus the cost of developing a new model

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 115

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are building an ML model to predict trends in the stock market based on a wide range of factors. While exploring the data, you notice that some features have a large range. You want to ensure that the features with the largest magnitude don't overfit the model. What should you do?

A. Standardize the data by transforming it with a logarithmic function.

B. Apply a principal component analysis (PCA) to minimize the effect of any particular feature.

C. Use a binning strategy to replace the magnitude of each feature with the appropriate bin number.

D. Normalize the data by scaling it to have values between 0 and 1.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 116
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for a biotech startup that is experimenting with deep learning ML models based on properties of biological organisms. Your team frequently works on early-stage experiments with new architectures of ML models, and writes custom TensorFlow ops in C++. You train your models on large datasets and large batch sizes. Your typical batch size has 1024 examples, and each example is about 1 MB in size. The average size of a network with all weights and embeddings is 20 GB. What hardware should you choose for your models?

A. A cluster with 2 n1-highcpu-64 machines, each with 8 NVIDIA Tesla V100 GPUs (128 GB GPU memory in total), and a n1-highcpu-64 machine with 64 vCPUs and 58 GB RAM

B. A cluster with 2 a2-megagpu-16g machines, each with 16 NVIDIA Tesla A100 GPUs (640 GB GPU memory in total), 96 vCPUs, and 1.4 TB RAM

C. A cluster with an n1-highcpu-64 machine with a v2-8 TPU and 64 GB RAM

D. A cluster with 4 n1-highcpu-96 machines, each with 96 vCPUs and 86 GB RAM

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 117
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are an ML engineer at an ecommerce company and have been tasked with building a model that predicts how much inventory the logistics team should order each month. Which approach should you take?

A. Use a clustering algorithm to group popular items together. Give the list to the logistics team so they can increase inventory of the popular items.

B. Use a regression model to predict how much additional inventory should be purchased each month. Give the results to the logistics team at the beginning of the month so they can increase inventory by the amount predicted by the model.

C. Use a time series forecasting model to predict each item's monthly sales. Give the results to the logistics team so they can base inventory on the amount predicted by the model.

D. Use a classification model to classify inventory levels as UNDER_STOCKED, OVER_STOCKED, and CORRECTLY_STOCKEGive the report to the logistics team each month so they can fine-tune inventory levels.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 118
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are building a TensorFlow model for a financial institution that predicts the impact of consumer spending on inflation globally. Due to the size and nature of the data, your model is long-running across all types of hardware, and you have built frequent checkpointing into the training process. Your organization has asked you to minimize cost. What hardware should you choose?

A. A Vertex AI Workbench user-managed notebooks instance running on an n1-standard-16 with 4 NVIDIA P100 GPUs

B. A Vertex AI Workbench user-managed notebooks instance running on an n1-standard-16 with an NVIDIA P100 GPU

C. A Vertex AI Workbench user-managed notebooks instance running on an n1-standard-16 with a non-preemptible v3-8 TPU

D. A Vertex AI Workbench user-managed notebooks instance running on an n1-standard-16 with a preemptible v3-8 TPU

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 119
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for a company that provides an anti-spam service that flags and hides spam posts on social media platforms. Your company currently uses a list of 200,000 keywords to identify suspected spam posts. If a post contains more than a few of these keywords, the post is identified as spam. You want to start using machine learning to flag spam posts for human review. What is the main advantage of implementing machine learning for this business case?

A. Posts can be compared to the keyword list much more quickly.

B. New problematic phrases can be identified in spam posts.

C. A much longer keyword list can be used to flag spam posts.

D. Spam posts can be flagged using far fewer keywords.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 120
Topic #: 1
[All Professional Machine Learning Engineer Questions]

One of your models is trained using data provided by a third-party data broker. The data broker does not reliably notify you of formatting changes in the data. You want to make your model training pipeline more robust to issues like this. What should you do?

A. Use TensorFlow Data Validation to detect and flag schema anomalies.

B. Use TensorFlow Transform to create a preprocessing component that will normalize data to the expected distribution, and replace values that don't match the schema with 0.

C. Use tf.math to analyze the data, compute summary statistics, and flag statistical anomalies.

D. Use custom TensorFlow functions at the start of your model training to detect and flag known formatting errors.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 121
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for a company that is developing a new video streaming platform. You have been asked to create a recommendation system that will suggest the next video for a user to watch. After a review by an AI Ethics team, you are approved to start development. Each video asset in your company's catalog has useful metadata (e.g., content type, release date, country), but you do not have any historical user event data. How should you build the recommendation system for the first version of the product?

A. Launch the product without machine learning. Present videos to users alphabetically, and start collecting user event data so you can develop a recommender model in the future.

B. Launch the product without machine learning. Use simple heuristics based on content metadata to recommend similar videos to users, and start collecting user event data so you can develop a recommender model in the future.

C. Launch the product with machine learning. Use a publicly available dataset such as MovieLens to train a model using the Recommendations AI, and then apply this trained model to your data.

D. Launch the product with machine learning. Generate embeddings for each video by training an autoencoder on the content metadata using TensorFlow. Cluster content based on the similarity of these embeddings, and then recommend videos from the same cluster.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 122
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You recently built the first version of an image segmentation model for a self-driving car. After deploying the model, you observe a decrease in the area under the curve (AUC) metric. When analyzing the video recordings, you also discover that the model fails in highly congested traffic but works as expected when there is less traffic. What is the most likely reason for this result?

A. The model is overfitting in areas with less traffic and underfitting in areas with more traffic.

B. AUC is not the correct metric to evaluate this classification model.

C. Too much data representing congested areas was used for model training.

D. Gradients become small and vanish while backpropagating from the output to input nodes.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 123

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are developing an ML model to predict house prices. While preparing the data, you discover that an important predictor variable, distance from the closest school, is often missing and does not have high variance. Every instance (row) in your data is important. How should you handle the missing data?

A. Delete the rows that have missing values.

B. Apply feature crossing with another column that does not have missing values.

C. Predict the missing values using linear regression.

D. Replace the missing values with zeros.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 124
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are an ML engineer responsible for designing and implementing training pipelines for ML models. You need to create an end-to-end training pipeline for a TensorFlow model. The TensorFlow model will be trained on several terabytes of structured data. You need the pipeline to include data quality checks before training and model quality checks after training but prior to deployment. You want to minimize development time and the need for infrastructure maintenance. How should you build and orchestrate your training pipeline?

A. Create the pipeline using Kubeflow Pipelines domain-specific language (DSL) and predefined Google Cloud components. Orchestrate the pipeline using Vertex AI Pipelines.

B. Create the pipeline using TensorFlow Extended (TFX) and standard TFX components. Orchestrate the pipeline using Vertex AI Pipelines.

C. Create the pipeline using Kubeflow Pipelines domain-specific language (DSL) and predefined Google Cloud components. Orchestrate the pipeline using Kubeflow Pipelines deployed on Google Kubernetes Engine.

D. Create the pipeline using TensorFlow Extended (TFX) and standard TFX components. Orchestrate the pipeline using Kubeflow Pipelines deployed on Google Kubernetes Engine.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 125
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You manage a team of data scientists who use a cloud-based backend system to submit training jobs. This system has become very difficult to administer, and you want to use a managed service instead. The data scientists you work with use many different frameworks, including Keras, PyTorch, theano, scikit-learn, and custom libraries. What should you do?

