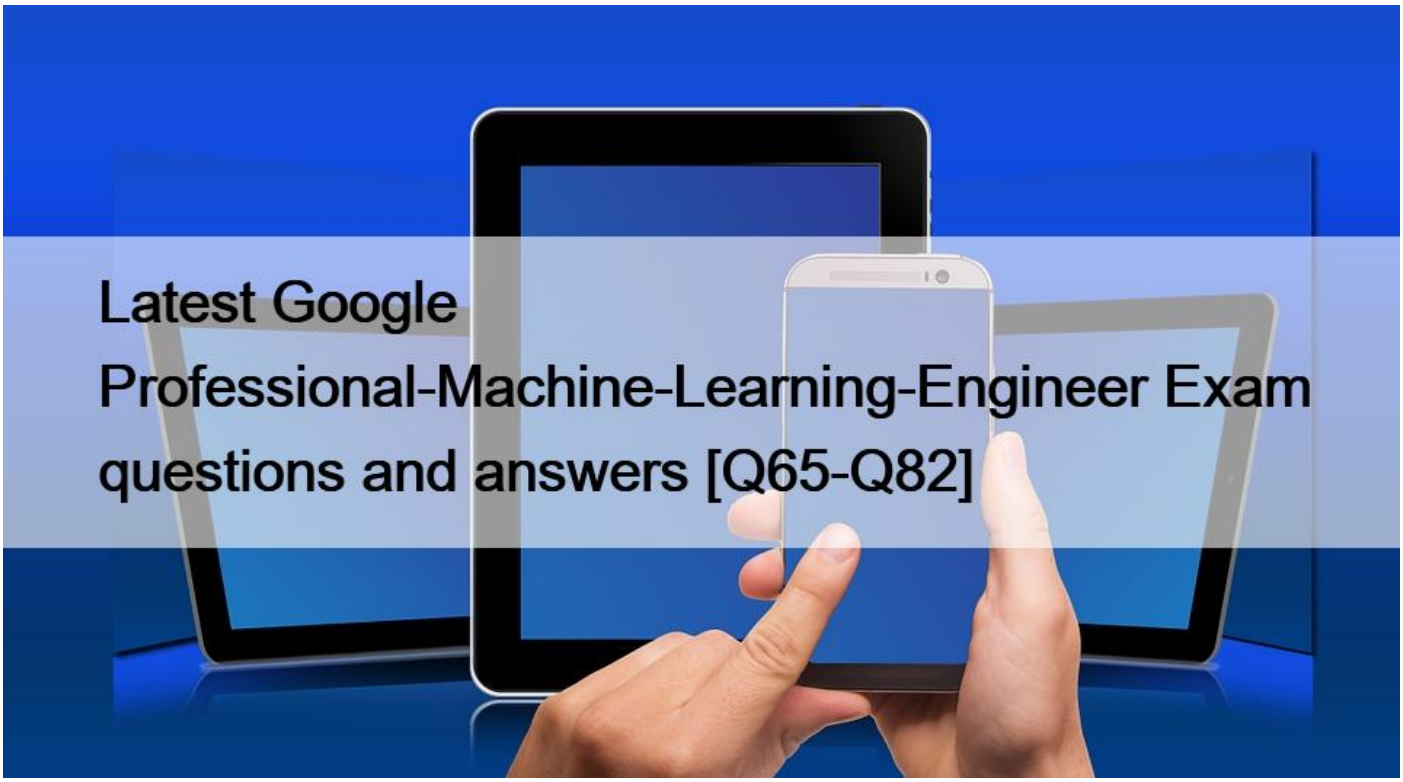
Latest Google Professional-Machine-Learning-Engineer Exam questions and answers [Q65-Q82



Latest Google Professional-Machine-Learning-Engineer Exam questions and answers
Test4Engine Professional-Machine-Learning-Engineer Exam Practice Test Questions (Updated 142 Questions)

How much Professional Machine Learning Engineer - Google Cost

The cost of the Professional Machine Learning Engineer - Google is $200. For more information related to exam price, please visit the official website Google Website as the cost of exams may be subjected to vary county-wise.

