Google Professional Machine Learning Engineer

1. You are an AI engineer working for a popular video streaming platform. You built a classification model using PyTorch to predict customer churn. Each week, the customer retention team plans to contact customers identified as at-risk for churning with personalized offers. You want to deploy the model while minimizing maintenance effort. What should you do?

A. Use Vertex AI's prebuilt containers for prediction. Deploy the container on Cloud Run to generate online predictions.

B. Use Vertex AI's prebuilt containers for prediction. Deploy the model on Google Kubernetes Engine (GKE), and configure the model for batch prediction.

C. Deploy the model to a Vertex AI endpoint, and configure the model for batch prediction. Schedule the batch prediction to run weekly.

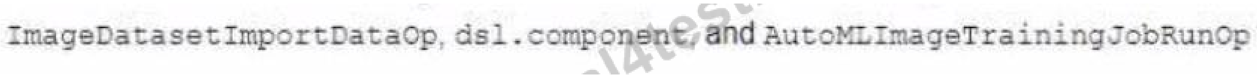
D. Deploy the model to a Vertex AI endpoint, and configure the model for online prediction. Schedule a job to query this endpoint weekly.

Answer(s): C

---

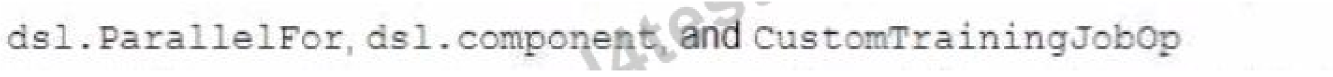
2. You are building a custom image classification model and plan to use Vertex AI Pipelines to implement the end-to-end training. Your dataset consists of images that need to be preprocessed before they can be used to train the model. The preprocessing steps include resizing the images, converting them 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. ○ **A.**



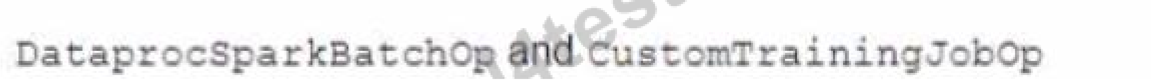
ImageDatasetImportDataOp, dsl.component, and AutoMLImageTrainingJobRunOp

B. ○ **B.**



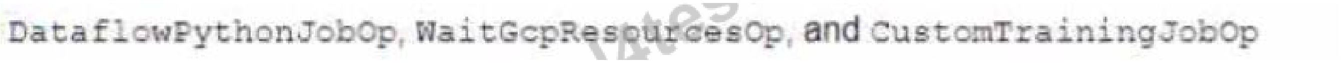
dsl.ParallelFor, dsl.component, and CustomTrainingJobOp

C. ○ **C.**



DataprocSparkBatchOp and CustomTrainingJobOp

D. ○ **D.**



DataflowPythonJobOp, WaitGcpResourcesOp, and CustomTrainingJobOp

Answer(s): D

---

3. You work at an organization that maintains a cloud-based communication platform that integrates conventional chat, voice, and video conferencing into one platform. The audio recordings are stored in Cloud Storage. All recordings have an 8 kHz sample rate and are more than one minute long. You need to implement a new feature in the platform that will automatically transcribe voice call recordings into a text for future applications, such as call summarization and sentiment analysis. How should you implement the voice call transcription feature following Google-recommended best practices?

A. Upsample the audio recordings to 16 kHz. and transcribe the audio by using the Speech-to-Text API with asynchronous recognition.

B. Upsample the audio recordings to 16 kHz. and transcribe the audio by using the Speech-to-Text API with synchronous recognition.

C. Use the original audio sampling rate, and transcribe the audio by using the Speech-to-Text API with synchronous recognition.

D. Use the original audio sampling rate, and transcribe the audio by using the Speech-to-Text API with asynchronous recognition.

Answer(s): A

---

4. 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 AI Vision Occupancy Analytics model.

B. Use the Vertex AI 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

Answer(s): A

---

5. Your company manages an ecommerce website. You developed an ML model that recommends additional products to users in near real time based on items currently in the user's cart. The workflow will include the following processes.

A. Write a Cloud Function that loads the model into memory for prediction Configure the function to be triggered when messages are sent to Pub/Sub.

