BigQuery ML Assignment

Introduction

This assignment will help you practice key BigQuery ML concepts, including connecting to public datasets, creating and evaluating machine learning models, and making predictions. You will use the New York City Yellow Taxi Trip dataset available in BigQuery to generate a fare prediction model.

All tasks should be completed using BigQuery's SQL interface to stay consistent with the learning objectives.

Dataset Information

You will use the New York City Yellow Taxi Trip dataset from BigQuery public datasets:

1. Project: 'bigquery-public-data'
2. Dataset: 'new_york'
3. Table: 'tlc_yellow_trips_2015'

Dataset Setup

Create a new BigQuery project and connect it to the specified dataset. All tasks should be completed within this project. Note: The dataset is very large, so we will work with a sample of the data for this assignment. Use SQL to randomly select 0.1% (1/1000) of the task dataset. You can achieve this by using the RAND() function to generate random values and filter rows accordingly. (Hint: WHERE RAND() < 0.001)

Tasks

Question 1: Exploring the Dataset

Objective: Understand the structure and contents of the NYC Yellow Taxi Trip dataset.

Instructions:

1. Data Exploration:
   - Write a SQL query to display the first 1000 rows of the 'tlc_yellow_trips_2015' table.
   - Identify and list the key columns that could be relevant for predicting taxi fares.

Hints:

Focus on columns such as 'trip_distance,' 'pickup_datetime,' 'dropoff_datetime,' 'passenger_count,' 'tolls_amount,' and 'fare_amount'.

Question 2: Data Preprocessing

Objective: Prepare the data for machine learning by handling missing values

Instructions:

1. Data Cleaning:
   - Write a SQL query to count the number of rows with missing values in key columns identified in Question 1.
   - Remove rows with missing values in 'trip_distance,' 'passenger_count,' 'tolls_amount,' and 'fare_amount'.

Hints:

Use 'IS NULL' and comparison operators to identify and filter out problematic rows.

Question 3: Feature Engineering

Objective: Create new features to improve the predictive power of the model.

Instructions:

1. Feature Creation:
   - Calculate the total fare by adding tolls_amount and fare_amount (AS total_fare), which will be used as the target for prediction.
   - Calculate the trip duration in minutes (AS trip_duration) by computing the difference between 'dropoff_datetime' and 'pickup_datetime'.
   - Extract the hour of the day from 'pickup_datetime' (AS pickup_hour) to capture potential time-based patterns.
   - Create a new feature for the day of the week from 'pickup_datetime' (AS pickup_day_of_week).

Hints:

For these calculations, use BigQuery's 'TIMESTAMP_DIFF,' 'EXTRACT,' and 'CAST' functions.

Question 4: Creating the Machine Learning Model

Objective: Build a linear regression model to predict taxi fares.

Instructions:

1. Model Creation:
   - Use the 'CREATE MODEL' statement to create a linear regression model named 'taxi_fare_model'.
   - Define the total_fare (calculated as tolls_amount + fare_amount) as the target.
   - Include the following features in the model: 'trip_distance,' 'trip_duration,' 'passenger_count,' 'pickup_hour,' and 'pickup_day_of_week.'
   - Split the data into training (80%) and testing (20%) sets using BigQuery ML's built-in split capabilities.

Hints:

```
CREATE OR REPLACE MODEL `your_project.your_dataset.taxi_fare_model`
OPTIONS(
   model_type = 'linear_reg',
   input_label_cols = ['total_fare'],
   DATA_SPLIT_METHOD = 'RANDOM',
   DATA_SPLIT_EVAL_FRACTION = 0.2
)
```

Question 5: Evaluating the Model

Objective: Assess the performance of the trained model.

Instructions:

1. Model Evaluation:
   - Use the 'ML.EVALUATE' function to evaluate the 'taxi_fare_model'.
   - Record metrics such as Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE).
   - Interpret these metrics to assess the model's performance.

Hints:

1. Lower MAE and RMSE values indicate better model performance.
2. Use a subset of the dataset not used in training for prediction to evaluate model generalization.

Submission Instructions

1. Share your BigQuery project link with 113356042@g.nccu.edu.tw by November 27, 2024, at 11:59 PM, ensuring the TA has viewing permissions.
2. Upload your sql file to the Moodle by November 27, 2024, at 11:59 PM