A. Use the Vertex AI Training to submit training jobs using any framework.

B. Configure Kubeflow to run on Google Kubernetes Engine and submit training jobs through TFJob.

C. Create a library of VM images on Compute Engine, and publish these images on a centralized repository.

D. Set up Slurm workload manager to receive jobs that can be scheduled to run on your cloud infrastructure.

Show Suggested Answer

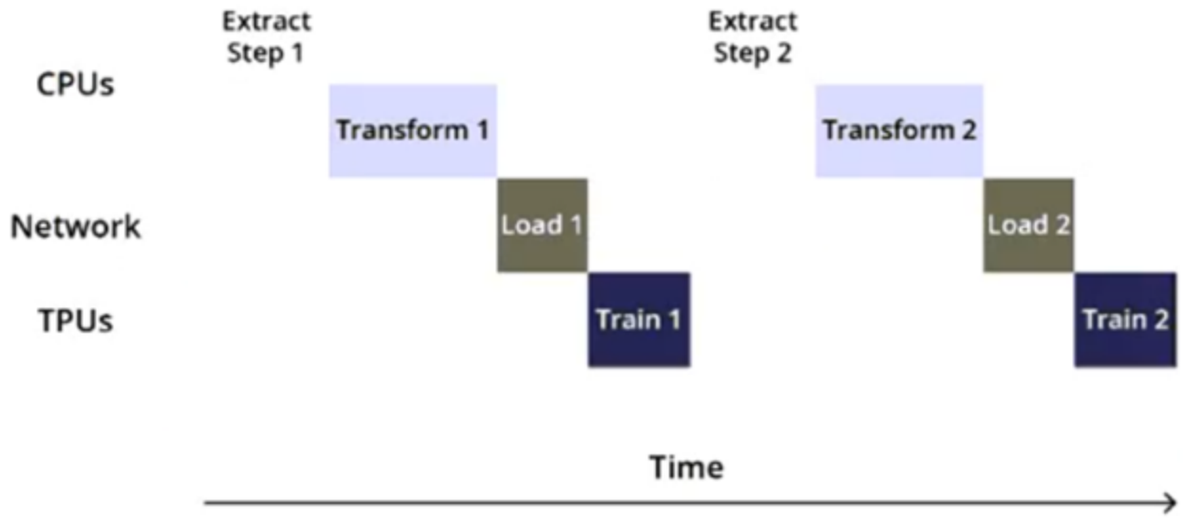
Actual exam question from Google's Professional Machine Learning Engineer

Question #: 126

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are training an object detection model using a Cloud TPU v2. Training time is taking longer than expected. Based on this simplified trace obtained with a Cloud TPU profile, what action should you take to decrease training time in a cost-efficient way?



A. Move from Cloud TPU v2 to Cloud TPU v3 and increase batch size.

B. Move from Cloud TPU v2 to 8 NVIDIA V100 GPUs and increase batch size.

C. Rewrite your input function to resize and reshape the input images.

D. Rewrite your input function using parallel reads, parallel processing, and prefetch.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 127

Topic #: 1

[All Professional Machine Learning Engineer Questions]

While performing exploratory data analysis on a dataset, you find that an important categorical feature has 5% null values. You want to minimize the bias that could result from the missing values. How should you handle the missing values?

A. Remove the rows with missing values, and upsample your dataset by 5%.

B. Replace the missing values with the feature's mean.

C. Replace the missing values with a placeholder category indicating a missing value.

D. Move the rows with missing values to your validation dataset.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 128
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are an ML engineer on an agricultural research team working on a crop disease detection tool to detect leaf rust spots in images of crops to determine the presence of a disease. These spots, which can vary in shape and size, are correlated to the severity of the disease. You want to develop a solution that predicts the presence and severity of the disease with high accuracy. What should you do?

A. Create an object detection model that can localize the rust spots.

B. Develop an image segmentation ML model to locate the boundaries of the rust spots.

C. Develop a template matching algorithm using traditional computer vision libraries.

D. Develop an image classification ML model to predict the presence of the disease.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 129
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You have been asked to productionize a proof-of-concept ML model built using Keras. The model was trained in a Jupyter notebook on a data scientist's local machine. The notebook contains a cell that performs data validation and a cell that performs model analysis. You need to orchestrate the steps contained in the notebook and automate the execution of these steps for weekly retraining. You expect much more training data in the future. You want your solution to take advantage of managed services while minimizing cost. What should you do?

A. Move the Jupyter notebook to a Notebooks instance on the largest N2 machine type, and schedule the execution of the steps in the Notebooks instance using Cloud Scheduler.

B. Write the code as a TensorFlow Extended (TFX) pipeline orchestrated with Vertex AI Pipelines. Use standard TFX components for data validation and model analysis, and use Vertex AI Pipelines for model retraining.

C. Rewrite the steps in the Jupyter notebook as an Apache Spark job, and schedule the execution of the job on ephemeral Dataproc clusters using Cloud Scheduler.

D. Extract the steps contained in the Jupyter notebook as Python scripts, wrap each script in an Apache Airflow BashOperator, and run the resulting directed acyclic graph (DAG) in Cloud Composer.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 130
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are working on a system log anomaly detection model for a cybersecurity organization. You have developed the model using TensorFlow, and you plan to use it for real-time prediction. You need to create a Dataflow pipeline to ingest data via Pub/Sub and write the results to BigQuery. You want to minimize the serving latency as much as possible. What should you do?

A. Containerize the model prediction logic in Cloud Run, which is invoked by Dataflow.

B. Load the model directly into the Dataflow job as a dependency, and use it for prediction.

C. Deploy the model to a Vertex AI endpoint, and invoke this endpoint in the Dataflow job.

D. Deploy the model in a TFServing container on Google Kubernetes Engine, and invoke it in the Dataflow job.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 131
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are an ML engineer at a mobile gaming company. A data scientist on your team recently trained a TensorFlow model, and you are responsible for deploying this model into a mobile application. You discover that the inference latency of the current model doesn't meet production requirements. You need to reduce the inference time by 50%, and you are willing to accept a small decrease in model accuracy in order to reach the latency requirement. Without training a new model, which model optimization technique for reducing latency should you try first?

A. Weight pruning

B. Dynamic range quantization

C. Model distillation

D. Dimensionality reduction

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 132

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You work on a data science team at a bank and are creating an ML model to predict loan default risk. You have collected and cleaned hundreds of millions of records worth of training data in a BigQuery table, and you now want to develop and compare multiple models on this data using TensorFlow and Vertex AI. You want to minimize any bottlenecks during the data ingestion state while considering scalability. What should you do?

A. Use the BigQuery client library to load data into a dataframe, and use tf.data.Dataset.from_tensor_slices() to read it.

B. Export data to CSV files in Cloud Storage, and use tf.data.TextLineDataset() to read them.

C. Convert the data into TFRecords, and use tf.data.TFRecordDataset() to read them.

D. Use TensorFlow I/O's BigQuery Reader to directly read the data.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 133
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You have recently created a proof-of-concept (POC) deep learning model. You are satisfied with the overall architecture, but you need to determine the value for a couple of hyperparameters. You want to perform hyperparameter tuning on Vertex AI to determine both the appropriate embedding dimension for a categorical feature used by your model and the optimal learning rate. You configure the following settings:
• For the embedding dimension, you set the type to INTEGER with a minValue of 16 and maxValue of 64.
• For the learning rate, you set the type to DOUBLE with a minValue of 10e-05 and maxValue of 10e-02.

You are using the default Bayesian optimization tuning algorithm, and you want to maximize model accuracy. Training time is not a concern. How should you set the hyperparameter scaling for each hyperparameter and the maxParallelTrials?

