Capstone Project Submission

Instructions:
i) Please fill in all the required information.
ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Sanjay Yadav
neer.ping@gmail.com

Contribution :

1. Data Analysis

2. Adding New Features

3. Feature Engineering and Selection

4. Regression Model Building

5. Model Validation & Selection

Please paste the GitHub Repo link.

Github Link:- https://github.com/sanjay2097/NYC-Taxi-Trip-Time-Prediction

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Problem Statement :

Our task is to build a model that predicts the total ride duration of taxi trips in New York City. The dataset is based on the 2016 NYC Yellow Cab trip record data made available in Big Query on Google Cloud Platform. The data was originally published by the NYC Taxi and Limousine Commission (TLC). The data was sampled and cleaned for the purposes of this project. Based on individual trip attributes, we should predict the duration of each trip in the test set.

NYC Taxi Data.csv - the training set (contains 1458644 trip records)

Approach :

In this project we covered various aspects of the Machine learning development cycle. We observed that the data exploration and variable analysis is a very important aspect of the whole cycle and should be done for thorough understanding of the data. We also cleaned the data while exploring as there were some outliers which should be treated before feature engineering. Further we did feature engineering to filter and gather only the

optimal features which are more significant and covered most of the variance in the dataset. Then finally we trained the models on the optimum featureset to get the results and validated the models through cross-checking performance metrics to find the optimum regression model for our dataset.

Conclusion :

- XGBoost regression model has the best performance metrics Adjusted R2 of 82% on training data and 80% on test data.
- The top 5 features according to the XGBoost model were : Distance,hour_5,hour_4,hour_1 and hour_6.
- Boosting regression models have performed much better than other models.