B. Create a pipeline in Vertex AI Pipelines that performs preprocessing, prediction and postprocessing Configure the pipeline to be triggered by a Cloud Function when messages are sent to Pub/Sub.

C. Expose the model as a Vertex AI endpoint Write a custom DoFn in a Dataflow job that calls the endpoint for prediction.

D. Use the RunInference API with watchFilePatterr. in a Dataflow job that wraps around the model and serves predictions.

Answer(s): D

---

6. You work at a gaming startup that has several terabytes of structured data in Cloud Storage. This data includes gameplay time data user metadata and game metadat a. You want to build a model that recommends new games to users that requires the least amount of coding. What should you do?

A. Load the data in BigQuery Use BigQuery ML to tram an Autoencoder model.

B. Load the data in BigQuery Use BigQuery ML to train a matrix factorization model.

C. Read data to a Vertex AI Workbench notebook Use TensorFlow to train a two-tower model.

D. Read data to a Vertex AI Workbench notebook Use TensorFlow to train a matrix factorization model.

Answer(s): B

---

7. You need to develop a custom TensorRow model that will be used for online predictions. The training data is stored in BigQuery. You need to apply instance-level data transformations to the data for model training and serving. You want to use the same preprocessing routine during model training and serving. How should you configure the preprocessing routine?

A. Create a BigQuery script to preprocess the data, and write the result to another BigQuery table.

B. Create a pipeline in Vertex AI Pipelines to read the data from BigQuery and preprocess it using a custom preprocessing component.

C. Create a preprocessing function that reads and transforms the data from BigQuery Create a Vertex AI custom prediction routine that calls the preprocessing function at serving time.

D. Create an Apache Beam pipeline to read the data from BigQuery and preprocess it by using TensorFlow Transform and Dataflow.

Answer(s): D

---

8. You need to train a ControlNet model with Stable Diffusion XL for an image editing use case. You want to train this model as quickly as possible. Which hardware configuration should you choose to train your model?

A. Configure one a2-highgpu-1g instance with an NVIDIA A100 GPU with 80 GB of RAM. Use float32 precision during model training.

B. Configure one a2-highgpu-1g instance with an NVIDIA A100 GPU with 80 GB of RAM. Use bfloat16 quantization during model training.

C. Configure four n1-standard-16 instances, each with one NVIDIA Tesla T4 GPU with 16 GB of RAM. Use float32 precision during model training.

D. Configure four n1-standard-16 instances, each with one NVIDIA Tesla T4 GPU with 16 GB of RAM. Use float16 quantization during model training.

Answer(s): A

---

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

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

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

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

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

Answer(s): B

---

10. You have recently used TensorFlow to train a classification model on tabular data You have created a Dataflow pipeline that can transform several terabytes of data into training or prediction datasets consisting of TFRecords. You now need to productionize the model, and you want the predictions to be automatically uploaded to a BigQuery table on a weekly schedule. What should you do?

A. Import the model into Vertex AI and deploy it to a Vertex AI endpoint On Vertex AI Pipelines create a pipeline that uses the Dataf lowPythonJobop and the Mcdei3archPredictoc components.

B. Import the model into Vertex AI and deploy it to a Vertex AI endpoint Create a Dataflow pipeline that reuses the data processing logic sends requests to the endpoint and then uploads predictions to a BigQuery table.

C. Import the model into Vertex AI On Vertex AI Pipelines, create a pipeline that uses the DataflowPythonJobOp and the ModelBatchPredictOp components.

D. Import the model into BigQuery Implement the data processing logic in a SQL query On Vertex AI Pipelines create a pipeline that uses the BigqueryQueryJobop and the EigqueryPredictModejobOp components.

Answer(s): C

---

11. You have developed an AutoML tabular classification model that identifies high-value customers who interact with your organization's website.

A. Configure the model deployment settings to use an n1-standard-32 machine type.

B. Configure the model deployment settings to use an n1-standard-4 machine type. Set the minReplicaCount value to 1 and the maxReplicaCount value to 8.

C. Configure the model deployment settings to use an n1-standard-4 machine type and a GPU accelerator. Set the minReplicaCount value to 1 and the maxReplicaCount value to 4.

D. Configure the model deployment settings to use an n1-standard-8 machine type and a GPU accelerator.

Answer(s): B

---

12. You work for a pet food company that manages an online forum Customers upload photos of their pets on the forum to share with others About 20 photos are uploaded daily You want to automatically and in near real time detect whether each uploaded photo has an animal You want to prioritize time and minimize cost of your application development and deployment What should you do?

