Engineer Data in Google Cloud: Challenge Lab

GSP327

Overview

In a challenge lab you're given a scenario and a set of tasks. Instead of following step-by-step instructions, you will use the skills learned from the labs in the quest to figure out how to complete the tasks on your own! An automated scoring system (shown on this page) will provide feedback on whether you have completed your tasks correctly.

When you take a challenge lab, you will not be taught new Google Cloud concepts. You are expected to extend your learned skills, like changing default values and reading and researching error messages to fix your own mistakes.

To score 100% you must successfully complete all tasks within the time period!

This lab is recommended for students who have enrolled in the Data Engineering quest. Are you ready for the challenge?

Topics tested:

- Create a new BigQuery table from existing data
- Clean data for ML Model using BigQuery, Dataprep or Dataflow
- Build and tune a model in BQML
- Perform a batch prediction into a new table with BQML

Setup

Before you click the Start Lab button

Read these instructions. Labs are timed and you cannot pause them. The timer, which starts when you click **Start Lab**, shows how long Google Cloud resources will be made available to you.

This Qwiklabs hands-on lab lets you do the lab activities yourself in a real cloud environment, not in a simulation or demo environment. It does so by giving you new, temporary credentials that you use to sign in and access Google Cloud for the duration of the lab.

What you need

To complete this lab, you need:

- Access to a standard internet browser (Chrome browser recommended).
- Time to complete the lab.
  Note: If you already have your own personal Google Cloud account or project, do not use it for this lab.

  Note: If you are using a Pixelbook, open an Incognito window to run this lab.

Challenge scenario

You have started a new role as a Data Engineer for TaxiCab Inc. You are expected to import some historical data to a working BigQuery dataset, and build a basic model that predicts fares based on information available when a new ride starts. Leadership is interested in building an app and estimating for users how much a ride will cost. The source data will be provided in your project.

You are expected to have the skills and knowledge for these tasks, so don't expect step-by-step guides to be provided.

Your challenge

As soon as you sit down at your desk and open your new laptop you receive your first assignment: build a basic BQML fare prediction model for leadership. Perform the following tasks to import and clean the data, then build the model and perform batch predictions with new data so that leadership can review model performance and make a go/no-go decision on deploying the app functionality.

Task 1: Clean your training data

You've already completed the first step, and have created a dataset `taxirides` and imported the historical data to table, `historical_taxi_rides_raw`. This is data prior for rides to 2015.

You may need to wait 1-3 minutes for the data to be fully populated in your project.
To complete this task you will need to:

- Clean the data in `historical_taxi_rides_raw` and make a copy to `taxi_training_data` in the same dataset. You can use BigQuery, DataPrep, DataFlow, etc. to create this table and clean the data. Make sure your target column is called `fare_amount`.
  Some helpful hints:

- You can see the source dataset in the BQ UI - familiarize yourself with the source schema first.
- As a hint for the data avilable at prediction time, familiarize yourself with the table `taxirides.report_prediction_data` which shows the format data will arrive at prediction time.
  Data Cleaning Tasks:

- Ensure `trip_distance` is greater than 0.
- Remove rows were `fare_amount` is very small (less than $2.5 for example).
- Ensure that the latitudes and longitudes are reasonable for the use case.
- Ensure `passenger_count` is greater than 0.
- Be sure to add `tolls_amount` and to `fare_amount` as the target variable since `total_amount` includes tips.
- Because the source dataset is large (>1 Billion rows), sample the dataset to less than 1 Million rows.
- Only copy fields that will be used in your model (`report_prediction_data` is a good guide).
  Click *Check my progress* to verify the objective.

If you don't get a green check mark, please click on the Score fly-out on the top right and click **Run Step** on the relevant step. You will see a hint pop up giving you advice.

Task 2: Create a BQML model called `taxirides.fare_model`

Based on the data you have in `taxirides.taxi_training_data`, build a BQML model that predicts `fare_amount`. Call the model `taxirides.fare_model`. Your model will need an RMSE of 10 or less to complete the task.

Some helpful hints:

- You can encapsulate any additional data transformationsn in a TRANSFORM() clause
- Keep in mind, only features in the `TRANSFORM()` clause will be passed to the model. You can use a `* EXCEPT(feature_to_leave_out)` to pass some or all of the features without explicitly calling them

- `ST_distance()` and `ST_GeogPoint()` GIS functions in BigQuery can be used to easily calculate euclidean distance (i.e. how far pickup to dropoff did the taxi travel):

```
ST_Distance(ST_GeogPoint(pickuplon, pickuplat), ST_GeogPoint(dropofflon, dropofflat))
AS euclidean
```

Click *Check my progress* to verify the objective.

If you don't get a green check mark, please click on the Score fly-out on the top right and click **Run Step** on the relevant step. You will see a hint pop up giving you advice.

Task 3: Perform a batch prediction on new data

Leadership is curious to see how well your model performs over new data, in this case, all of the data they've collected in 2015. This data is in `taxirides.report_prediction_data`. Only values known at prediction time are included in the table.

Use `ML.PREDICT` and your model to predict `fare_amount` and store your results in a table called `2015_fare_amount_predictions`.

Click *Check my progress* to verify the objective.

If you don't get a green check mark, please click on the Score fly-out on the top right and click **Run Step** on the relevant step. You will see a hint pop up giving you advice.

Congratulations!

Earn Your Next Skill Badge

This self-paced lab is part of the Engineer Data in Google Cloud quest. Completing this skill badge quest earns you the badge above, to recognize your achievement. Share your badge on your resume and social platforms, and announce your accomplishment using #GoogleCloudBadge.
This skill badge is part of Google Cloud's Data Engineer learning path. If you have already completed the other skill badge quests in this learning path, search the catalog for 20+ other skill badge quests in which you can enroll.

Google Cloud Training & Certification

...helps you make the most of Google Cloud technologies. Our classes include technical skills and best practices to help you get up to speed quickly and continue your learning journey. We offer fundamental to advanced level training, with on-demand, live, and virtual options to suit your busy schedule. Certifications help you validate and prove your skill and expertise in Google Cloud technologies.
Manual Last Updated May 4, 2021
Lab Last Tested June 11, 2020

solution:
https://www.youtube.com/watch?v=82su32ztgto
https://github.com/TechieZilla/Qwiklabs/blob/main/Engineer%20Data%20in%20Google%20Cloud:%20Challenge%20Lab%20%5BGSP327%5D