Number: Professional Machine Learning Engineer
Passing Score: 800
Time Limit: 120
File Version: 12.0

Exam Code: Google Professional Machine Learning Engineer
Exam Name: Google Professional Machine Learning Engineer

Exam A

QUESTION 1
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 Al Platform Training to execute the experiments Write the accuracy metrics to BigQuery, and query the results using the BigQueryAPI.
C. Use Al Platform Training to execute the experiments Write the accuracy metrics to Cloud Monitoring, and query the results using the Monitoring API.
D. Use Al Platform Notebooks to execute the experiments. Collect the results in a shared Google Sheets file, and query the results using the Google Sheets API

Correct Answer: A
Section:
Explanation:
https://codelabs.developers.google.com/codelabs/cloud-kubeflow-pipelines-gis Kubeflow Pipelines (KFP) helps solve these issues by providing a way to deploy robust, repeatable machine learning pipelines along with monitoring, auditing, version tracking, and reproducibility. Cloud AI Pipelines makes it easy to set up a KFP installation.
https://www.kubeflow.org/docs/components/pipelines/introduction/#what-is-kubeflow-pipelines
'Kubeflow Pipelines supports the export of scalar metrics. You can write a list of metrics to a local file to describe the performance of the model. The pipeline agent uploads the local file as your run-time metrics. You can view the uploaded metrics as a visualization in the Runs page for a particular experiment in the Kubeflow Pipelines UI.' https://www.kubeflow.org/docs/components/pipelines/sdk/pipelines-metrics/

QUESTION 2
You are developing an ML model intended to classify whether X-Ray images indicate bone fracture risk. You have trained on Api Resnet architecture on Vertex AI using a TPU as an accelerator, however you are unsatisfied with the trainning time and use 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 models accuracy. What should you do?

A. Configure your model to use bfloat16 instead float32
B. Reduce the global batch size from 1024 to 256
C. Reduce the number of layers in the model architecture
D. Reduce the dimensions of the images used un the model

Correct Answer: B
Section:

QUESTION 3
Your task is classify if a company logo is present on an image. You found out that 96% of a data does not include a logo. You are dealing with data imbalance problem. Which metric do you use to evaluate to model?

A. F1 Score
B. RMSE
C. F Score with higher precision weighting than recall
D. F Score with higher recall weighted than precision

Correct Answer: D
Section:

QUESTION 4
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

Correct Answer: B
Section:
Explanation:
https://cloud.google.com/bigquery-ml/docs/introduction

QUESTION 5
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 and Al Platform Prediction
C. Cloud Composer, BigQuery ML , and Al Platform Prediction
D. Cloud Composer, Al Platform Training with custom containers , and App Engine

Correct Answer: B
Section:

QUESTION 6
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

Correct Answer: A
Section:

QUESTION 7
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.

Correct Answer: C
Section:

QUESTION 8
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.

Correct Answer: B
Section:
Explanation:
AutoML Natural Language is a service that allows you to quickly build, test and deploy natural language processing (NLP) models without needing to have expertise in NLP or machine learning. You can use it to train a classifier on your corpus of written support cases, and then use the AutoML API to perform classification on new requests. Once the model is trained, it can be deployed as a REST API. This allows the classifier to be integrated into your pipeline and be easily consumed by other systems.

QUESTION 9
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

Correct Answer: A
Section:

QUESTION 10
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.

Correct Answer: C
Section:

QUESTION 11
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.

Correct Answer: D
Section:

QUESTION 12
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.

Correct Answer: C
Section:

QUESTION 13
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.

Vdumps

Correct Answer: A
Section:

QUESTION 14
You need to execute a batch prediction on 100million 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.

Correct Answer: A
Section:

QUESTION 15
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.

Correct Answer: C
Section:

QUESTION 16
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 24hours.
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.

Correct Answer: C
Section:

QUESTION 17
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.

Correct Answer: B
Section:
Explanation:
This would allow you to tailor the training process to your specific needs and requirements, and it would also allow for more flexible experimentation with different training strategies.
Additionally, creating a custom training loop could result in faster training times compared to using a single GPU or the distributed training strategies currently available in Keras.

QUESTION 18
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.

Correct Answer: D

Section:

QUESTION 19
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.

Correct Answer: A
Section:

QUESTION 20
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.

Correct Answer: A
Section:
Explanation:
This is the most likely problem that is occurring based on the information provided. Training-serving skew occurs when the distribution of the data used for training and the data used for serving the model in production are different. This can result in a drop in model performance when the model is deployed to production. It's also possible that the model is overfitting during training.
It is not a problem of insufficient amount of data because the data is split by using the BigQuery and it's not a problem of sharing some records between tables because it is not mentioned that the data is shared in the question.
The problem D is also not correct as the RAND() function is used to split the data but it doesn't mean that every record in the validation table will also be in the training table.

QUESTION 21
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,

Correct Answer: A
Section:

Vdumps

Explanation:
Pub/Sub is a messaging service that can be used to stream data into BigQuery in real-time. Configuring Pub/Sub to stream the user activity data into BigQuery would ensure real-time ingestion of the data. Source: Google Cloud

QUESTION 22
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.

Correct Answer: A
Section:
Explanation:
A should be the way to go for large datasets --This is also good but it is legacy way of checking:- NFORMATION_SCHEMA contains these views for table metadata: TABLES and TABLE_OPTIONS for metadata about tables. COLUMNS and COLUMN_FIELD_PATHS for metadata about columns and fields. PARTITIONS for metadata about table partitions (Preview)

QUESTION 23
You are 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.

Correct Answer: B
Section:
Explanation:
https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9

QUESTION 24
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.

Correct Answer: C
Section:
Explanation:
https://medium.com/google-cloud/secure-cloud-run-cloud-functions-and-app-engine-with-api-key-73c57bededd1

QUESTION 25
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.

