ML Project Assignment

Summary

In Lab 11, you have learned how to create and evaluate a machine learning model, then predict using the model. In this project you are asked to practice the skills on a different dataset - Chicago Taxi Trips, this dataset includes taxi trips from 2013 to the present, reported to the City of Chicago.

Your task here is to

1. use the python scikit-learn to investigate the similarities and relationships that PCA can provide. It will not be possible for you to analyze the entire data set in this way so you should perform your analysis on *the first 1000 rows* of the table. Ideally the output of this phase should guide the direction you take in the second, ML, part of the assignment. You will need to decide which columns should be part of your analysis and which should be ignored. See some of the tutorial links we have posted in the lecture slides for help in deciding what are appropriate columns to consider.

2. based on the information you gained from step 1, create a model[1] to predict the taxi fare ("fare" column in the dataset). Note that you may need to clean the data, pick a list of features (feature engineering), and then design your model.

Please email your zipped solution pack to the lecturer by the end of week 14 with subject "cs6502: ml proj".

Note that your solution pack should contain

- the query you use to selected the first 1000 rows
- key steps for your PCA (setup, command, etc)
- the sql for model creation
- the sql to evaluate the model
- the sql to predict using the model above
- link of the BigQuery commands you composed[2]
- screenshot of the model evaluation report[3]

You will be assessed on the quality of the information you glean and feed into the ML part and how you present this to us. Again, see how this is done from the tutorials. (Note, as the tutorials make clear the R statistical package has very good support for this type of analysis. If you wish to take this task on using R then this will be ok, too.)

Clean Up

After finishing your assignment, do make sure to clean up (delete the dataset you created) to avoid costs.

[1] https://cloud.google.com/bigquery-ml/docs/bigqueryml-intro#supported_models_in

[2] https://cloud.google.com/bigquery/docs/saving-sharing-queries

[3] in the BigQuery console, click on the model your created in the left panel, the model details is present in the bottom right window, click the "Evaluation" tab to get the model evaluation report