Machine Learning Challenge

1. Problem Statement

Use GYC Green Taxi data to implement a machine learning system to predict the expected tip amount for a trip.

2. Data

2017 GYC Green Taxi Data: https://data.cityofnewyork.us/Transportation/2017-Green-Taxi-Trip-Data/5gj9-2kzx

You may use the Jan 2017 data for your model training and Feb 2017 data for model evaluation.

3. Specification of Deliverables

3.1 Data Pre-Processing Code

Use Spark (in Java, Scale, or PySpark) for data pre-processing (cleansing, transformation, splitting traning/test dataset etc). Save the preprocessed data.

3.2 Machine Learning Model and Model Serving

Build a machine learning model based on pre-processed dataset. The machine learning code might include following parts:

- Model Training: input will be pre-processed training dataset; output will a machine learning models to solve the problem described in Section 1.
    - You are encouraged to use SparkML to train the model. However, other library or framework is acceptable.
    - Export trained model as a file.
- Batch Scoring Code: code to call your model in offline batch mode.
    - Take a data file as input
    - Predict the tips for trip data stored in the input file and save results to an output file.

3.3 Report

Write a short report with regards to the data pre-processing, model training, and batch scoring, which shall include the following topics:

- Discuss of data and steps you put into your pre-processing pipelines.

- Discuss of batch scoring pipelines and any optimization can be done to improve the production runtime performance with the increase of the production workload.

4. Assessment Criteria

4.1 General Criteria

We will assess the solution and deliverables based on rationality of overall solution over specific technologies and model accuracy. Good overall understanding of machine learning process is valued. Spending effort to tuning the model to gain 1% more of accuracy will be appreciated but not compulsory.

4.2 Solution and Coding Guidelines

The whole solution will be assessed under below criteria and constrains:

- Programming Language: Python, Java, or Scala
- Library/Framework: Please use Spark whenever possible.
- Good coding practices with respect to the coding language chosen