Correct Answer: C
Section:
Explanation:
https://cloud.google.com/deep-learning-vm/docs/cli#creating_an_instance_with_one_or_more_gpus
https://cloud.google.com/deep-learning-vm/docs/introduction#pre-installed_packages
'speed up model training' will make us biased towards GPU,TPU options by options eliminations we may need to stay away of any manual installations , so using preconfigered deep learning will speed up time to market

QUESTION 26
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 data set. 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.

Correct Answer: B
Section:

QUESTION 27
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

Correct Answer: D
Section:
Explanation:
https://cloud.google.com/data-fusion/docs/concepts/overview#using_the_code-free_web_ui

QUESTION 28
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

Correct Answer: D
Section:
Explanation:
Google has bought this software and support for this tool is not good. SQL can work in Cloud fusion pipelines too but I would prefer to use a single tool like Bigquery to both transform and store data.

QUESTION 29
You are building a real-time prediction engine that streams files which may contain Personally Identifiable Information (Pll) to Google Cloud. You want to use the Cloud Data Loss Prevention (DLP) API to scan the files. How should you ensure that the Pll is not accessible by unauthorized individuals?

A. Stream all files to Google CloudT 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.
E. Periodically conduct a bulk scan of that bucket using the DLP API, and move the data to either the Sensitive or Non-Sensitive bucket

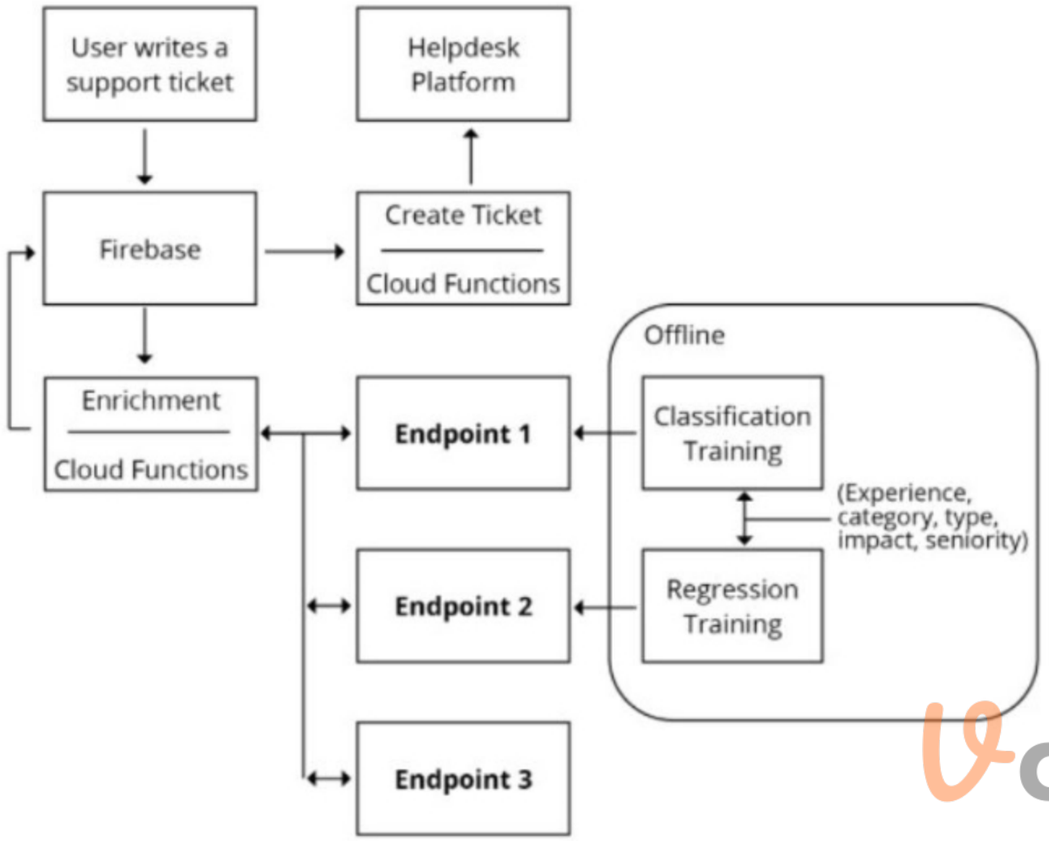
Correct Answer: A
Section:

QUESTION 30
You are designing an ML recommendation model for shoppers on your company's ecommerce website. You will use Recommendations Al to build, test, and deploy your system. How should you develop recommendations that increase revenue while following best practices?

A. Use the 'Other Products You May Like' recommendation type to increase the click-through rate
B. Use the 'Frequently Bought Together' 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.

Correct Answer: B
Section:
Explanation:
Frequently bought together' recommendations aim to up-sell and cross-sell customers by providing product.

QUESTION 31
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 = Vertex Al. 2 = Vertex Al. 3 = AutoML Natural Language
B. 1 = Vertex Al. 2 = Vertex Al. 3 = Cloud Natural Language API
C. 1 = Vertex Al. 2 = Vertex Al. 3 = AutoML Vision
D. 1 = Cloud Natural Language API. 2 = Vertex Al, 3 = Cloud Vision API

Correct Answer: B
Section:
Explanation:
https://cloud.google.com/architecture/architecture-of-a-serverless-ml-model#architecture
The architecture has the following flow:
A user writes a ticket to Firebase, which triggers a Cloud Function.
-The Cloud Function calls 3 different endpoints to enrich the ticket:
-An AI Platform endpoint, where the function can predict the priority.
-An AI Platform endpoint, where the function can predict the resolution time.
-The Natural Language API to do sentiment analysis and word salience.
-For each reply, the Cloud Function updates the Firebase real-time database.
-The Cloud Function then creates a ticket into the helpdesk platform using the RESTful API.