A. Use UNIT_LINEAR_SCALE for the embedding dimension, UNIT_LOG_SCALE for the learning rate, and a large number of parallel trials.

B. Use UNIT_LINEAR_SCALE for the embedding dimension, UNIT_LOG_SCALE for the learning rate, and a small number of parallel trials.

C. Use UNIT_LOG_SCALE for the embedding dimension, UNIT_LINEAR_SCALE for the learning rate, and a large number of parallel trials.

D. Use UNIT_LOG_SCALE for the embedding dimension, UNIT_LINEAR_SCALE for the learning rate, and a small number of parallel trials.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 134
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are the Director of Data Science at a large company, and your Data Science team has recently begun using the Kubeflow Pipelines SDK to orchestrate their training pipelines. Your team is struggling to integrate their custom Python code into the Kubeflow Pipelines SDK. How should you instruct them to proceed in order to quickly integrate their code with the Kubeflow Pipelines SDK?

A. Use the func_to_container_op function to create custom components from the Python code.

B. Use the predefined components available in the Kubeflow Pipelines SDK to access Dataproc, and run the custom code there.

C. Package the custom Python code into Docker containers, and use the load_component_from_file function to import the containers into the pipeline.

D. Deploy the custom Python code to Cloud Functions, and use Kubeflow Pipelines to trigger the Cloud Function.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 135

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You work for the AI team of an automobile company, and you are developing a visual defect detection model using TensorFlow and Keras. To improve your model performance, you want to incorporate some image augmentation functions such as translation, cropping, and contrast tweaking. You randomly apply these functions to each training batch. You want to optimize your data processing pipeline for run time and compute resources utilization. What should you do?

A. Embed the augmentation functions dynamically in the tf.Data pipeline.

B. Embed the augmentation functions dynamically as part of Keras generators.

C. Use Dataflow to create all possible augmentations, and store them as TFRecords.

D. Use Dataflow to create the augmentations dynamically per training run, and stage them as TFRecords.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 136
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for an online publisher that delivers news articles to over 50 million readers. You have built an AI model that recommends content for the company's weekly newsletter. A recommendation is considered successful if the article is opened within two days of the newsletter's published date and the user remains on the page for at least one minute.

All the information needed to compute the success metric is available in BigQuery and is updated hourly. The model is trained on eight weeks of data, on average its performance degrades below the acceptable baseline after five weeks, and training time is 12 hours. You want to ensure that the model's performance is above the acceptable baseline while minimizing cost. How should you monitor the model to determine when retraining is necessary?

A. Use Vertex AI Model Monitoring to detect skew of the input features with a sample rate of 100% and a monitoring frequency of two days.

B. Schedule a cron job in Cloud Tasks to retrain the model every week before the newsletter is created.

C. Schedule a weekly query in BigQuery to compute the success metric.

D. Schedule a daily Dataflow job in Cloud Composer to compute the success metric.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 137
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You deployed an ML model into production a year ago. Every month, you collect all raw requests that were sent to your model prediction service during the previous month. You send a subset of these requests to a human labeling service to evaluate your model's performance. After a year, you notice that your model's performance sometimes degrades significantly after a month, while other times it takes several months to notice any decrease in performance. The labeling service is costly, but you also need to avoid large performance degradations. You want to determine how often you should retrain your model to maintain a high level of performance while minimizing cost. What should you do?

A. Train an anomaly detection model on the training dataset, and run all incoming requests through this model. If an anomaly is detected, send the most recent serving data to the labeling service.

B. Identify temporal patterns in your model's performance over the previous year. Based on these patterns, create a schedule for sending serving data to the labeling service for the next year.

C. Compare the cost of the labeling service with the lost revenue due to model performance degradation over the past year. If the lost revenue is greater than the cost of the labeling service, increase the frequency of model retraining; otherwise, decrease the model retraining frequency.

D. Run training-serving skew detection batch jobs every few days to compare the aggregate statistics of the features in the training dataset with recent serving data. If skew is detected, send the most recent serving data to the labeling service.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 138
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for a company that manages a ticketing platform for a large chain of cinemas. Customers use a mobile app to search for movies they're interested in and purchase tickets in the app. Ticket purchase requests are sent to Pub/Sub and are processed with a Dataflow streaming pipeline configured to conduct the following steps:
1. Check for availability of the movie tickets at the selected cinema.
2. Assign the ticket price and accept payment.
3. Reserve the tickets at the selected cinema.
4. Send successful purchases to your database.

Each step in this process has low latency requirements (less than 50 milliseconds). You have developed a logistic regression model with BigQuery ML that predicts whether offering a promo code for free popcorn increases the chance of a ticket purchase, and this prediction should be added to the ticket purchase process. You want to identify the simplest way to deploy this model to production while adding minimal latency. What should you do?

A. Run batch inference with BigQuery ML every five minutes on each new set of tickets issued.

B. Export your model in TensorFlow format, and add a tfx_bsl.public.beam.RunInference step to the Dataflow pipeline.

C. Export your model in TensorFlow format, deploy it on Vertex AI, and query the prediction endpoint from your streaming pipeline.

D. Convert your model with TensorFlow Lite (TFLite), and add it to the mobile app so that the promo code and the incoming request arrive together in Pub/Sub.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 139
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work on a team in a data center that is responsible for server maintenance. Your management team wants you to build a predictive maintenance solution that uses monitoring data to detect potential server failures. Incident data has not been labeled yet. What should you do first?

A. Train a time-series model to predict the machines' performance values. Configure an alert if a machine's actual performance values significantly differ from the predicted performance values.

B. Develop a simple heuristic (e.g., based on z-score) to label the machines' historical performance data. Use this heuristic to monitor server performance in real time.

C. Develop a simple heuristic (e.g., based on z-score) to label the machines' historical performance data. Train a model to predict anomalies based on this labeled dataset.

D. Hire a team of qualified analysts to review and label the machines' historical performance data. Train a model based on this manually labeled dataset.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 140
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for a retailer that sells clothes to customers around the world. You have been tasked with ensuring that ML models are built in a secure manner. Specifically, you need to protect sensitive customer data that might be used in the models. You have identified four fields containing sensitive data that are being used by your data science team: AGE, IS_EXISTING_CUSTOMER, LATITUDE_LONGITUDE, and SHIRT_SIZE. What should you do with the data before it is made available to the data science team for training purposes?

A. Tokenize all of the fields using hashed dummy values to replace the real values.

B. Use principal component analysis (PCA) to reduce the four sensitive fields to one PCA vector.

C. Coarsen the data by putting AGE into quantiles and rounding LATITUDE_LONGTTUDE into single precision. The other two fields are already as coarse as possible.

D. Remove all sensitive data fields, and ask the data science team to build their models using non-sensitive data.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 141
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for a magazine publisher and have been tasked with predicting whether customers will cancel their annual subscription. In your exploratory data analysis, you find that 90% of individuals renew their subscription every year, and only 10% of individuals cancel their subscription. After training a NN Classifier, your model predicts those who cancel their subscription with 99% accuracy and predicts those who renew their subscription with 82% accuracy. How should you interpret these results?

A. This is not a good result because the model should have a higher accuracy for those who renew their subscription than for those who cancel their subscription.

B. This is not a good result because the model is performing worse than predicting that people will always renew their subscription.

C. This is a good result because predicting those who cancel their subscription is more difficult, since there is less data for this group.

