Create ML Models with BigQuery ML: Challenge Lab

GSP341

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 Create ML Models with BigQuery ML quest. Are you ready for the challenge?
Topics tested

- Create a new BigQuery dataset which will store your BigQuery ML models.
- Create forecasting (linear regression) models in BigQuery ML.
- Evaluate the performance of your machine learning models.
- Make predictions of trip duration using BigQuery ML models.

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 junior member of the Data Science department Jooli Inc. Your team is working on a number of machine learning initiatives related to urban mobility services. You are expected to help with the development and assessment of data sets and machine learning models to help provide insights based on real work data sets.

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

One of the projects you are working on needs to provide analysis based on real world data that will help in the selection of new bicycle models for public bike share systems. Your role in this project is to develop and evaluate machine learning models that can predict average trip durations for bike schemes using the public data from Austin's public bike share scheme to train and evaluate your models.

Two of the senior data scientists in your team have different theories on what factors are important in determining the duration of a bike share trip and you have been asked to prioritise these to start. The first data scientist maintains that the key factors are the start station, the location of the start station, the day of the week and the hour the trip started. While the second data scientist argues that this is an over complication and the key factors are simply start station, subscriber type, and the hour the trip started.

You have been asked to develop a machine learning model based on each of these input features. Given the fact that stay-at-home orders were in place for Austin during parts of 2020 as a result of COVID-19 you will be working on data from previous years. You have been instructed to train your models on data from 2018 and then evaluate them against data from 2019 on the basis of Mean Absolute Error and the square root of Mean Squared Error.

You can access the public data for the Austin bike share scheme in your project by opening this link to the Austin bike share dataset in the browser tab for your lab.
As a final step you must create and run a query that uses the model that includes subscriber type as a feature, to predict the average trip duration for all trips from the busiest bike sharing station in 2019 (based on the number of trips per station in 2019) where the subscriber type is 'Single Trip'.

Task 1: Create a dataset to store your machine learning models

Create a new dataset in which you can store your machine learning models.

Check a new dataset has been created.

Check my progress

If you don't get a green check mark, click on the **Score** fly-out on the top right and click **Run Step** on the relevant step. A hint pop up opens to give you advice.

Task 2: Create a forecasting BigQuery machine learning model

Create the first machine learning model to predict the trip duration for bike trips. The features of this model must incorporate the starting station name, the hour the trip started, the weekday of the trip, and the address of the start station labeled as `location`. You must use 2018 data only to train this model.

Check that a BigQuery machine learning model has been created.

Check my progress

If you don't get a green check mark, click on the **Score** fly-out on the top right and click **Run Step** on the relevant step. A hint pop up opens to give you advice.

Task 3: Create the second machine learning model

Create the second machine learning model to predict the trip duration for bike trips. The features of this model must incorporate the starting station name, the bike share subscriber type and the start time for the trip. You must also use 2018 data only to train this model.

Check that a second BigQuery machine learning model has been created.

Check my progress

If you don't get a green check mark, click on the **Score** fly-out on the top right and click **Run Step** on the relevant step. A hint pop up opens to give you advice.

Task 4: Evaluate the two machine learning models

Evaluate each of the machine learning models against 2019 data only using separate queries. Your queries must report both the Mean Absolute Error and the Root Mean Square Error.

Confirm that both machine learning models have been evaluated.

Check my progress

If you don't get a green check mark, click on the **Score** fly-out on the top right and click **Run Step** on the relevant step. A hint pop up opens to give you advice.

Task 5: Use the subscriber type machine learning model to predict average trip durations

When both models have been created and evaulated, use the second model, that uses `subscriber_type` as a feature, to predict average trip length for trips from the busiest bike sharing station in 2019 where the subscriber type is `Single Trip`.

Check that predictions have been made successfully for single trip subscribers at the busiest bike hire station in 2019.

Check my progress

If you don't get a green check mark, click on the **Score** fly-out on the top right and click **Run Step** on the relevant step. A hint pop up opens to give you advice.

Tips and Tricks

- **Tip 1**. You will need to combine the information from both Austin bike share tables in the public dataset to create your first model by means of a JOIN statement.

- **Tip 2**. You must train both models using 2018 data only and evaluate the models using 2019 data only.

- **Tip 3**. You must choose a model type that is suitable for forecasting label values. When evaluating the models you should use `SELECT SQRT(mean_squared_error) AS rmse, mean_absolute_error FROM ML.EVALUATE(...)` to return the specific model performance metrics the data scientists want to use.

- **Tip 4**. Your prediction queries must return the average of the predicted value output by the model for the trip duration and not just the average of the actual trip duration.

Congratulations!

Earn Your Next Skill Badge

This self-paced lab is part of the Create ML Models with BigQuery ML skill badge 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 quest is part of Google Cloud's Data Analyst 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 that you can enroll in.

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 April 14, 2021
Lab Last Tested April 14, 2021

Solution:

https://www.youtube.com/watch?v=Upnsuq2Lg1I

https://github.com/GirishSharma5956/gsp341-new/blob/main/gsp341.txt