QUESTION 32

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.

Correct Answer: C
Section:
Explanation:
https://cloud.google.com/architecture/architecture-for-mlops-using-tfx-kubeflow-pipelines-and-cloud-build#triggering-and-scheduling-kubeflow-pipelines

QUESTION 33
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 Vertex 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.

Correct Answer: A
Section:
Explanation:
AI platform also contains kubeflow pipelines. you don't need to set up infrastructure to use it. For D you need to set up a kubernetes cluster engine. The question asks us to minimize infrastructure overheard.

QUESTION 34
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 AutoML 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

Correct Answer: B
Section:

QUESTION 35
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

Correct Answer: B
Section:
Explanation:
https://www.oecd.org/finance/Impact-Big-Data-AI-in-the-Insurance-Sector.pdf
https://medium.com/artefact-engineering-and-data-science/including-ethics-best-practices-in-your-data-science-project-from-day-one-c15b26c2bf99

QUESTION 36
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 30 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 set is from 30 days after your validation set

Correct Answer: D
Section:
Explanation:
https://cloud.google.com/automl-tables/docs/data-best-practices#time

QUESTION 37
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 customers1 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

Correct Answer: D
Section:
Explanation:
Firebase is designed for exactly this sort of scenario. Also, it would not be possible to create millions of pubsub topics due to GCP quotas https://cloud.google.com/pubsub/quotas#quotas
https://firebase.google.com/docs/cloud-messaging

QUESTION 38
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?

IT Certification Exams - Questions & Answers | Vdumps.com

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.

Correct Answer: A
Section:
Explanation:
https://cloud.google.com/ai-platform/prediction/docs/batch-predict

QUESTION 39
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 the 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.

Correct Answer: A
Section:
Explanation:
https://www.tensorflow.org/tfx/guide/evaluator

QUESTION 40
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 I AM 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.

Correct Answer: C
Section:
Explanation:
https://cloud.google.com/ai-platform/prediction/docs/resource-labels#overview_of_labels
You can add labels to your AI Platform Prediction jobs, models, and model versions, then use those labels to organize resources into categories when viewing or monitoring the resources. For example, you can label jobs by team (such as engineering or research) and development phase (prod or test), then filter the jobs based on the team and phase. Labels are also available on operations, but these labels are derived from the resource to which the operation applies. You cannot add or update labels on an operation.
https://cloud.google.com/ai-platform/prediction/docs/sharing-models.

QUESTION 41
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 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.

Correct Answer: B
Section:
Explanation:
https://cloud.google.com/ai-platform/prediction/docs/ai-explanations/overview#sampled-shapley

QUESTION 42
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 Al Platform to perform transfer learning
D. Use an established text classification model on Al Platform as-is to classify support requests

Correct Answer: C
Section:
Explanation:
the model cannot work as-is as the classes to predict will likely not be the same; we need to use transfer learning to retrain the last layer and adapt it to the classes we need

QUESTION 43
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:

Vdumps

```
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, ...]
    ...
...
```

Vdumps

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: [TextC1, TextC2, ...]

C)

Distribute sentences randomly across the train-test-eval subsets:

Train set: [SentenceA11, SentenceA21, Sentence B11, SentenceB21, SentenceC11, SentenceD21, ...]

Test set: [SentenceA12, SentenceA22, Sentence B12, SentenceC22, SentenceC12, SentenceD22, ...]

Eval set: [SentenceA13, SentenceA23, Sentence B13, SentenceC23, SentenceC13, SentenceD31, ...]

D)

Distribute paragraphs of texts (i.e., chunks of consecutive sentences) across the train-test-eval subsets:

Train set: [SentenceA11, SentenceA12, Sentence D11, SentenceD12, ...]

Test set: [SentenceA13, SentenceB13, Sentence B21, SentenceD23, SentenceC12, SentenceD13, ...]

Eval set: [SentenceA11, SentenceA22, Sentence B13, SentenceD22, SentenceC23, SentenceD11, ...]

A. Option A
B. Option B
C. Option C
D. Option D

Correct Answer: B
Section:
Explanation:
If we just put inside the Training set , Validation set and Test set , randomly Text, Paragraph or sentences the model will have the ability to learn specific qualities about The Author's use of language beyond just his own articles. Therefore the model will mixed up different opinions. Rather if we divided things up a the author level, so that given authors were only on the training data, or only in the test data or only in the validation data. The model will find more difficult to get a high accuracy on the test validation (What is correct and have more sense!). Because it will need to really focus in author by author articles rather than get a single political affiliation based on a bunch of mixed articles from different authors. https://developers.google.com/machine-learning/crash-course/18th-century-literature
For example, suppose you are training a model with purchase data from a number of stores. You know, however, that the model will be used primarily to make predictions for stores that are not in the training data. To ensure that the model can generalize to unseen stores, you should segregate your data sets by stores. In other words, your test set should include only stores different from the evaluation set, and the evaluation set should include only stores different from the training set. https://cloud.google.com/automl-tables/docs/prepare#ml-use

QUESTION 44
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. Increase the size of the training batch
B. Decrease the size of the training batch
C. Increase the learning rate hyperparameter
D. Decrease the learning rate hyperparameter

Correct Answer: D
Section:
Explanation:
https://developers.google.com/machine-learning/crash-course/introduction-to-neural-networks/playground-exercises

QUESTION 45
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? (Choose One Correct Answer)

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

Correct Answer: C
Section:

QUESTION 46
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.

Correct Answer: D
Section:

QUESTION 47
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

Correct Answer: B
Section:

QUESTION 48
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_STOCKED. Give the report to the logistics team each month so they can fine-tune inventory levels.