Understanding functional and technical aspects of Professional Machine Learning Engineer - Google Data Preparation and Processing

The following will be discussed in **Google Professional-Machine-Learning-Engineer exam dumps**:

- Encoding structured data types- Data exploration (EDA)- Class imbalance- Feature selection- Feature crosses- Managing large samples (TFRecords)- Data ingestion- Data validation- Database migration- Evaluation of data quality and feasibility- Monitoring/changing deployed pipelines- Design data pipelines- Handling missing data- Ingestion of various file types (e.g. Csv, json, img, parquet or databases, Hadoop/Spark)- Streaming data (e.g. from IoT devices)- Visualization- Feature engineering- Handling outliers- Statistical fundamentals at scale- Data leakage and augmentation **Q65.** A gaming company has launched an online game where people can start playing for free, but they need to pay if they choose to use certain features. The company needs to build an automated system to predict whether or not a new user will become a paid user within 1 year. The company has gathered a labeled dataset from 1 million users.

The training dataset consists of 1,000 positive samples (from users who ended up paying within 1 year) and

999,000 negative samples (from users who did not use any paid features). Each data sample consists of 200 features including user age, device, location, and play patterns.

Using this dataset for training, the Data Science team trained a random forest model that converged with over

99% accuracy on the training set. However, the prediction results on a test dataset were not satisfactory Which of the following approaches should the Data Science team take to mitigate this issue? (Choose two.)
* Add more deep trees to the random forest to enable the model to learn more features.
* Include a copy of the samples in the test dataset in the training dataset.
* Generate more positive samples by duplicating the positive samples and adding a small amount of noise to the duplicated data.
* Change the cost function so that false negatives have a higher impact on the cost value than false positives.
* Change the cost function so that false positives have a higher impact on the cost value than false negatives.

Q66. 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?
* Create the pipeline using Kubeflow Pipelines domain-specific language (DSL) and predefined Google Cloud components. Orchestrate the pipeline using Vertex AI Pipelines.
* Create the pipeline using TensorFlow Extended (TFX) and standard TFX components. Orchestrate the pipeline using Vertex AI Pipelines.
* 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.
* Create the pipeline using TensorFlow Extended (TFX) and standard TFX components. Orchestrate the pipeline using Kubeflow Pipelines deployed on Google Kubernetes Engine.

Q67. 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?
* Distribute the dataset with tf.distribute.Strategy.experimental_distribute_dataset
* Create a custom training loop.
* Use a TPU with tf.distribute.TPUStrategy.
* Increase the batch size.

Q68. A data scientist wants to use Amazon Forecast to build a forecasting model for inventory demand for a retail company. The company has provided a dataset of historic inventory demand for its products as a .csv file stored in an Amazon S3 bucket. The table below shows a sample of the dataset.

timestamp	item_id	demand	category	lead_time
2019-12-14	uni_000736	120	hardware	90
2020-01-31	uni_003429	98	hardware	30
2020-03-04	uni_000211	234	accessories	10

How should the data scientist transform the data?
* Use ETL jobs in AWS Glue to separate the dataset into a target time series dataset and an item metadata dataset. Upload both

datasets as .csv files to Amazon S3.
* Use a Jupyter notebook in Amazon SageMaker to separate the dataset into a related time series dataset and an item metadata dataset. Upload both datasets as tables in Amazon Aurora.
* Use AWS Batch jobs to separate the dataset into a target time series dataset, a related time series dataset, and an item metadata dataset. Upload them directly to Forecast from a local machine.
* Use a Jupyter notebook in Amazon SageMaker to transform the data into the optimized protobuf recordIO format. Upload the dataset in this format to Amazon S3.

Q69. 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?
* Use the Vertex AI Training to submit training jobs using any framework.
* Configure Kubeflow to run on Google Kubernetes Engine and submit training jobs through TFJob.
* Create a library of VM images on Compute Engine, and publish these images on a centralized repository.
* Set up Slurm workload manager to receive jobs that can be scheduled to run on your cloud infrastructure.