A. Send user-submitted images to the Cloud Vision API Use object localization to identify all objects in the image and compare the results against a list of animals.

B. Download an object detection model from TensorFlow Hub. Deploy the model to a Vertex Al endpoint. Send new user-submitted images to the model endpoint to classify whether each photo has an animal.

C. Manually label previously submitted images with bounding boxes around any animals Build an AutoML object detection model by using Vertex Al Deploy the model to a Vertex Al endpoint Send new user-submitted images to your model endpoint to detect whether each photo has an animal.

D. Manually label previously submitted images as having animals or not Create an image dataset on Vertex Al Train a classification model by using Vertex AutoML to distinguish the two classes Deploy the model to a Vertex Al endpoint Send new user-submitted images to your model endpoint to classify whether each photo has an animal.

Answer(s): A

---

13. 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.

Answer(s): B

---

14. You work for a global footwear retailer and need to predict when an item will be out of stock based on historical inventory dat a. 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.

Answer(s): A

---

15. 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 AI 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.

Answer(s): A

---

16. You developed a Python module by using Keras to train a regression model. You developed two model architectures, linear regression and deep neural network (DNN). within the same module. You are using the - raining_method argument to select one of the two methods, and you are using the Learning_rate-and num_hidden_layers arguments in the DNN. You plan to use Vertex AI's hypertuning service with a Budget to perform 100 trials. You want to identify the model architecture and hyperparameter values that minimize training loss and maximize model performance What should you do?

A. Run one hypertuning job for 100 trials. Set num hidden_layers as a conditional hypetparameter based on its parent hyperparameter training_mothod. and set learning rate as a non-conditional hyperparameter

B. Run two separate hypertuning jobs. a linear regression job for 50 trials, and a DNN job for 50 trials Compare their final performance on a common validation set. and select the set of hyperparameters with the least training loss

C. Run one hypertuning job with training_method as the hyperparameter for 50 trials Select the architecture with the lowest training loss. and further hypertune It and its corresponding hyperparameters for 50 trials

D. Run one hypertuning job for 100 trials Set num_hidden_layers and learning_rate as conditional hyperparameters based on their parent hyperparameter training method.

Answer(s): D

---

17. You are analyzing customer data for a healthcare organization that is stored in Cloud Storage. The data contains personally identifiable information (PII) You need to perform data exploration and preprocessing while ensuring the security and privacy of sensitive fields What should you do?

A. Use the Cloud Data Loss Prevention (DLP) API to de-identify the PI! before performing data exploration and preprocessing.

B. Use customer-managed encryption keys (CMEK) to encrypt the Pll data at rest and decrypt the Pll data during data exploration and preprocessing.

C. Use a VM inside a VPC Service Controls security perimeter to perform data exploration and preprocessing.

D. Use Google-managed encryption keys to encrypt the Pll data at rest, and decrypt the Pll data during data exploration and preprocessing.

Answer(s): A

---

18. You are developing a custom TensorFlow classification model based on tabular dat a. Your raw data is stored in BigQuery contains hundreds of millions of rows, and includes both categorical and numerical features. You need to use a MaxMin scaler on some numerical features, and apply a one-hot encoding to some categorical features such as SKU names. Your model will be trained over multiple epochs. You want to minimize the effort and cost of your solution. What should you do?

A. 1 Write a SQL query to create a separate lookup table to scale the numerical features.2. Deploy a TensorFlow-based model from Hugging Face to BigQuery to encode the text features.3. Feed the resulting BigQuery view into Vertex AI Training.

B. 1 Use BigQuery to scale the numerical features.2. Feed the features into Vertex AI Training.3 Allow TensorFlow to perform the one-hot text encoding.

C. 1 Use TFX components with Dataflow to encode the text features and scale the numerical features.2 Export results to Cloud Storage as TFRecords.3 Feed the data into Vertex AI Training.

D. 1 Write a SQL query to create a separate lookup table to scale the numerical features.2 Perform the one-hot text encoding in BigQuery.3. Feed the resulting BigQuery view into Vertex AI Training.

Answer(s): C

---

19. 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

Answer(s): A

---

20. You work for a retail company. You have created a Vertex AI 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.

Answer(s): B