Correct Answer: B
Section:

QUESTION 49
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

Correct Answer: B
Section:

QUESTION 50
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.

Correct Answer: A
Section:
Explanation:
Containerizing the model prediction logic in Cloud Run allows for easy and efficient deployment of the model, and allows it to be invoked by Dataflow. Cloud Run is a fully managed service that allows you to run stateless containers in a serverless environment. It automatically scales instances up and down based on the traffic, which can minimize the serving latency.
Additionally, Dataflow can easily invoke Cloud Run services via HTTP requests, making it simple to integrate into your pipeline. This allows the Dataflow pipeline to focus on data ingestion and processing, while the Cloud Run service handles the real-time predictions.
While it is possible to load the model directly into the Dataflow job as a dependency, this approach can increase the complexity of the pipeline and could lead to increased latency. Other options, such as deploying the model to a Vertex AI endpoint or a TFServing container on GKE, would also work but this option is the most optimal for minimizing the serving latency.

QUESTION 51
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

Correct Answer: C
Section:

QUESTION 52
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.

Correct Answer: D
Section:
Explanation:
TensorFlow I/O's BigQuery Reader allows you to directly read data from BigQuery tables into your TensorFlow model without the need to export the data to a separate file format. This can minimize any bottlenecks during the data ingestion stage and also it can increase the scalability. By using BigQuery Reader, you can easily read large amounts of data from BigQuery and use it to train your model without having to worry about the performance impact of reading from a dataframe or CSV file.
You can use the tfio.BigQueryRecordDataset which will return a dataset of dictionaries, and where each key corresponds to a table column and each value corresponds to the value in that column.

QUESTION 53
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.

Correct Answer: B
Section:

QUESTION 54
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.

Correct Answer: D
Section:

QUESTION 55
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.

Correct Answer: C
Section:

QUESTION 56
You work for an online publisher that delivers news articles to over 50million 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 12hours. 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.

Correct Answer: C
Section:
Explanation:
Scheduling a weekly query in BigQuery to compute the success metric is a cost-effective way to monitor the model's performance. BigQuery allows you to run complex queries on large datasets in a cost-effective and performant manner. By using BigQuery, you can compute the success metric on a regular basis without incurring the additional costs of other services such as Vertex AI or Cloud Composer.
Additionally, by scheduling the query to run weekly, you can ensure that you are monitoring the model's performance in a timely manner, while still providing enough time for the model to degrade below the acceptable baseline. You can then use the results of the query to determine when retraining is necessary.

QUESTION 57
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.

Correct Answer: A
Section:

QUESTION 58
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 50milliseconds). 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.

Correct Answer: A
Section:

QUESTION 59
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.

Correct Answer: A
Section:

QUESTION 60
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%.

Correct Answer: B
Section:
Explanation:
In this case, the model has a high accuracy of 99% for identifying customers who cancel their subscriptions, but a lower accuracy of 82% for identifying customers who renew their subscriptions. However, this does not necessarily mean that the model is performing well, because 90% of the customers renew their subscription, so if the model always predicts that customers will renew, it will be correct 90% of the time. Therefore, the model's performance is worse than the baseline of always predicting that customers will renew their subscription.
https://en.wikipedia.org/wiki/Imbalanced_data
https://machinelearningmastery.com/baseline-performance-machine-learning-algorithms/

QUESTION 61
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 Al 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 Al managed dataset.
D. Use a Vertex Al 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 Al managed dataset.

Correct Answer: A
Section:

QUESTION 62
You developed a Vertex Al 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 Al 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 Al 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 Al Pipelines.
D. Trigger GitHub Actions to run the tests launch a Cloud Build workflow to build custom Dicker images, push the images to Artifact Registry, and launch the pipeline in Vertex Al Pipelines.

Correct Answer: C
Section:

QUESTION 63
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. ONE_HOT_ENCODER function on the categorical features, and select the encoded categorical features and non-categorical features as inputs to create your model.

Correct Answer: D
Section:

QUESTION 64
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 Al Pipelines with the Kubeflow Pipelines SDK to create a pipeline that reads the images from Cloud Storage and trains the model.
B. Use Vertex Al Pipelines with TensorFlow Extended (TFX) to create a pipeline that reads the images from Cloud Storage and trams the model.
C. Import the labeled images as a managed dataset in Vertex Al: and use AutoML to tram the model.
D. Convert the image dataset to a tabular format using Dataflow Load the data into BigQuery and use BigQuery ML to tram the model.

Correct Answer: A
Section:

QUESTION 65
You are developing a mode! 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 2 answers

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.

Correct Answer: B, D
Section:

QUESTION 66
You need to deploy a scikit-learn 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 Al prediction endpoint Set the max replica count to 1
B. Deploy an online Vertex Al prediction endpoint Set the max replica count to 100
C. Deploy an online Vertex Al prediction endpoint with one GPU per replica Set the max replica count to 1.
D. Deploy an online Vertex Al prediction endpoint with one GPU per replica Set the max replica count to 100.

Correct Answer: C
Section:

QUESTION 67
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 tram 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 M-standard-4 VM with 4 NVIDIA P100 GPUs SSH into the VM and use Parameter Server Strategy to train the model.
D. Configure a M-standard-4 VM with 4 NVIDIA P100 GPUs SSH into the VM and use MultiWorkerMirroredStrategy to train the model.

Correct Answer: B
Section:

QUESTION 68
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 teaming rates Choose the grid search option
B. 1 Create a Vertex Al 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 Al 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 Al Experiments 2. Create a Vertex Al pipeline with a custom model training job as part of the pipeline. Configure the pipelines parameters to include those you are investigating 3. Submit multiple runs to the same experiment using different values for the parameters

Vdumps

Correct Answer: B
Section:

QUESTION 69
You received a training-serving skew alert from a Vertex Al Model Monitoring job running in production. You retrained the model with more recent training data, and deployed it back to the Vertex Al 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 Al 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 Al endpoint

Correct Answer: D
Section:

QUESTION 70
You developed a custom model by using Vertex Al 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 Al 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-rare 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.

