Professional Machine Learning Engineer

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

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

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

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

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

Answer(s): A

---

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

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

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

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

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

Answer(s): C

---

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

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

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

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

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

Answer(s): C

---

4. 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.

Answer(s): B

---

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

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

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

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

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

Answer(s): A

---

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

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

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

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

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

Answer(s): C

---

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

A. Configure AutoML Tables to perform the classification task.

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

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

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

Answer(s): A

---

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

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

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

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

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

Answer(s): A

---

9. You are developing ML models with AI Platform for image segmentation on CT scans. You frequently update your model architectures based on the newest available research papers, and have to rerun training on the same dataset to benchmark their performance. You want to

minimize computation costs and manual intervention while having version control for your code. What should you do?

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

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

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

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

Answer(s): C

---

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

A. Categorical hinge

B. Binary cross-entropy

C. Categorical cross-entropy

D. Sparse categorical cross-entropy

Answer(s): D

---

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

A. Use the "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.

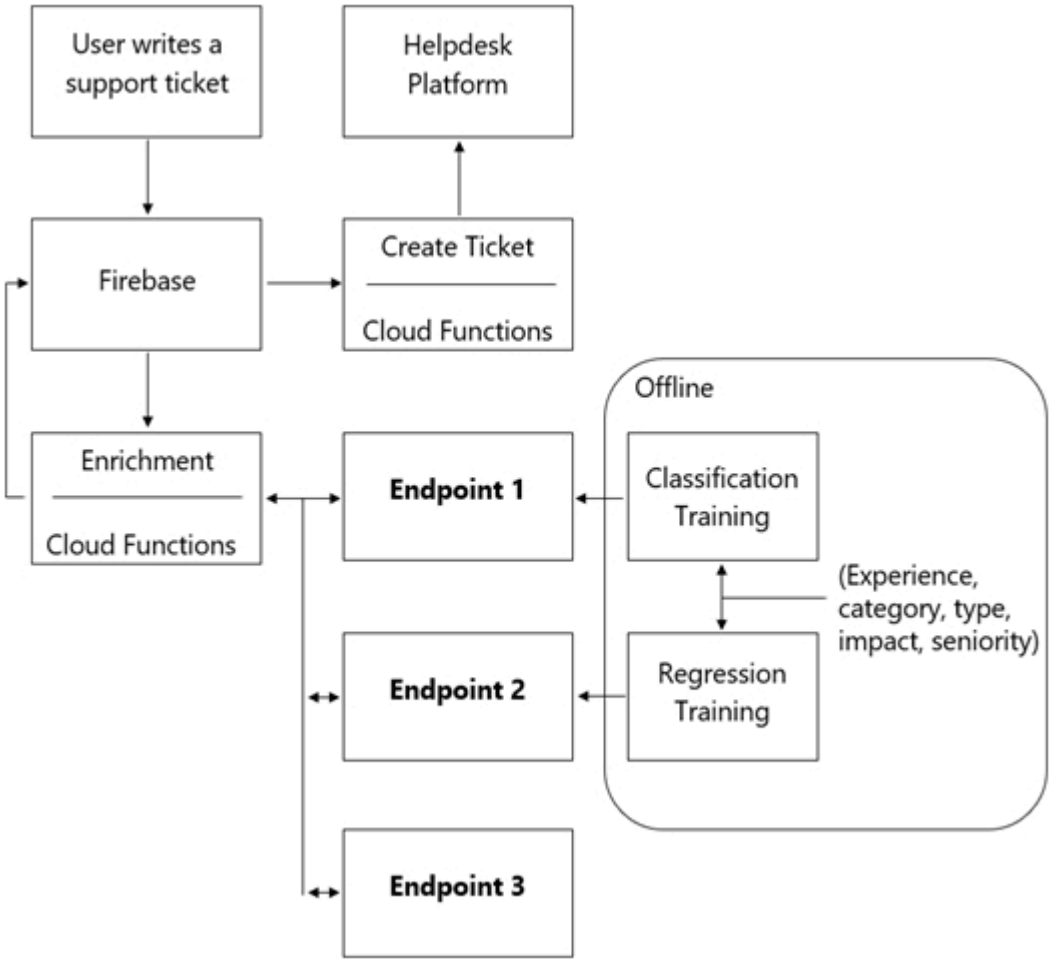
C. Import your user events and then your product catalog to make sure you have the highest quality event stream.

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.

Answer(s): B

---

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



Which endpoints should the Enrichment Cloud Functions call?

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

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

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

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

Answer(s): C

---

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

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

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

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

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

Answer(s): C

---

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

A. Poor data quality

B. Lack of model retraining

C. Too few layers in the model for capturing information

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

Answer(s): B

---

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

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

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

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

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

Answer(s): D

---

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

A. Classification

B. Reinforcement Learning

C. Recurrent Neural Networks (RNN)

D. Convolutional Neural Networks (CNN)

Answer(s): C

---

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

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

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

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

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

Answer(s): A

---

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

A. Manually combine all columns that contain a time signal into an array. AIlow 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. AIlow 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. AIlow AutoML to split your data based on the time signal provided, and reserve the more recent data for the validation and testing sets.

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

Answer(s): D

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

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

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

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

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

Answer(s): B

---

20. You are training an LSTM-based model on AI Platform to summarize text using the following job submission script:

```
gcloud ai-platform jobs submit training $JOB_NAME \
--package-path $TRAINER_PACKAGE_PATH \
--module-name $MAIN_TRAINER_MODULE \
--job-dir $JOB_DIR \
--region $REGION \
--scale-tier basic \
-- \
--epochs 20 \
--batch_size=32 \
--learning_rate=0.001 \
```

You want to ensure that training time is minimized without significantly compromising the accuracy of your model. What should you do?

A. Modify the ‘epochs’ parameter.

B. Modify the ‘scale-tier’ parameter.

C. Modify the ‘batch size’ parameter.

D. Modify the ‘learning rate’ parameter.

Answer(s): B
