Sample Question PMLE and Solutions – Exam Topic

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

Discussion:
https://cloud.google.com/solutions/building-anomaly-detection-dataflow-bigqueryml-dlp

A, because the in the question mentioned “store the results for analytics and visualisation”. Meanwhile, BigQuery is OLAP

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.

Discussion
Define the optimal route is the key

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.

Discussion
https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data#downsampling-and-upweighting

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.**

Discussion

PySpark is a Python API for Apache Spark.
Apache Spark is an open-source unified analytics engine for large-scale data processing, with built in modules for SQL, streaming, machine learning, and graph processing.
Also BigQuery is Serverless

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.

Discussion
https://cloud.google.com/ai-platform/training/docs/getting-started-pytorch

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

Discussion:
We need to extend, because we don't have any information with new product in model trained

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.

https://cloud.google.com/bigquery-ml/docs/logistic-regression-prediction

Discussion
Key: without writing code, EDA, feature selection. AutoML can be a codeless services

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.

Discussion
Not B because didn’t mention any data stored in BQ, D and C isn't suitable

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.

https://cloud.google.com/architecture/architecture-for-mlops-using-tfx-kubeflow-pipelines-and-cloud-build#cicd_architecture

Discussion

A. Cloud Storage aims not to store the code
B. Uneffective, not automated to use CLI
C. Automate
D. No need to use cloud composer

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**

https://stats.stackexchange.com/questions/326065/cross-entropy-vs-sparse-cross-entropy-when-to-use-one-over-the-other
https://machinelearningmastery.com/how-to-choose-loss-functions-when-training-deep-learning-neural-networks/

Discussion

Loss Function
Categorical Hinge is an alternative to cross-entropy for binary classification problems is the hinge loss function, primarily developed for use with Support Vector Machine (SVM) models. Can be used also for multiclass, less common tasks for Image Classification.
Binary cross-entropy or we can call just cross entropy is the default loss function to use for binary classification problems
Categorical cross-entropy is loss function for multi-class classification, when the labels are one hot-encoded. Examples (for a 3-class classification): [1,0,0] , [0,1,0], [0,0,1]
Sparse categorical cross-entropy is is a loss function for multi-class classification, when the labels are integers. Examples for above 3-class classification problem: [1], [2], [3]

Given the label(`['drivers_license', 'passport', 'credit_card']`), like `0`, `1`, and `2`, then Sparse Categorical Cross-Entropy is indeed the appropriate choice.