Correct Answer: D
Section:

QUESTION 71
You have recently trained a scikit-learn model that you plan to deploy on Vertex Al. 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 Al Model Registry by using a prebuilt scikit-learn prediction container 2 Deploy your model to Vertex Al Endpoints, and create a Vertex Al batch prediction job that uses the instanceConfig.inscanceType 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 sci-kit learn model container to Vertex Al Model Registry 3 Deploy your model to Vertex Al Endpoints, and create a Vertex Al batch prediction job
C. 1. Create a custom container for your sci-kit learn model, 2 Define a custom serving function for your model 3 Upload your model and custom container to Vertex Al Model Registry 4 Deploy your model to Vertex Al Endpoints, and create a Vertex Al batch prediction job
D. 1 Create a custom container for your sci-kit learn model. 2 Upload your model and custom container to Vertex Al Model Registry 3 Deploy your model to Vertex Al Endpoints, and create a Vertex Al batch prediction job that uses the instanceConfig. instanceType setting to transform your input data

Correct Answer: B
Section:

QUESTION 72
You work for a food product company. Your company's historical sales data is stored in BigQuery You need to use Vertex Al'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 min-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?

IT Certification Exams - Questions & Answers | Vdumps.com

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.

Correct Answer: B
Section:

QUESTION 73
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.

Correct Answer: D
Section:

QUESTION 74
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.

Correct Answer: C
Section:

QUESTION 75
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 Kubemetes 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 contain 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.

Correct Answer: D
Section:

QUESTION 76
You recently deployed a pipeline in Vertex Al Pipelines that trains and pushes a model to a Vertex Al 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 Al 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 arrets 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

Correct Answer: C
Section:

QUESTION 77
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 projects 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 peering's, and add Vertex Al to a service perimeter
B. Create a Cloud Run endpoint as a proxy to the data Use Identity and Access Management (1AM) authentication to secure access to the endpoint from the training job.
C. Configure VPC Peering with Vertex Al and specify the network of the training job
D. Download the data to a Cloud Storage bucket before calling the training job

vdumps

Correct Answer: B
Section:

QUESTION 78
You are deploying a new version of a model to a production Vertex Al 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 Al 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 Al 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 Al 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 Al Model Registry 2 Deploy the new model to the existing endpoint

Correct Answer: C
Section:

QUESTION 79
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

Correct Answer: D
Section:

QUESTION 80
You work for a retail company. You have a managed tabular dataset in Vertex Al 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 Al manual split, using the store name feature to assign one store for each set.
B. Use Vertex Al default data split.
C. Use Vertex Al chronological split and specify the sales timestamp feature as the time vanable.
D. Use Vertex Al random split assigning 70% of the rows to the training set, 10% to the validation set, and 20% to the test set.

Correct Answer: A
Section:

QUESTION 81
You have developed a BigQuery ML model that predicts customer churn and deployed the model to Vertex Al 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 Al 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 Al 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 Al Model Monitoring job configured to monitor prediction drift. 2. Configure alert monitoring to publish a message to a Pub/Sub queue when a monitonng alert is detected. 3. Use a Cloud Function to monitor the Pub/Sub queue, and trigger retraining in BigQuery
D. 1. Create a Vertex Al 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.

Correct Answer: C
Section:

QUESTION 82
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 Al 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 Al 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 Al 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 creare_cusrora_rraining_job_from_ccraponent Utility that launches the custom model training job.
D. Create a Vertex Al Pipelines job to link and run both components 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 Al Pipelines job to link and run both components

Correct Answer: C
Section:

QUESTION 83
You recently deployed a scikit-learn model to a Vertex Al 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 autcscalir.gMetricspecs configuration to a higher value
D. Change the model's machine type to one that utilizes GPUs.

Correct Answer: B
Section:

QUESTION 84
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 Al 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 Al Model Registry and deploy the model to a Vertex Ai endpoint. 2. Create a Vertex Al Model Monitoring job with feature drift detection as the monitoring objective, and provide an instance schema.
B. 1 Upload the model to Vertex Al Model Registry and deploy the model to a Vertex Al endpoint. 2 Create a Vertex Al 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 Al Model Registry and deploy the model to a Vertex Al endpoint. 3. Create a Vertex Al 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 Al Model Registry and deploy the model to a Vertex Al endpoint. 3. Create a Vertex Al Model Monitoring job with feature skew detection as the monitoring objective.

Correct Answer: C
Section:

QUESTION 85
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 Al 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 Al batch prediction job to generate predictions for the exported data.
D. Configure a Vertex Al batch prediction job to apply the model to the historical data in BigQuery

Correct Answer: D
Section:

QUESTION 86
You recently deployed a model to a Vertex Al endpoint Your data drifts frequently so you have enabled request-response logging and created a Vertex Al 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 scnpt 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 scheduieconfig of the monitoring job.

Correct Answer: C
Section:

QUESTION 87
You work for a retail company. You have created a Vertex Al forecast model that produces monthly item sales predictions. You want to quickly create a report that will help to explain how the model calculates the predictions. You have one month of recent actual sales data that was not included in the training dataset. How should you generate data for your report?

A. Create a batch prediction job by using the actual sales data Compare the predictions to the actuals in the report.
B. Create a batch prediction job by using the actual sates data and configure the job settings to generate feature attributions. Compare the results in the report.
C. Generate counterfactual examples by using the actual sales data Create a batch prediction job using the actual sales data and the counterfactual examples Compare the results in the report.
D. Train another model by using the same training dataset as the original and exclude some columns. Using the actual sales data create one batch prediction job by using the new model and another one with the original model Compare the two sets of predictions in the report.