D. This is a good result because the accuracy across both groups is greater than 80%.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 142

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You have built a model that is trained on data stored in Parquet files. You access the data through a Hive table hosted on Google Cloud. You preprocessed these data with PySpark and exported it as a CSV file into Cloud Storage. After preprocessing, you execute additional steps to train and evaluate your model. You want to parametrize this model training in Kubeflow Pipelines. What should you do?

A. Remove the data transformation step from your pipeline.

B. Containerize the PySpark transformation step, and add it to your pipeline.

C. Add a ContainerOp to your pipeline that spins a Dataproc cluster, runs a transformation, and then saves the transformed data in Cloud Storage.

D. Deploy Apache Spark at a separate node pool in a Google Kubernetes Engine cluster. Add a ContainerOp to your pipeline that invokes a corresponding transformation job for this Spark instance.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 143
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You have developed an ML model to detect the sentiment of users' posts on your company's social media page to identify outages or bugs. You are using Dataflow to provide real-time predictions on data ingested from Pub/Sub. You plan to have multiple training iterations for your model and keep the latest two versions live after every run. You want to split the traffic between the versions in an 80:20 ratio, with the newest model getting the majority of the traffic. You want to keep the pipeline as simple as possible, with minimal management required. What should you do?

A. Deploy the models to a Vertex AI endpoint using the traffic-split=0=80, PREVIOUS_MODEL_ID=20 configuration.

B. Wrap the models inside an App Engine application using the --splits PREVIOUS_VERSION=0.2, NEW_VERSION=0.8 configuration

C. Wrap the models inside a Cloud Run container using the REVISION1=20, REVISION2=80 revision configuration.

D. Implement random splitting in Dataflow using beam.Partition() with a partition function calling a Vertex AI endpoint.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 144

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are developing an image recognition model using PyTorch based on ResNet50 architecture. Your code is working fine on your local laptop on a small subsample. Your full dataset has 200k labeled images. You want to quickly scale your training workload while minimizing cost. You plan to use 4 V100 GPUs. What should you do?

A. Create a Google Kubernetes Engine cluster with a node pool that has 4 V100 GPUs. Prepare and submit a TFJob operator to this node pool.

B. Create a Vertex AI Workbench user-managed notebooks instance with 4 V100 GPUs, and use it to train your model.

C. Package your code with Setuptools, and use a pre-built container. Train your model with Vertex AI using a custom tier that contains the required GPUs.

D. Configure a Compute Engine VM with all the dependencies that launches the training. Train your model with Vertex AI using a custom tier that contains the required GPUs.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 145

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You have trained a DNN regressor with TensorFlow to predict housing prices using a set of predictive features. Your default precision is tf.float64, and you use a standard TensorFlow estimator:

```
estimator = tf.estimator.DNNRegressor(
      feature_columns=[YOUR_LIST_OF_FEATURES],
      hidden_units=[1024, 512, 256],
      dropout=None)
```

Your model performs well, but just before deploying it to production, you discover that your current serving latency is 10ms @ 90 percentile and you currently serve on CPUs. Your production requirements expect a model latency of 8ms @ 90 percentile. You're willing to accept a small decrease in performance in order to reach the latency requirement.

Therefore your plan is to improve latency while evaluating how much the model's prediction decreases. What should you first try to quickly lower the serving latency?

A. Switch from CPU to GPU serving.

B. Apply quantization to your SavedModel by reducing the floating point precision to tf.float16.

C. Increase the dropout rate to 0.8 and retrain your model.

D. Increase the dropout rate to 0.8 in _PREDICT mode by adjusting the TensorFlow Serving parameters.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 146
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work on the data science team at a manufacturing company. You are reviewing the company's historical sales data, which has hundreds of millions of records. For your exploratory data analysis, you need to calculate descriptive statistics such as mean, median, and mode; conduct complex statistical tests for hypothesis testing; and plot variations of the features over time. You want to use as much of the sales data as possible in your analyses while minimizing computational resources. What should you do?

A. Visualize the time plots in Google Data Studio. Import the dataset into Vertex AI Workbench user-managed notebooks. Use this data to calculate the descriptive statistics and run the statistical analyses.

B. Spin up a Vertex AI Workbench user-managed notebooks instance and import the dataset. Use this data to create statistical and visual analyses.

C. Use BigQuery to calculate the descriptive statistics. Use Vertex AI Workbench user-managed notebooks to visualize the time plots and run the statistical analyses.

D. Use BigQuery to calculate the descriptive statistics, and use Google Data Studio to visualize the time plots. Use Vertex AI Workbench user-managed notebooks to run the statistical analyses.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 147
Topic #: 1
[All Professional Machine Learning Engineer Questions]

Your data science team needs to rapidly experiment with various features, model architectures, and hyperparameters. They need to track the accuracy metrics for various experiments and use an API to query the metrics over time. What should they use to track and report their experiments while minimizing manual effort?

A. Use Vertex AI Pipelines to execute the experiments. Query the results stored in MetadataStore using the Vertex AI API.

B. Use Vertex AI Training to execute the experiments. Write the accuracy metrics to BigQuery, and query the results using the BigQuery API.

C. Use Vertex AI Training to execute the experiments. Write the accuracy metrics to Cloud Monitoring, and query the results using the Monitoring API.

D. Use Vertex AI Workbench user-managed notebooks to execute the experiments. Collect the results in a shared Google Sheets file, and query the results using the Google Sheets API.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 148
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are training an ML model using data stored in BigQuery that contains several values that are considered Personally Identifiable Information (PII). You need to reduce the sensitivity of the dataset before training your model. Every column is critical to your model. How should you proceed?

A. Using Dataflow, ingest the columns with sensitive data from BigQuery, and then randomize the values in each sensitive column.

B. Use the Cloud Data Loss Prevention (DLP) API to scan for sensitive data, and use Dataflow with the DLP API to encrypt sensitive values with Format Preserving Encryption.

C. Use the Cloud Data Loss Prevention (DLP) API to scan for sensitive data, and use Dataflow to replace all sensitive data by using the encryption algorithm AES-256 with a salt.

D. Before training, use BigQuery to select only the columns that do not contain sensitive data. Create an authorized view of the data so that sensitive values cannot be accessed by unauthorized individuals.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 149
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You recently deployed an ML model. Three months after deployment, you notice that your model is underperforming on certain subgroups, thus potentially leading to biased results. You suspect that the inequitable performance is due to class imbalances in the training data, but you cannot collect more data. What should you do? (Choose two.)

A. Remove training examples of high-performing subgroups, and retrain the model.

B. Add an additional objective to penalize the model more for errors made on the minority class, and retrain the model

C. Remove the features that have the highest correlations with the majority class.

D. Upsample or reweight your existing training data, and retrain the model

E. Redeploy the model, and provide a label explaining the model's behavior to users.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 150

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are working on a binary classification ML algorithm that detects whether an image of a classified scanned document contains a company's logo. In the dataset, 96% of examples don't have the logo, so the dataset is very skewed. Which metric would give you the most confidence in your model?

A. Precision

B. Recall

C. RMSE

D. F1 score

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 151

Topic #: 1

[All Professional Machine Learning Engineer Questions]

While running a model training pipeline on Vertex AI, you discover that the evaluation step is failing because of an out-of-memory error. You are currently using TensorFlow Model Analysis (TFMA) with a standard Evaluator TensorFlow Extended (TFX) pipeline component for the evaluation step. You want to stabilize the pipeline without downgrading the evaluation quality while minimizing infrastructure overhead. What should you do?