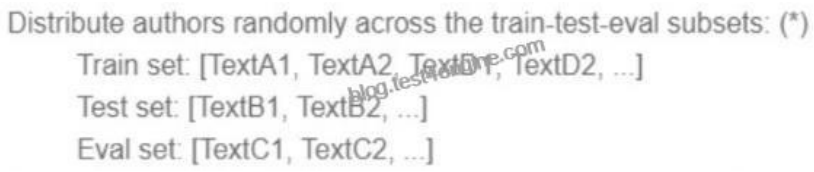
Q70. Your team is working on an NLP research project to predict political affiliation of authors based on articles they have written. You have a large training dataset that is structured like this:

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

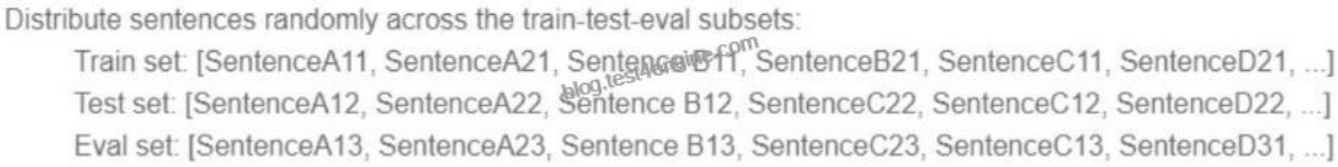

A)

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

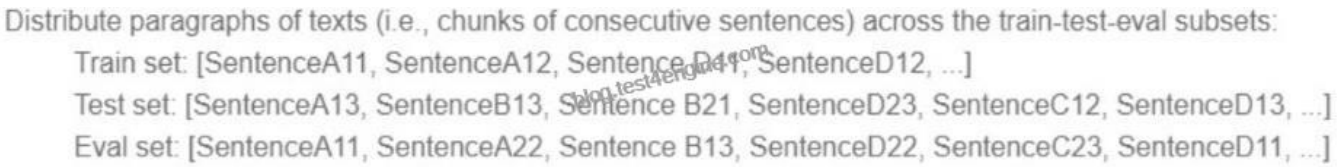
B)



C)



D)



* Option A
* Option B
* Option C
* Option D

Q71. An online reseller has a large, multi-column dataset with one column missing 30% of its data. A Machine Learning Specialist believes that certain columns in the dataset could be used to reconstruct the missing data.

Which reconstruction approach should the Specialist use to preserve the integrity of the dataset?
* Listwise deletion
* Last observation carried forward
* Multiple imputation
* Mean substitution

Explanation/Reference: https://worldwidescience.org/topicpages/i/imputing+missing+values.html

Q72. 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?
* 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.

* 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.
* 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
* 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

Q73. You are developing ML models with Al 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?
* Use Cloud Functions to identify changes to your code in Cloud Storage and trigger a retraining job
* Use the gcloud command-line tool to submit training jobs on Al Platform when you update your code
* Use Cloud Build linked with Cloud Source Repositories to trigger retraining when new code is pushed to the repository
* Create an automated workflow in Cloud Composer that runs daily and looks for changes in code in Cloud Storage using a sensor.

Q74. 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?
* Use the TFX ModelValidator tools to specify performance metrics for production readiness
* Use k-fold cross-validation as a validation strategy to ensure that your model is ready for production.
* Use the last relevant week of data as a validation set to ensure that your model is performing accurately on current data
* Use the entire dataset and treat the area under the receiver operating characteristics curve (AUC ROC) as the main metric.

Q75. A Machine Learning Specialist working for an online fashion company wants to build a data ingestion solution for the company’s Amazon S3-based data lake.