Correct Answer: B
Section:

QUESTION 88
Your team has a model deployed to a Vertex Al endpoint You have created a Vertex Al pipeline that automates the model training process and is triggered by a Cloud Function. You need to prioritize keeping the model up-to-date, but also minimize retraining costs. How should you configure retraining'?

A. Configure Pub/Sub to call the Cloud Function when a sufficient amount of new data becomes available.
B. Configure a Cloud Scheduler job that calls the Cloud Function at a predetermined frequency that fits your team's budget.
C. Enable model monitoring on the Vertex Al endpoint Configure Pub/Sub to call the Cloud Function when anomalies are detected.
D. Enable model monitoring on the Vertex Al endpoint Configure Pub/Sub to call the Cloud Function when feature drift is detected.

Correct Answer: D
Section:

QUESTION 89
Your company stores a large number of audio files of phone calls made to your customer call center in an on-premises database. Each audio file is in wav format and is approximately 5 minutes long. You need to analyze these audio files for customer sentiment. You plan to use the Speech-to-Text API. You want to use the most efficient approach. What should you do?

A. 1 Upload the audio files to Cloud Storage 2. Call the speech: Iongrunningrecognize API endpoint to generate transcriptions 3. Call the predict method of an AutoML sentiment analysis model to analyze the transcriptions
B. 1 Upload the audio files to Cloud Storage 2 Call the speech: Iongrunningrecognize API endpoint to generate transcriptions. 3 Create a Cloud Function that calls the Natural Language API by using the analyzesentiment method
C. 1 Iterate over your local Tiles in Python 2. Use the Speech-to-Text Python library to create a speech.RecognitionAudio object and set the content to the audio file data 3. Call the speech: recognize API endpoint to generate transcriptions 4. Call the predict method of an AutoML sentiment analysis model to analyze the transcriptions
D. 1 Iterate over your local files in Python 2 Use the Speech-to-Text Python Library to create a speech.RecognitionAudio object, and set the content to the audio file data 3. Call the speech: lengrunningrecognize API endpoint to generate transcriptions 4 Call the Natural Language API by using the analyzesenriment method

Correct Answer: B
Section:

QUESTION 90
You work for a social media company. You want to create a no-code image classification model for an iOS mobile application to identify fashion accessories You have a labeled dataset in Cloud Storage You need to configure a training workflow that minimizes cost and serves predictions with the lowest possible latency What should you do?

A. Train the model by using AutoML, and register the model in Vertex Al Model Registry Configure your mobile application to send batch requests during prediction.
B. Train the model by using AutoML Edge and export it as a Core ML model Configure your mobile application to use the mlmodel file directly.
C. Train the model by using AutoML Edge and export the model as a TFLite model Configure your mobile application to use the tflite file directly
D. Train the model by using AutoML, and expose the model as a Vertex Al endpoint Configure your mobile application to invoke the endpoint during prediction.

Correct Answer: B
Section:

QUESTION 91
You work for a retail company. You have been asked to develop a model to predict whether a customer will purchase a product on a given day. Your team has processed the company's sales data, and created a table with the following rows:
* Customer_id
* Product_id
* Date
* Days_since_last_purchase (measured in days)
* Average_purchase_frequency (measured in 1/days)
* Purchase (binary class, if customer purchased product on the Date)
You need to interpret your models results for each individual prediction. What should you do?

A. Create a BigQuery table Use BigQuery ML to build a boosted tree classifier Inspect the partition rules of the trees to understand how each prediction flows through the trees.
B. Create a Vertex Al tabular dataset Train an AutoML model to predict customer purchases Deploy the model to a Vertex Al endpoint and enable feature attributions Use the 'explain' method to get feature attribution values for each individual prediction.
C. Create a BigQuery table Use BigQuery ML to build a logistic regression classification model Use the values of the coefficients of the model to interpret the feature importance with higher values corresponding to more importance.
D. Create a Vertex Al tabular dataset Train an AutoML model to predict customer purchases Deploy the model to a Vertex Al endpoint. At each prediction enable L1 regularization to detect non-informative features.

Correct Answer: B
Section:

QUESTION 92
You work for a company that captures live video footage of checkout areas in their retail stores You need to use the live video footage to build a mode! to detect the number of customers waiting for service in near real time You want to implement a solution quickly and with minimal effort How should you build the model?

A. Use the Vertex Al Vision Occupancy Analytics model.
B. Use the Vertex Al Vision Person/vehicle detector model
C. Train an AutoML object detection model on an annotated dataset by using Vertex AutoML
D. Train a Seq2Seq+ object detection model on an annotated dataset by using Vertex AutoML

Correct Answer: A
Section:

QUESTION 93
You are building a MLOps platform to automate your company's ML experiments and model retraining. You need to organize the artifacts for dozens of pipelines How should you store the pipelines' artifacts'?

IT Certification Exams - Questions & Answers | Vdumps.com

A. Store parameters in Cloud SQL and store the models' source code and binaries in GitHub
B. Store parameters in Cloud SQL store the models' source code in GitHub, and store the models' binaries in Cloud Storage.
C. Store parameters in Vertex ML Metadata store the models' source code in GitHub and store the models' binaries in Cloud Storage.
D. Store parameters in Vertex ML Metadata and store the models source code and binaries in GitHub.

Correct Answer: C
Section:

QUESTION 94
You work for a telecommunications company You're building a model to predict which customers may fail to pay their next phone bill. The purpose of this model is to proactively offer at-risk customers assistance such as service discounts and bill deadline extensions. The data is stored in BigQuery, and the predictive features that are available for model training include
- Customer_id -Age
- Salary (measured in local currency) -Sex
-Average bill value (measured in local currency)
- Number of phone calls in the last month (integer) -Average duration of phone calls (measured in minutes)
You need to investigate and mitigate potential bias against disadvantaged groups while preserving model accuracy What should you do?