A. Include the flag -runner=DataflowRunner in beam_pipeline_args to run the evaluation step on Dataflow.

B. Move the evaluation step out of your pipeline and run it on custom Compute Engine VMs with sufficient memory.

C. Migrate your pipeline to Kubeflow hosted on Google Kubernetes Engine, and specify the appropriate node parameters for the evaluation step.

D. Add tfma.MetricsSpec () to limit the number of metrics in the evaluation step.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 152

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are developing an ML model using a dataset with categorical input variables. You have randomly split half of the data into training and test sets. After applying one-hot encoding on the categorical variables in the training set, you discover that one categorical variable is missing from the test set. What should you do?

A. Use sparse representation in the test set.

B. Randomly redistribute the data, with 70% for the training set and 30% for the test set

C. Apply one-hot encoding on the categorical variables in the test data

D. Collect more data representing all categories

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 153

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You work for a bank and are building a random forest model for fraud detection. You have a dataset that includes transactions, of which 1% are identified as fraudulent. Which data transformation strategy would likely improve the performance of your classifier?

A. Modify the target variable using the Box-Cox transformation.

B. Z-normalize all the numeric features.

C. Oversample the fraudulent transaction 10 times.

D. Log transform all numeric features.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 154

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are developing a classification model to support predictions for your company's various products. The dataset you were given for model development has class imbalance You need to minimize false positives and false negatives What evaluation metric should you use to properly train the model?

A. F1 score

B. Recall

C. Accuracy

D. Precision

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 155

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are training an object detection machine learning model on a dataset that consists of three million X-ray images, each roughly 2 GB in size. You are using Vertex AI Training to run a custom training application on a Compute Engine instance with 32-cores, 128 GB of RAM, and 1 NVIDIA P100 GPU. You notice that model training is taking a very long time. You want to decrease training time without sacrificing model performance. What should you do?

A. Increase the instance memory to 512 GB, and increase the batch size.

B. Replace the NVIDIA P100 GPU with a K80 GPU in the training job.

C. Enable early stopping in your Vertex AI Training job.

D. Use the tf.distribute.Strategy API and run a distributed training job.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 156
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You need to build classification workflows over several structured datasets currently stored in BigQuery. Because you will be performing the classification several times, you want to complete the following steps without writing code: exploratory data analysis, feature selection, model building, training, and hyperparameter tuning and serving. What should you do?

A. Train a TensorFlow model on Vertex AI.

B. Train a classification Vertex AutoML model.

C. Run a logistic regression job on BigQuery ML.

D. Use scikit-learn in Vertex AI Workbench user-managed notebooks with pandas library.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 157
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You recently developed a deep learning model. To test your new model, you trained it for a few epochs on a large dataset. You observe that the training and validation losses barely changed during the training run. You want to quickly debug your model. What should you do first?

A. Verify that your model can obtain a low loss on a small subset of the dataset

B. Add handcrafted features to inject your domain knowledge into the model

C. Use the Vertex AI hyperparameter tuning service to identify a better learning rate

D. Use hardware accelerators and train your model for more epochs

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 158
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are a data scientist at an industrial equipment manufacturing company. You are developing a regression model to estimate the power consumption in the company's manufacturing plants based on sensor data collected from all of the plants. The sensors collect tens of millions of records every day. You need to schedule daily training runs for your model that use all the data collected up to the current date. You want your model to scale smoothly and require minimal development work. What should you do?

A. Develop a custom TensorFlow regression model, and optimize it using Vertex AI Training.

B. Develop a regression model using BigQuery ML.

C. Develop a custom scikit-learn regression model, and optimize it using Vertex AI Training.

D. Develop a custom PyTorch regression model, and optimize it using Vertex AI Training.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 159
Topic #: 1
[All Professional Machine Learning Engineer Questions]

Your organization manages an online message board. A few months ago, you discovered an increase in toxic language and bullying on the message board. You deployed an automated text classifier that flags certain comments as toxic or harmful. Now some users are reporting that benign comments referencing their religion are being misclassified as abusive. Upon further inspection, you find that your classifier's false positive rate is higher for comments that reference certain underrepresented religious groups. Your team has a limited budget and is already overextended. What should you do?

A. Add synthetic training data where those phrases are used in non-toxic ways.

B. Remove the model and replace it with human moderation.

C. Replace your model with a different text classifier.

D. Raise the threshold for comments to be considered toxic or harmful.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 160
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for a magazine distributor and need to build a model that predicts which customers will renew their subscriptions for the upcoming year. Using your company's historical data as your training set, you created a TensorFlow model and deployed it to Vertex AI. You need to determine which customer attribute has the most predictive power for each prediction served by the model. What should you do?

A. Stream prediction results to BigQuery. Use BigQuery's CORR(X1, X2) function to calculate the Pearson correlation coefficient between each feature and the target variable.

B. Use Vertex Explainable AI. Submit each prediction request with the explain' keyword to retrieve feature attributions using the sampled Shapley method.

C. Use Vertex AI Workbench user-managed notebooks to perform a Lasso regression analysis on your model, which will eliminate features that do not provide a strong signal.

D. Use the What-If tool in Google Cloud to determine how your model will perform when individual features are excluded. Rank the feature importance in order of those that caused the most significant performance drop when removed from the model.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 161
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are an ML engineer at a manufacturing company. You are creating a classification model for a predictive maintenance use case. You need to predict whether a crucial machine will fail in the next three days so that the repair crew has enough time to fix the machine before it breaks. Regular maintenance of the machine is relatively inexpensive, but a failure would be very costly. You have trained several binary classifiers to predict whether the machine will fail, where a prediction of 1 means that the ML model predicts a failure.

You are now evaluating each model on an evaluation dataset. You want to choose a model that prioritizes detection while ensuring that more than 50% of the maintenance jobs triggered by your model address an imminent machine failure. Which model should you choose?

A. The model with the highest area under the receiver operating characteristic curve (AUC ROC) and precision greater than 0.5

B. The model with the lowest root mean squared error (RMSE) and recall greater than 0.5.

C. The model with the highest recall where precision is greater than 0.5.

D. The model with the highest precision where recall is greater than 0.5.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 162
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You built a custom ML model using scikit-learn. Training time is taking longer than expected. You decide to migrate your model to Vertex AI Training, and you want to improve the model's training time. What should you try out first?

A. Train your model in a distributed mode using multiple Compute Engine VMs.

B. Train your model using Vertex AI Training with CPUs.

C. Migrate your model to TensorFlow, and train it using Vertex AI Training.

D. Train your model using Vertex AI Training with GPUs.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 163
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are an ML engineer at a retail company. You have built a model that predicts a coupon to offer an ecommerce customer at checkout based on the items in their cart. When a customer goes to checkout, your serving pipeline, which is hosted on Google Cloud, joins the customer's existing cart with a row in a BigQuery table that contains the customers' historic purchase behavior and uses that as the model's input. The web team is reporting that your model is returning predictions too slowly to load the coupon offer with the rest of the web page. How should you speed up your model's predictions?

A. Attach an NVIDIA P100 GPU to your deployed model's instance.

B. Use a low latency database for the customers' historic purchase behavior.

C. Deploy your model to more instances behind a load balancer to distribute traffic.

D. Create a materialized view in BigQuery with the necessary data for predictions.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 164
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for a small company that has deployed an ML model with autoscaling on Vertex AI to serve online predictions in a production environment. The current model receives about 20 prediction requests per hour with an average response time of one second. You have retrained the same model on a new batch of data, and now you are canary testing it, sending ~10% of production traffic to the new model. During this canary test, you notice that prediction requests for your new model are taking between 30 and 180 seconds to complete. What should you do?

A. Submit a request to raise your project quota to ensure that multiple prediction services can run concurrently.

B. Turn off auto-scaling for the online prediction service of your new model. Use manual scaling with one node always available.

C. Remove your new model from the production environment. Compare the new model and existing model codes to identify the cause of the performance bottleneck.

D. Remove your new model from the production environment. For a short trial period, send all incoming prediction requests to BigQuery. Request batch predictions from your new model, and then use the Data Labeling Service to validate your model's performance before promoting it to production.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 165
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You want to train an AutoML model to predict house prices by using a small public dataset stored in BigQuery. You need to prepare the data and want to use the simplest, most efficient approach. What should you do?

A. Write a query that preprocesses the data by using BigQuery and creates a new table. Create a Vertex AI managed dataset with the new table as the data source.

B. Use Dataflow to preprocess the data. Write the output in TFRecord format to a Cloud Storage bucket.

C. Write a query that preprocesses the data by using BigQuery. Export the query results as CSV files, and use those files to create a Vertex AI managed dataset.

D. Use a Vertex AI Workbench notebook instance to preprocess the data by using the pandas library. Export the data as CSV files, and use those files to create a Vertex AI managed dataset.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 166
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You developed a Vertex AI ML pipeline that consists of preprocessing and training steps and each set of steps runs on a separate custom Docker image. Your organization uses GitHub and GitHub Actions as CI/CD to run unit and integration tests. You need to automate the model retraining workflow so that it can be initiated both manually and when a new version of the code is merged in the main branch. You want to minimize the steps required to build the workflow while also allowing for maximum flexibility. How should you configure the CI/CD workflow?

A. Trigger a Cloud Build workflow to run tests, build custom Docker images, push the images to Artifact Registry, and launch the pipeline in Vertex AI Pipelines.

B. Trigger GitHub Actions to run the tests, launch a job on Cloud Run to build custom Docker images, push the images to Artifact Registry, and launch the pipeline in Vertex AI Pipelines.

C. Trigger GitHub Actions to run the tests, build custom Docker images, push the images to Artifact Registry, and launch the pipeline in Vertex AI Pipelines.

D. Trigger GitHub Actions to run the tests, launch a Cloud Build workflow to build custom Docker images, push the images to Artifact Registry, and launch the pipeline in Vertex AI Pipelines.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 167
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are working with a dataset that contains customer transactions. You need to build an ML model to predict customer purchase behavior. You plan to develop the model in BigQuery ML, and export it to Cloud Storage for online prediction. You notice that the input data contains a few categorical features, including product category and payment method. You want to deploy the model as quickly as possible. What should you do?

A. Use the TRANSFORM clause with the ML.ONE_HOT_ENCODER function on the categorical features at model creation and select the categorical and non-categorical features.

B. Use the ML.ONE_HOT_ENCODER function on the categorical features and select the encoded categorical features and non-categorical features as inputs to create your model.

C. Use the CREATE MODEL statement and select the categorical and non-categorical features.

D. Use the ML.MULTI_HOT_ENCODER function on the categorical features, and select the encoded categorical features and non-categorical features as inputs to create your model.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 168

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You need to develop an image classification model by using a large dataset that contains labeled images in a Cloud Storage bucket. What should you do?

A. Use Vertex AI Pipelines with the Kubeflow Pipelines SDK to create a pipeline that reads the images from Cloud Storage and trains the model.

B. Use Vertex AI Pipelines with TensorFlow Extended (TFX) to create a pipeline that reads the images from Cloud Storage and trains the model.

C. Import the labeled images as a managed dataset in Vertex AI and use AutoML to train the model.

D. Convert the image dataset to a tabular format using Dataflow Load the data into BigQuery and use BigQuery ML to train the model.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 169
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are developing a model to detect fraudulent credit card transactions. You need to prioritize detection, because missing even one fraudulent transaction could severely impact the credit card holder. You used AutoML to tram a model on users' profile information and credit card transaction data After training the initial model, you notice that the model is failing to detect many fraudulent transactions. How should you adjust the training parameters in AutoML to improve model performance? (Choose two.)

A. Increase the score threshold

B. Decrease the score threshold.

C. Add more positive examples to the training set

D. Add more negative examples to the training set

E. Reduce the maximum number of node hours for training

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 170

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You need to deploy a scikit-leam classification model to production. The model must be able to serve requests 24/7, and you expect millions of requests per second to the production application from 8 am to 7 pm. You need to minimize the cost of deployment. What should you do?

A. Deploy an online Vertex AI prediction endpoint. Set the max replica count to 1

B. Deploy an online Vertex AI prediction endpoint. Set the max replica count to 100

C. Deploy an online Vertex AI prediction endpoint with one GPU per replica. Set the max replica count to 1

D. Deploy an online Vertex AI prediction endpoint with one GPU per replica. Set the max replica count to 100

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 171

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You work with a team of researchers to develop state-of-the-art algorithms for financial analysis. Your team develops and debugs complex models in TensorFlow. You want to maintain the ease of debugging while also reducing the model training time. How should you set up your training environment?

A. Configure a v3-8 TPU VM. SSH into the VM to train and debug the model.

B. Configure a v3-8 TPU node. Use Cloud Shell to SSH into the Host VM to train and debug the model.

C. Configure a n1 -standard-4 VM with 4 NVIDIA P100 GPUs. SSH into the VM and use ParameterServerStraregv to train the model.

D. Configure a n1-standard-4 VM with 4 NVIDIA P100 GPUs. SSH into the VM and use MultiWorkerMirroredStrategy to train the model.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 172
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You created an ML pipeline with multiple input parameters. You want to investigate the tradeoffs between different parameter combinations. The parameter options are
• Input dataset
• Max tree depth of the boosted tree regressor
• Optimizer learning rate

You need to compare the pipeline performance of the different parameter combinations measured in F1 score, time to train, and model complexity. You want your approach to be reproducible, and track all pipeline runs on the same platform. What should you do?

A. 1. Use BigQueryML to create a boosted tree regressor, and use the hyperparameter tuning capability.
2. Configure the hyperparameter syntax to select different input datasets: max tree depths, and optimizer learning rates. Choose the grid search option.

B. 1. Create a Vertex AI pipeline with a custom model training job as part of the pipeline. Configure the pipeline's parameters to include those you are investigating.
2. In the custom training step, use the Bayesian optimization method with F1 score as the target to maximize.

C. 1. Create a Vertex AI Workbench notebook for each of the different input datasets.
2. In each notebook, run different local training jobs with different combinations of the max tree depth and optimizer learning rate parameters.
3. After each notebook finishes, append the results to a BigQuery table.

D. 1. Create an experiment in Vertex AI Experiments.
2. Create a Vertex AI pipeline with a custom model training job as part of the pipeline. Configure the pipeline's parameters to include those you are investigating.
3. Submit multiple runs to the same experiment, using different values for the parameters.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 173
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You received a training-serving skew alert from a Vertex AI Model Monitoring job running in production. You retrained the model with more recent training data, and deployed it back to the Vertex AI endpoint, but you are still receiving the same alert. What should you do?

A. Update the model monitoring job to use a lower sampling rate.

B. Update the model monitoring job to use the more recent training data that was used to retrain the model.

C. Temporarily disable the alert. Enable the alert again after a sufficient amount of new production traffic has passed through the Vertex AI endpoint.

D. Temporarily disable the alert until the model can be retrained again on newer training data. Retrain the model again after a sufficient amount of new production traffic has passed through the Vertex AI endpoint.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 174

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You developed a custom model by using Vertex AI to forecast the sales of your company's products based on historical transactional data. You anticipate changes in the feature distributions and the correlations between the features in the near future. You also expect to receive a large volume of prediction requests. You plan to use Vertex AI Model Monitoring for drift detection and you want to minimize the cost. What should you do?

A. Use the features for monitoring. Set a monitoring-frequency value that is higher than the default.

B. Use the features for monitoring. Set a prediction-sampling-rate value that is closer to 1 than 0.

C. Use the features and the feature attributions for monitoring. Set a monitoring-frequency value that is lower than the default.

D. Use the features and the feature attributions for monitoring. Set a prediction-sampling-rate value that is closer to 0 than 1.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 175
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You have recently trained a scikit-learn model that you plan to deploy on Vertex AI. This model will support both online and batch prediction. You need to preprocess input data for model inference. You want to package the model for deployment while minimizing additional code. What should you do?

A. 1. Upload your model to the Vertex AI Model Registry by using a prebuilt scikit-ieam prediction container.
2. Deploy your model to Vertex AI Endpoints, and create a Vertex AI batch prediction job that uses the instanceConfig.instanceType setting to transform your input data.

B. 1. Wrap your model in a custom prediction routine (CPR). and build a container image from the CPR local model.
2. Upload your scikit learn model container to Vertex AI Model Registry.
3. Deploy your model to Vertex AI Endpoints, and create a Vertex AI batch prediction job

C. 1. Create a custom container for your scikit learn model.
2. Define a custom serving function for your model.
3. Upload your model and custom container to Vertex AI Model Registry.
4. Deploy your model to Vertex AI Endpoints, and create a Vertex AI batch prediction job.

D. 1. Create a custom container for your scikit learn model.
2. Upload your model and custom container to Vertex AI Model Registry.
3. Deploy your model to Vertex AI Endpoints, and create a Vertex AI batch prediction job that uses the instanceConfig.instanceType setting to transform your input data.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 176
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for a food product company. Your company's historical sales data is stored in BigQuery.You need to use Vertex AI's custom training service to train multiple TensorFlow models that read the data from BigQuery and predict future sales. You plan to implement a data preprocessing algorithm that performs mm-max scaling and bucketing on a large number of features before you start experimenting with the models. You want to minimize preprocessing time, cost, and development effort. How should you configure this workflow?

A. Write the transformations into Spark that uses the spark-bigquery-connector, and use Dataproc to preprocess the data.

B. Write SQL queries to transform the data in-place in BigQuery.

C. Add the transformations as a preprocessing layer in the TensorFlow models.

D. Create a Dataflow pipeline that uses the BigQueryIO connector to ingest the data, process it, and write it back to BigQuery.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 177
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You have created a Vertex AI pipeline that includes two steps. The first step preprocesses 10 TB data completes in about 1 hour, and saves the result in a Cloud Storage bucket. The second step uses the processed data to train a model. You need to update the model's code to allow you to test different algorithms. You want to reduce pipeline execution time and cost while also minimizing pipeline changes. What should you do?

A. Add a pipeline parameter and an additional pipeline step. Depending on the parameter value, the pipeline step conducts or skips data preprocessing, and starts model training.

B. Create another pipeline without the preprocessing step, and hardcode the preprocessed Cloud Storage file location for model training.

C. Configure a machine with more CPU and RAM from the compute-optimized machine family for the data preprocessing step.

D. Enable caching for the pipeline job, and disable caching for the model training step.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 178
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for a bank. You have created a custom model to predict whether a loan application should be flagged for human review. The input features are stored in a BigQuery table. The model is performing well, and you plan to deploy it to production. Due to compliance requirements the model must provide explanations for each prediction. You want to add this functionality to your model code with minimal effort and provide explanations that are as accurate as possible. What should you do?

A. Create an AutoML tabular model by using the BigQuery data with integrated Vertex Explainable AI.

B. Create a BigQuery ML deep neural network model and use the ML.EXPLAIN_PREDICT method with the num_integral_steps parameter.

C. Upload the custom model to Vertex AI Model Registry and configure feature-based attribution by using sampled Shapley with input baselines.

D. Update the custom serving container to include sampled Shapley-based explanations in the prediction outputs.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 179
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You recently used XGBoost to train a model in Python that will be used for online serving. Your model prediction service will be called by a backend service implemented in Golang running on a Google Kubernetes Engine (GKE) cluster. Your model requires pre and postprocessing steps. You need to implement the processing steps so that they run at serving time. You want to minimize code changes and infrastructure maintenance, and deploy your model into production as quickly as possible. What should you do?

A. Use FastAPI to implement an HTTP server. Create a Docker image that runs your HTTP server, and deploy it on your organization's GKE cluster.

B. Use FastAPI to implement an HTTP server. Create a Docker image that runs your HTTP server, Upload the image to Vertex AI Model Registry and deploy it to a Vertex AI endpoint.

C. Use the Predictor interface to implement a custom prediction routine. Build the custom container, upload the container to Vertex AI Model Registry and deploy it to a Vertex AI endpoint.

D. Use the XGBoost prebuilt serving container when importing the trained model into Vertex AI. Deploy the model to a Vertex AI endpoint. Work with the backend engineers to implement the pre- and postprocessing steps in the Golang backend service.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 180
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You recently deployed a pipeline in Vertex AI Pipelines that trains and pushes a model to a Vertex AI endpoint to serve real-time traffic. You need to continue experimenting and iterating on your pipeline to improve model performance. You plan to use Cloud Build for CI/CD You want to quickly and easily deploy new pipelines into production, and you want to minimize the chance that the new pipeline implementations will break in production. What should you do?

A. Set up a CI/CD pipeline that builds and tests your source code. If the tests are successful, use the Google. Cloud console to upload the built container to Artifact Registry and upload the compiled pipeline to Vertex AI Pipelines.

B. Set up a CI/CD pipeline that builds your source code and then deploys built artifacts into a pre-production environment. Run unit tests in the pre-production environment. If the tests are successful deploy the pipeline to production.

C. Set up a CI/CD pipeline that builds and tests your source code and then deploys built artifacts into a pre-production environment. After a successful pipeline run in the pre-production environment, deploy the pipeline to production.

D. Set up a CI/CD pipeline that builds and tests your source code and then deploys built artifacts into a pre-production environment. After a successful pipeline run in the pre-production environment, rebuild the source code and deploy the artifacts to production.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 181
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for a bank with strict data governance requirements. You recently implemented a custom model to detect fraudulent transactions. You want your training code to download internal data by using an API endpoint hosted in your project's network. You need the data to be accessed in the most secure way, while mitigating the risk of data exfiltration. What should you do?

A. Enable VPC Service Controls for peerings, and add Vertex AI to a service perimeter.

B. Create a Cloud Run endpoint as a proxy to the data. Use Identity and Access Management (IAM) authentication to secure access to the endpoint from the training job.

C. Configure VPC Peering with Vertex AI, and specify the network of the training job.

D. Download the data to a Cloud Storage bucket before calling the training job.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 182
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are deploying a new version of a model to a production Vertex AI endpoint that is serving traffic. You plan to direct all user traffic to the new model. You need to deploy the model with minimal disruption to your application. What should you do?

A. 1. Create a new endpoint
2. Create a new model. Set it as the default version. Upload the model to Vertex AI Model Registry
3. Deploy the new model to the new endpoint
4. Update Cloud DNS to point to the new endpoint

B. 1. Create a new endpoint
2. Create a new model. Set the parentModel parameter to the model ID of the currently deployed model and set it as the default version. Upload the model to Vertex AI Model Registry
3. Deploy the new model to the new endpoint, and set the new model to 100% of the traffic.

C. 1. Create a new model. Set the parentModel parameter to the model ID of the currently deployed model. Upload the model to Vertex AI Model Registry.
2. Deploy the new model to the existing endpoint, and set the new model to 100% of the traffic

D. 1. Create a new model. Set it as the default version. Upload the model to Vertex AI Model Registry
2. Deploy the new model to the existing endpoint

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 183

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are training an ML model on a large dataset. You are using a TPU to accelerate the training process. You notice that the training process is taking longer than expected. You discover that the TPU is not reaching its full capacity. What should you do?

A. Increase the learning rate

B. Increase the number of epochs

C. Decrease the learning rate

D. Increase the batch size

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 184

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You work for a retail company. You have a managed tabular dataset in Vertex AI that contains sales data from three different stores. The dataset includes several features, such as store name and sale timestamp. You want to use the data to train a model that makes sales predictions for a new store that will open soon. You need to split the data between the training, validation, and test sets. What approach should you use to split the data?

A. Use Vertex AI manual split, using the store name feature to assign one store for each set

B. Use Vertex AI default data split

C. Use Vertex AI chronological split, and specify the sales timestamp feature as the time variable

D. Use Vertex AI random split, assigning 70% of the rows to the training set, 10% to the validation set, and 20% to the test set

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 185
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You have developed a BigQuery ML model that predicts customer chum, and deployed the model to Vertex AI Endpoints. You want to automate the retraining of your model by using minimal additional code when model feature values change. You also want to minimize the number of times that your model is retrained to reduce training costs. What should you do?

A. 1 Enable request-response logging on Vertex AI Endpoints
2. Schedule a TensorFlow Data Validation job to monitor prediction drift
3. Execute model retraining if there is significant distance between the distributions

B. 1. Enable request-response logging on Vertex AI Endpoints
2. Schedule a TensorFlow Data Validation job to monitor training/serving skew
3. Execute model retraining if there is significant distance between the distributions

C. 1. Create a Vertex AI Model Monitoring job configured to monitor prediction drift
2. Configure alert monitoring to publish a message to a Pub/Sub queue when a monitoring alert is detected
3. Use a Cloud Function to monitor the Pub/Sub queue, and trigger retraining in BigQuery

D. 1. Create a Vertex AI Model Monitoring job configured to monitor training/serving skew
2. Configure alert monitoring to publish a message to a Pub/Sub queue when a monitoring alert is detected
3. Use a Cloud Function to monitor the Pub/Sub queue, and trigger retraining in BigQuery

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 186
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You have been tasked with deploying prototype code to production. The feature engineering code is in PySpark and runs on Dataproc Serverless. The model training is executed by using a Vertex AI custom training job. The two steps are not connected, and the model training must currently be run manually after the feature engineering step finishes. You need to create a scalable and maintainable production process that runs end-to-end and tracks the connections between steps. What should you do?

A. Create a Vertex AI Workbench notebook. Use the notebook to submit the Dataproc Serverless feature engineering job. Use the same notebook to submit the custom model training job. Run the notebook cells sequentially to tie the steps together end-to-end.

B. Create a Vertex AI Workbench notebook. Initiate an Apache Spark context in the notebook and run the PySpark feature engineering code. Use the same notebook to run the custom model training job in TensorFlow. Run the notebook cells sequentially to tie the steps together end-to-end.

C. Use the Kubeflow pipelines SDK to write code that specifies two components:
- The first is a Dataproc Serverless component that launches the feature engineering job
- The second is a custom component wrapped in the create_custom_training_job_from_component utility that launches the custom model training job
Create a Vertex AI Pipelines job to link and run both components

D. Use the Kubeflow pipelines SDK to write code that specifies two components
- The first component initiates an Apache Spark context that runs the PySpark feature engineering code
- The second component runs the TensorFlow custom model training code
Create a Vertex AI Pipelines job to link and run both components.

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 187
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You recently deployed a scikit-learn model to a Vertex AI endpoint. You are now testing the model on live production traffic. While monitoring the endpoint, you discover twice as many requests per hour than expected throughout the day. You want the endpoint to efficiently scale when the demand increases in the future to prevent users from experiencing high latency. What should you do?

A. Deploy two models to the same endpoint, and distribute requests among them evenly

B. Configure an appropriate minReplicaCount value based on expected baseline traffic

C. Set the target utilization percentage in the autoscailngMetricSpecs configuration to a higher value

D. Change the model's machine type to one that utilizes GPUs

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 188
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work at a bank. You have a custom tabular ML model that was provided by the bank's vendor. The training data is not available due to its sensitivity. The model is packaged as a Vertex AI Model serving container, which accepts a string as input for each prediction instance. In each string, the feature values are separated by commas. You want to deploy this model to production for online predictions and monitor the feature distribution over time with minimal effort. What should you do?

A. 1. Upload the model to Vertex AI Model Registry, and deploy the model to a Vertex AI endpoint
2. Create a Vertex AI Model Monitoring job with feature drift detection as the monitoring objective, and provide an instance schema

B. 1. Upload the model to Vertex AI Model Registry, and deploy the model to a Vertex AI endpoint
2. Create a Vertex AI Model Monitoring job with feature skew detection as the monitoring objective, and provide an instance schema

C. 1. Refactor the serving container to accept key-value pairs as input format
2. Upload the model to Vertex AI Model Registry, and deploy the model to a Vertex AI endpoint
3. Create a Vertex AI Model Monitoring job with feature drift detection as the monitoring objective.

D. 1. Refactor the serving container to accept key-value pairs as input format
2. Upload the model to Vertex AI Model Registry, and deploy the model to a Vertex AI endpoint
3. Create a Vertex AI Model Monitoring job with feature skew detection as the monitoring objective

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 189

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are implementing a batch inference ML pipeline in Google Cloud. The model was developed using TensorFlow and is stored in SavedModel format in Cloud Storage. You need to apply the model to a historical dataset containing 10 TB of data that is stored in a BigQuery table. How should you perform the inference?

A. Export the historical data to Cloud Storage in Avro format. Configure a Vertex AI batch prediction job to generate predictions for the exported data

B. Import the TensorFlow model by using the CREATE MODEL statement in BigQuery ML. Apply the historical data to the TensorFlow model

C. Export the historical data to Cloud Storage in CSV format. Configure a Vertex AI batch prediction job to generate predictions for the exported data

D. Configure a Vertex AI batch prediction job to apply the model to the historical data in BigQuery

Show Suggested Answer

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 190
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You recently deployed a model to a Vertex AI endpoint. Your data drifts frequently, so you have enabled request-response logging and created a Vertex AI Model Monitoring job. You have observed that your model is receiving higher traffic than expected. You need to reduce the model monitoring cost while continuing to quickly detect drift. What should you do?

A. Replace the monitoring job with a DataFlow pipeline that uses TensorFlow Data Validation (TFDV)

B. Replace the monitoring job with a custom SQL script to calculate statistics on the features and predictions in BigQuery

C. Decrease the sample_rate parameter in the RandomSampleConfig of the monitoring job

D. Increase the monitor_interval parameter in the ScheduleConfig of the monitoring job

Show Suggested Answer