Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email, and Contribution:
1) Yogesh Reddy Yogeshreddy005@gmail.com • Data Wrangling • Exploratory Data Analysis • Handling outlier data • Data Preprocessing • Model Building

Please paste the GitHub Repo link.

Github Link :https://github.com/Yogs005/NYC_Taxi_Trip_Prediction

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

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 NYC Taxi and Limousine commission (TLC). The data totally has 1458644 rows and 10 columns.

Some of the features have outliers and each of the feature is analyzed carefully and we have handled the outliers using 1.5 rule where we used IQR. We have visualized the pickup latitude and longitude and drop-off longitude and latitude using scatter plot . we have added new features for pickup day and drop off day, pickup month and drop off month. For this we have used pickup_datetime and dropoff_datetime feature, we have extracted datetime using pandas datetime feature. We have also added two new features pickup_timezone and dropoff_timezone, this two features have five time zones based on pickup and drop off times. We have found several insights using all these features and plotted several plots on pickups and drop offs based on time zones and number of pickups and drop offs based on week days and total pickups and drop offs based on month. We have added new feature which is distance based on pickup and drop off latitude and longitude we found the distance using geopy library

. We have normalized distance and trip_duration for the model building process. We used heatmap to visualize the correlation between each feature. For model building we have used train test split to split data for training and testing. We have used Linear regression, Decision Tree, random forest and boosting methods Adaboost, gradient boosting and Xgboost. Out of all this models Randomforest gave better accuracy of 90%. As it was very big data we faced difficulties in performing some operations but this helped us a lot to learn several things.