A. Determine whether there is a meaningful correlation between the sensitive features and the other features Train a BigQuery ML boosted trees classification model and exclude the sensitive features and any meaningfully correlated features
B. Train a BigQuery ML boosted trees classification model with all features Use the ml. global explain method to calculate the global attribution values for each feature of the model If the feature importance value for any of the sensitive features exceeds a threshold, discard the model and tram without this feature
C. Train a BigQuery ML boosted trees classification model with all features Use the ml. exflain_predict method to calculate the attribution values for each feature for each customer in a test set If for any individual customer the importance value for any feature exceeds a predefined threshold, discard the model and train the model again without this feature.
D. Define a fairness metric that is represented by accuracy across the sensitive features Train a BigQuery ML boosted trees classification model with all features Use the trained model to make predictions on a test set Join the data back with the sensitive features, and calculate a fairness metric to investigate whether it meets your requirements.

Correct Answer: A
Section:

QUESTION 95
You recently trained a XGBoost model that you plan to deploy to production for online inference Before sending a predict request to your model's binary you need to perform a simple data preprocessing step This step exposes a REST API that accepts requests in your internal VPC Service Controls and returns predictions You want to configure this preprocessing step while minimizing cost and effort What should you do?

A. Store a pickled model in Cloud Storage Build a Flask-based app packages the app in a custom container image, and deploy the model to Vertex Al Endpoints.
B. Build a Flask-based app. package the app and a pickled model in a custom container image, and deploy the model to Vertex Al Endpoints.
C. Build a custom predictor class based on XGBoost Predictor from the Vertex Al SDK. package it and a pickled model in a custom container image based on a Vertex built-in image, and deploy the model to Vertex Al Endpoints.
D. Build a custom predictor class based on XGBoost Predictor from the Vertex Al SDK and package the handler in a custom container image based on a Vertex built-in container image Store a pickled model in Cloud Storage and deploy the model to Vertex Al Endpoints.

Correct Answer: C
Section:

QUESTION 96
You work at a bank. You need to develop a credit risk model to support loan application decisions You decide to implement the model by using a neural network in TensorFlow Due to regulatory requirements, you need to be able to explain the models predictions based on its features When the model is deployed, you also want to monitor the model's performance overtime You decided to use Vertex Al for both model development and deployment What should you do?

A. Use Vertex Explainable AI with the sampled Shapley method, and enable Vertex AI Model Monitoring to check for feature distribution drift.
B. Use Vertex Explainable AI with the sampled Shapley method, and enable Vertex AI Model Monitoring to check for feature distribution skew.
C. Use Vertex Explainable AI with the XRAI method, and enable Vertex AI Model Monitoring to check for feature distribution drift.
D. Use Vertex Explainable AI with the XRAI method and enable Vertex AI Model Monitoring to check for feature distribution skew.

Correct Answer: A
Section:

QUESTION 97
You are investigating the root cause of a misclassification error made by one of your models. You used Vertex AI Pipelines to tram and deploy the model. The pipeline reads data from BigQuery. creates a copy of the data in Cloud Storage in TFRecord format trains the model in Vertex AI Training on that copy, and deploys the model to a Vertex AI endpoint. You have identified the specific version of that model that misclassified: and you need to recover the data this model was trained on. How should you find that copy of the data'?

A. Use Vertex AI Feature Store Modify the pipeline to use the feature store; and ensure that all training data is stored in it Search the feature store for the data used for the training.
B. Use the lineage feature of Vertex AI Metadata to find the model artifact Determine the version of the model and identify the step that creates the data copy, and search in the metadata for its location.
C. Use the logging features in the Vertex AI endpoint to determine the timestamp of the models deployment Find the pipeline run at that timestamp Identify the step that creates the data copy; and search in the logs for its location.
D. Find the job ID in Vertex AI Training corresponding to the training for the model Search in the logs of that job for the data used for the training.

Correct Answer: B
Section:

QUESTION 98
You work for a manufacturing company. You need to train a custom image classification model to detect product defects at the end of an assembly line Although your model is performing well some images in your holdout set are consistently mislabeled with high confidence You want to use Vertex AI to understand your model's results What should you do?

A. Configure feature-based explanations by using Integrated Gradients. Set `visualization` type to `PIXELS`, and set `clip_percent_upperbound` to 95.
B. Create an index by using Vertex AI Matching Engine. Query the index with your mislabeled images.
C. Configure feature-based explanations by using XRAI. Set `visualization` type to `OUTLINES`, and set `polarity` to positive.
D. Configure example-based explanations. Specify the embedding output layer to be used for the latent space representation.

Correct Answer: A
Section:

QUESTION 99
You are using Keras and TensorFlow to develop a fraud detection model Records of customer transactions are stored in a large table in BigQuery. You need to preprocess these records in a cost-effective and efficient way

before you use them to train the model. The trained model will be used to perform batch inference in BigQuery. How should you implement the preprocessing workflow?

A. Implement a preprocessing pipeline by using Apache Spark, and run the pipeline on Dataproc Save the preprocessed data as CSV files in a Cloud Storage bucket.
B. Load the data into a pandas DataFrame Implement the preprocessing steps using panda's transformations. and train the model directly on the DataFrame.
C. Perform preprocessing in BigQuery by using SQL Use the BigQueryClient in TensorFlow to read the data directly from BigQuery.
D. Implement a preprocessing pipeline by using Apache Beam, and run the pipeline on Dataflow Save the preprocessed data as CSV files in a Cloud Storage bucket.