The Specialist wants to create a set of ingestion mechanisms that will enable future capabilities comprised of:

* Real-time analytics

* Interactive analytics of historical data

* Clickstream analytics

* Product recommendations

Which services should the Specialist use?
* AWS Glue as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for real- time data insights; Amazon Kinesis Data Firehose for delivery to Amazon ES for clickstream analytics; Amazon EMR to generate personalized product recommendations
* Amazon Athena as the data catalog: Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for near-real-time data insights; Amazon Kinesis Data Firehose for clickstream analytics; AWS Glue to generate personalized product recommendations
* AWS Glue as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for historical data insights; Amazon Kinesis Data Firehose for delivery to Amazon ES for clickstream analytics; Amazon EMR to generate personalized product recommendations
* Amazon Athena as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for historical data insights; Amazon DynamoDB streams for clickstream analytics; AWS Glue to generate personalized product recommendations

Q76. You need to design a customized deep neural network in Keras that will predict customer purchases based on their purchase history. You want to explore model performance using multiple model architectures, store training data, and be able to compare the evaluation metrics in the same dashboard. What should you do?
* Create multiple models using AutoML Tables
* Automate multiple training runs using Cloud Composer
* Run multiple training jobs on Al Platform with similar job names
* Create an experiment in Kubeflow Pipelines to organize multiple runs
https://www.kubeflow.org/docs/components/pipelines/concepts/experiment/
https://www.kubeflow.org/docs/components/pipelines/concepts/run/

Q77. A Machine Learning team runs its own training algorithm on Amazon SageMaker. The training algorithm requires external assets. The team needs to submit both its own algorithm code and algorithm-specific parameters to Amazon SageMaker.

What combination of services should the team use to build a custom algorithm in Amazon SageMaker?

(Choose two.)
* AWS Secrets Manager
* AWS CodeStar
* Amazon ECR
* Amazon ECS
* Amazon S3

Q78. A Machine Learning Specialist is developing a custom video recommendation model for an application. The dataset used to train this model is very large with millions of data points and is hosted in an Amazon S3 bucket.

The Specialist wants to avoid loading all of this data onto an Amazon SageMaker notebook instance because it would take hours to move and will exceed the attached 5 GB Amazon EBS volume on the notebook instance.

Which approach allows the Specialist to use all the data to train the model?
* Load a smaller subset of the data into the SageMaker notebook and train locally. Confirm that the training code is executing and the model parameters seem reasonable. Initiate a SageMaker training job using the full dataset from the S3 bucket using Pipe input mode.
* Launch an Amazon EC2 instance with an AWS Deep Learning AMI and attach the S3 bucket to the instance. Train on a small amount of the data to verify the training code and hyperparameters. Go back to Amazon SageMaker and train using the full dataset
* Use AWS Glue to train a model using a small subset of the data to confirm that the data will be compatible with Amazon SageMaker. Initiate a SageMaker training job using the full dataset from the S3 bucket using Pipe input mode.
* Load a smaller subset of the data into the SageMaker notebook and train locally. Confirm that the training code is executing and the model parameters seem reasonable. Launch an Amazon EC2 instance with an AWS Deep Learning AMI and attach the S3 bucket to train the full dataset.

Q79. 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?
* 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.
* This is not a good result because the model is performing worse than predicting that people will always renew their subscription.
* This is a good result because predicting those who cancel their subscription is more difficult, since there is less data for this group.
* This is a good result because the accuracy across both groups is greater than 80%.

Q80. 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?
* 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
* 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
* 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
* 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

Q81. A logistics company needs a forecast model to predict next month’s inventory requirements for a single item in

10 warehouses. A machine learning specialist uses Amazon Forecast to develop a forecast model from 3 years of monthly data. There is no missing data. The specialist selects the DeepAR+ algorithm to train a predictor. The predictor means absolute percentage error (MAPE) is much larger than the MAPE produced by the current human forecasters.

Which changes to the CreatePredictor API call could improve the MAPE? (Choose two.)
* Set PerformAutoML to true.
* Set ForecastHorizon to 4.
* Set ForecastFrequency to W for weekly.
* Set PerformHPO to true.
* Set FeaturizationMethodName to filling.
Explanation/Reference: https://docs.aws.amazon.com/forecast/latest/dg/forecast.dg.pdf

Q82. 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?
* Increase the size of the training batch
* Decrease the size of the training batch
* Increase the learning rate hyperparameter
* Decrease the learning rate hyperparameter
https://developers.google.com/machine-learning/crash-course/introduction-to-neural-networks/playground-exercises