Correct Answer: D
Section:

QUESTION 100
You are training models in Vertex Al by using data that spans across multiple Google Cloud Projects You need to find track, and compare the performance of the different versions of your models Which Google Cloud services should you include in your ML workflow?

A. Dataplex. Vertex Al Feature Store and Vertex Al TensorBoard
B. Vertex Al Pipelines, Vertex Al Feature Store, and Vertex Al Experiments
C. Dataplex. Vertex Al Experiments, and Vertex Al ML Metadata
D. Vertex Al Pipelines: Vertex Al Experiments and Vertex Al Metadata

Correct Answer: D
Section:

QUESTION 101
You need to use TensorFlow to train an image classification model. Your dataset is located in a Cloud Storage directory and contains millions of labeled images Before training the model, you need to prepare the data. You want the data preprocessing and model training workflow to be as efficient scalable, and low maintenance as possible. What should you do?

A. 1 Create a Dataflow job that creates sharded TFRecord files in a Cloud Storage directory. 2 Reference tf .data.TFRecordDataset in the training script. 3. Train the model by using Vertex Al Training with a V100 GPU.
B. 1 Create a Dataflow job that moves the images into multiple Cloud Storage directories, where each directory is named according to the corresponding label. 2 Reference tfds.fclder_da-asst.imageFclder in the training script. 3. Train the model by using Vertex AI Training with a V100 GPU.
C. 1 Create a Jupyter notebook that uses an n1-standard-64, V100 GPU Vertex Al Workbench instance. 2 Write a Python script that creates sharded TFRecord files in a directory inside the instance 3. Reference tf. da-a.TFRecrrdDataset in the training script. 4. Train the model by using the Workbench instance.
D. 1 Create a Jupyter notebook that uses an n1-standard-64, V100 GPU Vertex Al Workbench instance. 2 Write a Python scnpt that copies the images into multiple Cloud Storage directories, where each directory is named according to the corresponding label. 3 Reference tf ds. f older_dataset. imageFolder in the training script. 4. Train the model by using the Workbench instance.

Correct Answer: A
Section:

QUESTION 102
You are building a custom image classification model and plan to use Vertex Al Pipelines to implement the end-to-end training. Your dataset consists of images that need to be preprocessed before they can be used to train the model. The preprocessing steps include resizing the images, converting them to grayscale, and extracting features. You have already implemented some Python functions for the preprocessing tasks. Which components should you use in your pipeline'?

A. `DataprocSparkBatchOp` and `CustomTrainingJobOp`
B. `DataflowPythonJobOp`, `WaitGcpResourcesOp`, and `CustomTrainingJobOp`

C.

```
dsl.ParallelFor, dsl.component, and CustomTrainingJobOp
```

D.

```
ImageDatasetImportDataOp, dsl.component, and AutoMLImageTrainingJobRunOp
```

Correct Answer: B
Section:

QUESTION 103
You work for a retail company that is using a regression model built with BigQuery ML to predict product sales. This model is being used to serve online predictions Recently you developed a new version of the model that uses a different architecture (custom model) Initial analysis revealed that both models are performing as expected You want to deploy the new version of the model to production and monitor the performance over the next two months You need to minimize the impact to the existing and future model users How should you deploy the model?

A. Import the new model to the same Vertex Al Model Registry as a different version of the existing model. Deploy the new model to the same Vertex Al endpoint as the existing model, and use traffic splitting to route 95% of production traffic to the BigQuery ML model and 5% of production traffic to the new model.
B. Import the new model to the same Vertex Al Model Registry as the existing model Deploy the models to one Vertex Al endpoint Route 95% of production traffic to the BigQuery ML model and 5% of production traffic to the new model
C. Import the new model to the same Vertex Al Model Registry as the existing model Deploy each model to a separate Vertex Al endpoint.
D. Deploy the new model to a separate Vertex Al endpoint Create a Cloud Run service that routes the prediction requests to the corresponding endpoints based on the input feature values.

Correct Answer: A
Section:

QUESTION 104
You work for a large retailer and you need to build a model to predict customer churn. The company has a dataset of historical customer data, including customer demographics, purchase history, and website activity. You need to create the model in BigQuery ML and thoroughly evaluate its performance. What should you do?

A. Create a linear regression model in BigQuery ML and register the model in Vertex Al Model Registry Evaluate the model performance in Vertex Al.
B. Create a logistic regression model in BigQuery ML and register the model in Vertex Al Model Registry. Evaluate the model performance in Vertex Al.
C. Create a linear regression model in BigQuery ML Use the ml. evaluate function to evaluate the model performance.
D. Create a logistic regression model in BigQuery ML Use the ml.confusion_matrix function to evaluate the model performance.

Correct Answer: B
Section:

QUESTION 105
You are developing a model to identify traffic signs in images extracted from videos taken from the dashboard of a vehicle. You have a dataset of 100 000 images that were cropped to show one out of ten different traffic signs. The images have been labeled accordingly for model training and are stored in a Cloud Storage bucket You need to be able to tune the model during each training run. How should you train the model?

A. Train a model for object detection by using Vertex Al AutoML.
B. Train a model for image classification by using Vertex Al AutoML.
C. Develop the model training code for object detection and tram a model by using Vertex Al custom training.
D. Develop the model training code for image classification and train a model by using Vertex Al custom training.

Correct Answer: C

Section:

QUESTION 106
You have deployed a scikit-learn model to a Vertex Al endpoint using a custom model server. You enabled auto scaling; however, the deployed model fails to scale beyond one replica, which led to dropped requests. You notice that CPU utilization remains low even during periods of high load. What should you do?

A. Attach a GPU to the prediction nodes.
B. Increase the number of workers in your model server.
C. Schedule scaling of the nodes to match expected demand.
D. Increase the minReplicaCount in your DeployedModel configuration.

Correct Answer: B
Section: