Capstone Project 2
NYC TAXI TIME PREDICTION

Team Members

Pruthvi Raj
Yogesh Reddy
Piyush A. Lanjewar

INTRODUCTION:

Given the data of Taxi trip of entire NYC for the year 2016(contains ~1.5 mil trip records) where we need to closely analyse the data and predict the total duration of taxi trips based on the previous records and shows user the estimated time that particular trip takes.

PROBLEM STATEMENT:

The taxi trip data provided by leading taxi providers in NYC which are NYC Taxi, Limousine Commission which as more than 1.4 million records of taxi trips in the NYC for the year 2016. The following data which includes geometric locations of both pickup and drop. By analysing the provided data and predict the estimated time taken for each trip.

Data Summary:

The provided data consists of 1.4mil records of taxi trips across NYC in 2016. We need to filter a lot of data to move ahead in the visualization part and in the analysing part.

- There are 11 columns and 1458644 rows.

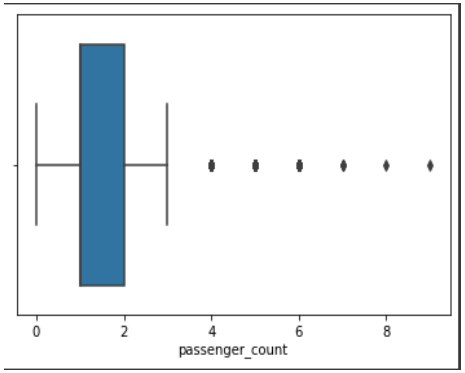
Column Features:

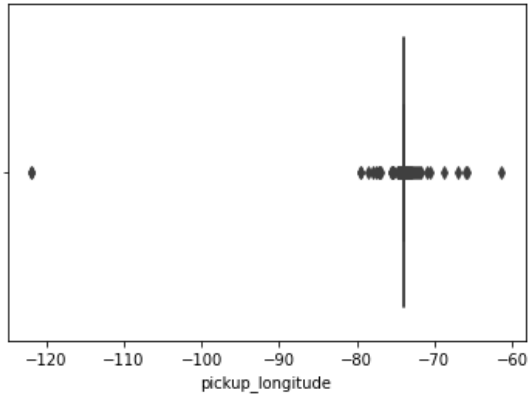
- ID : Provides the respective id foe each taxi booking.
- Vendor ID : Specifies the record from which source(1 - NYC Taxi, 2- Limousine Commission)
- Pickup_datetime : Date and Time of pickup data
- Dropoff_datetime : Date and Time of Dropoff data
- Passenger_count : Indicates the No. of Passengers in the trip
- Pickup_longitude : longitude of the pickup location
- Pickup_latitude : latitude of pickup location
- Dropoff_longitude : longitude of drop location
- Dropoff_latitude : latitude of drop location
- Store_and_fwd_flag : This flag indicates whether the trip was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server(Y - dStore and Forward, N - not Store and Forward)
- Trip_duration : Duration of the trip

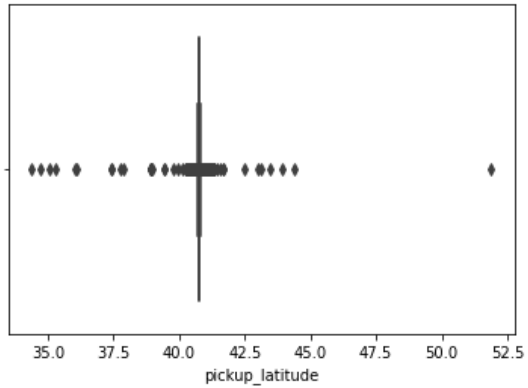
EDA

- The given data has total of 10 features where seven where numerical features.
- There were no null values in the dataset.
- There were outliers in the following features
  - Passenger Count
  - Pickup latitude
  - Pickup longitude
  - Dropoff latitude
  - Dropoff longitude
  - Trip Duration

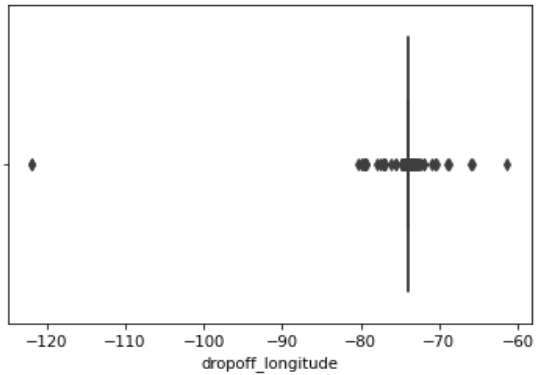
Outliers:

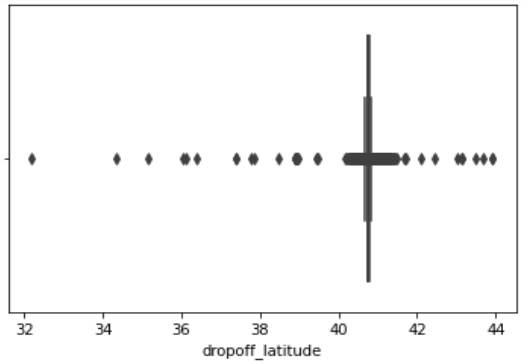




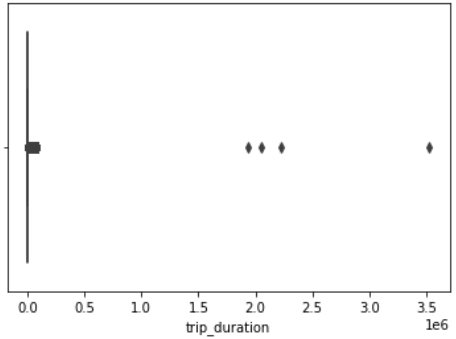


Passenger Count



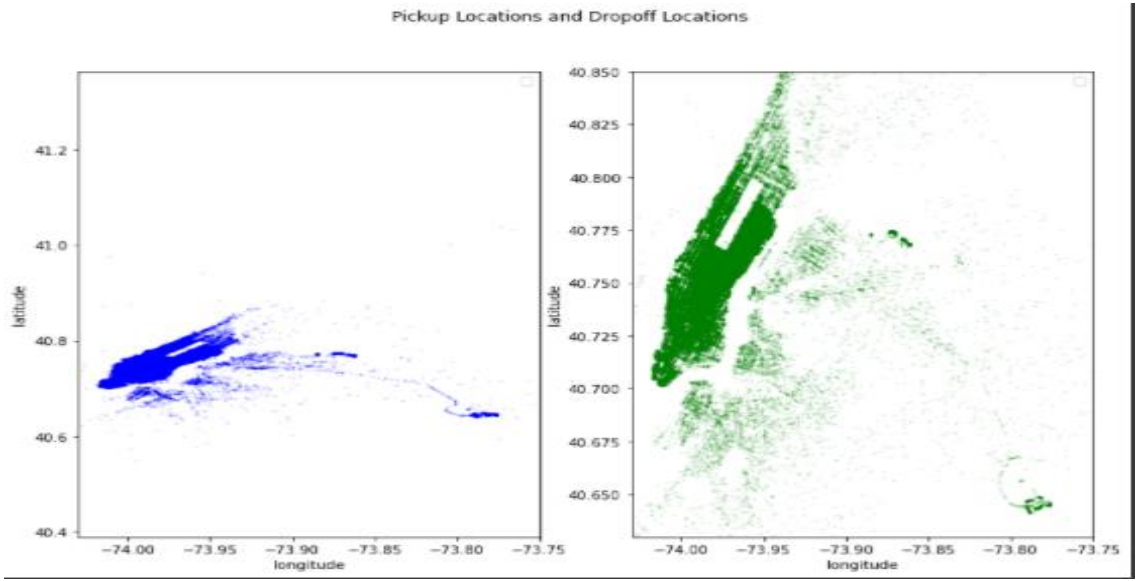


Pickup and Dropoff Locations

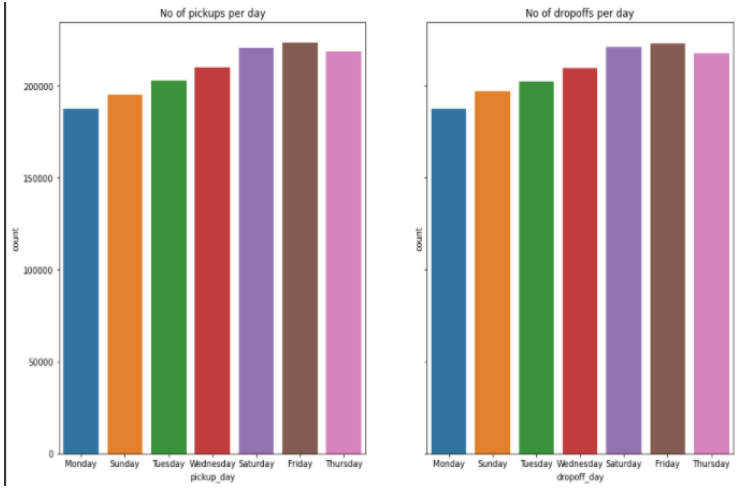


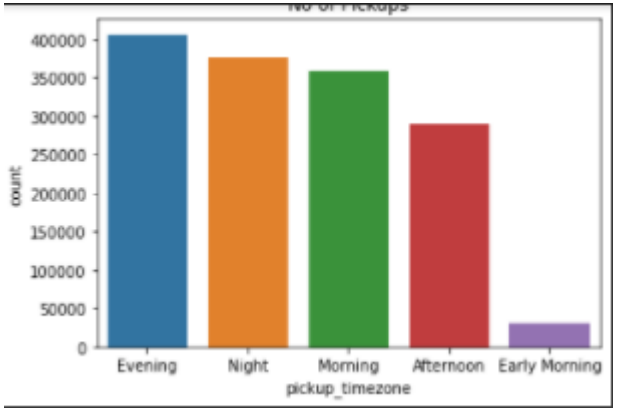
- Trip Duration

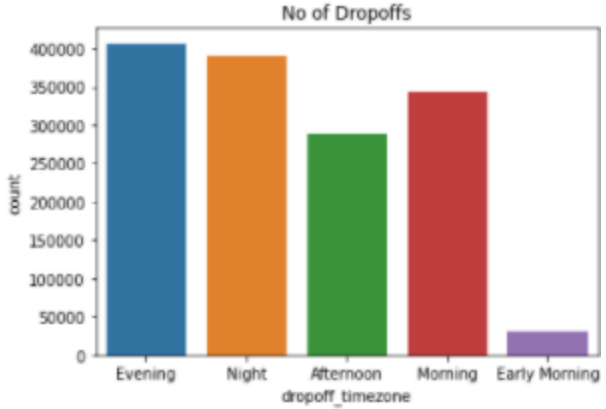
- Here we can see outliers of the each feature, this affects the accuracy of our models.



Here pickup and dropoff locations are plotted using scatter plot, as this is huge data it is difficult to plot on a map.

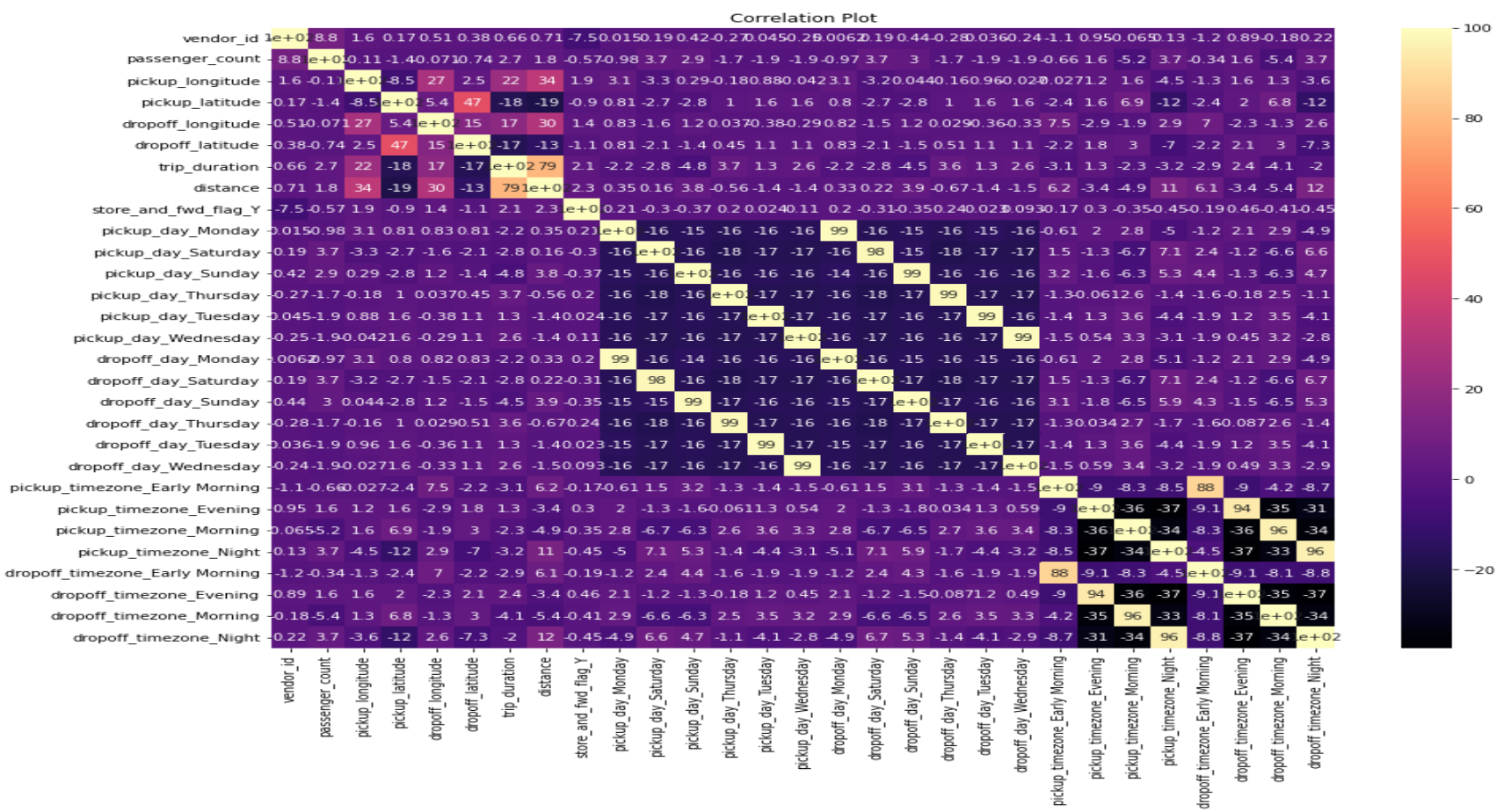






Here we can see there are more pickups and drop-offs on Friday.

There are more pickups and drop-offs in the evening and less in the early mornings.



We have plot correlation of the features using heat map.

Data Preprocessing:

- Here we have added extra features derived from the given data, following are the new features added
  - Pickup date time and drop-off date time
  - Pickup and drop-off day
  - Pickup and drop-off timezone
  - Distance
  - Pickup and drop-off month

- For distance calculation we have used pickup and drop-off locations.
- We have used geopy library to calculate distances using latitude and longitude data.

HANDLING OUTLIERS:

- There we several outliers in the data which affects the accuracy of the model so we have removed outlier data.
- We have removed outliers by using IQR method.

The trip_duration and distance feature were skewed so we have normalized using np.log10 function.

ONE HOT ENCODING:

There we few features which were categorical and not numerical, so we have used one hot encoding technique on those features and changed to numerical values.

Model Creation:

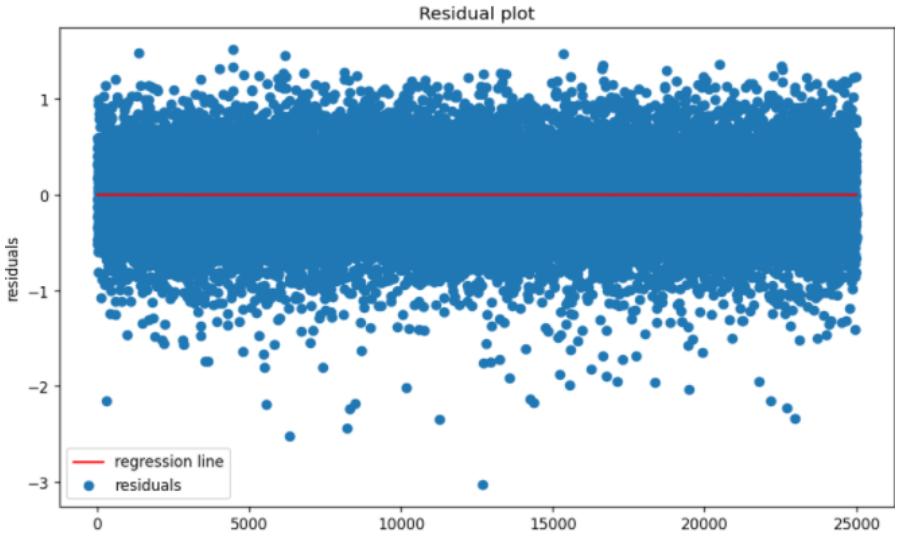
As the target feature outcome is continuous, so this is a regression problem. We have created some regression models. The following are the models we have created

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- Decision Tree Regressor with GridSearch CV
- Xgboost
- Gradient Boost
- AdaBoost

Linear Regression:

The following are the outcomes of the linear regression model

- R2 Value: 67.25%
- Adjusted R2 : 67.22%
- Root Mean Squared Error: 0.161

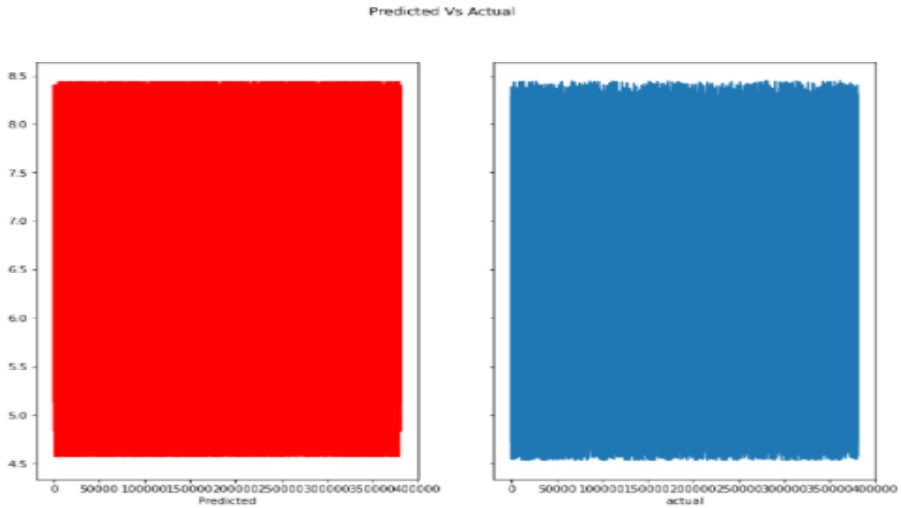


Here is the plot of the model, where the red line is the best fit line and points are the residual.

Decision Tree Regressor:

The following are the observations of the Decision Tree Regressor Model

- Accuracy : 69.9%
- Root Mean Squared Error: 0.3839
- R2 Value: 0.7006
- Adjusted R2: 0.7030

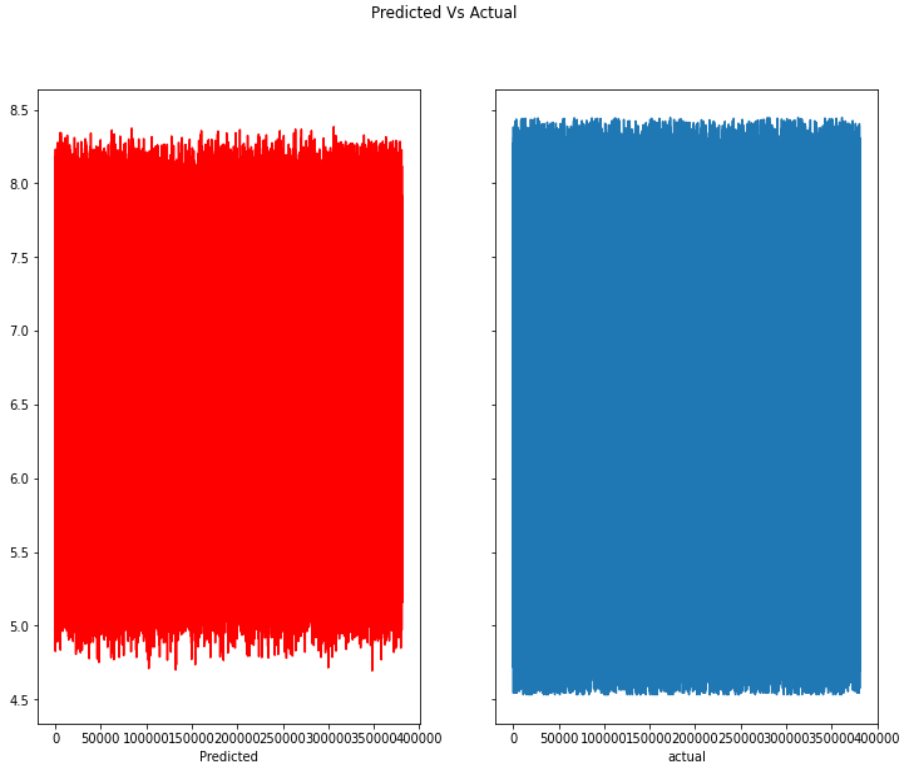


The above plot shows the predicted and actual values of Decision Tree Regressor Model

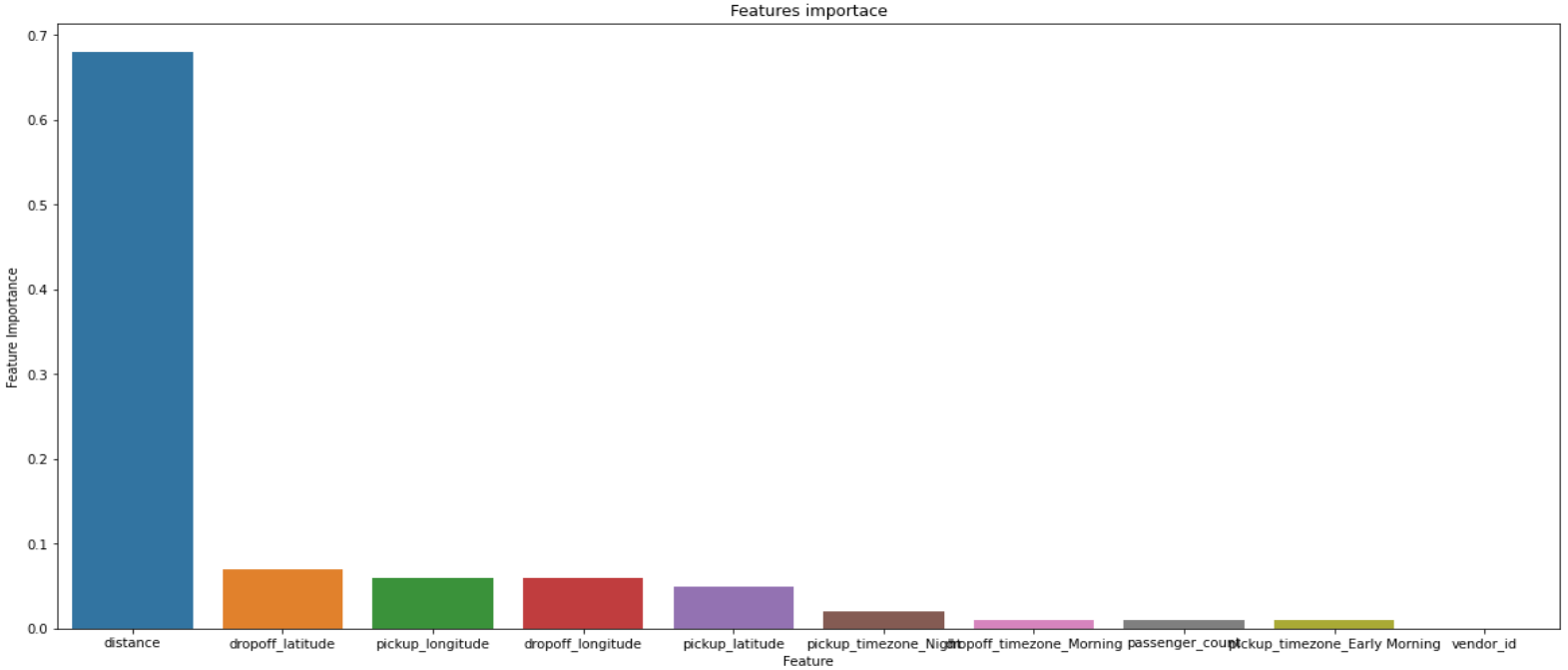
Random Forest Regressor:

The following are the outcomes of the Random Forest Regressor

- Root Mean Squared Error: 0.3494
- R2 Value : 0.75
- Adjusted R2: 0.89



The following plot shows the predicted and actual values of Random Forest Regressor Model

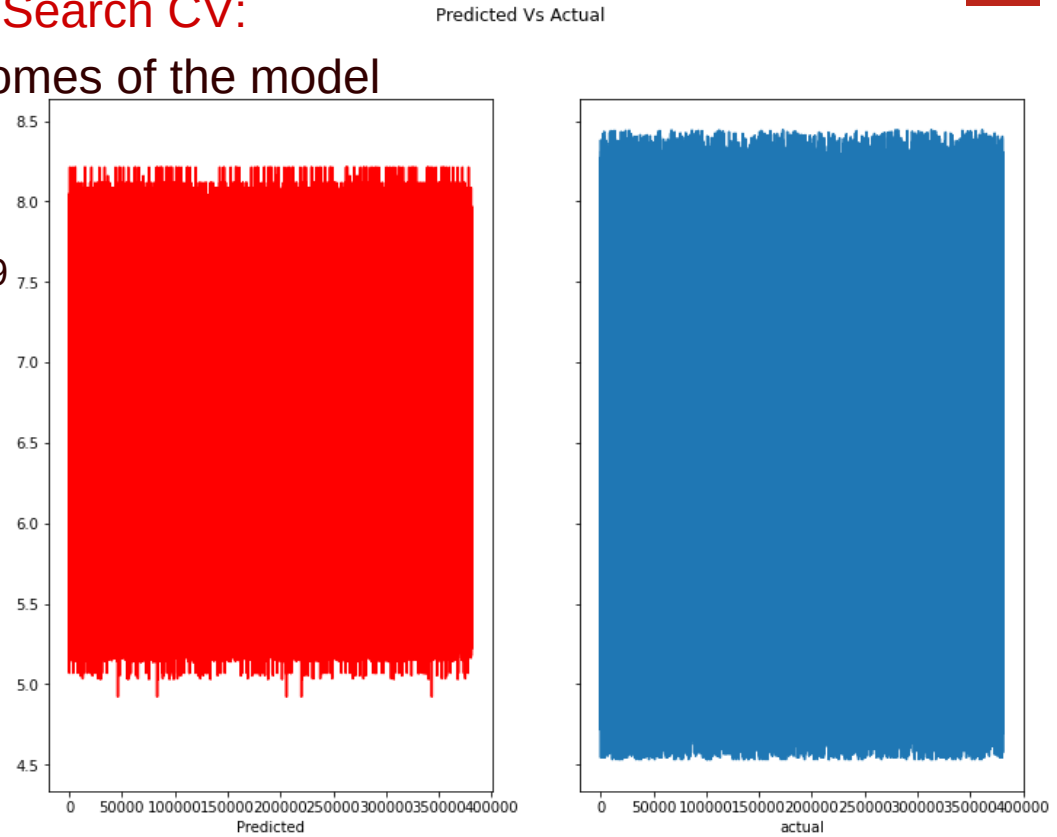


This plot gives the details of the important features that were helpful for the better accuracy and out of all the features distance feature has more importance.

Decsion Tree Regressor With GridSearch CV:

The following are the outcomes of the model

- Accuracy: 0.699
- Root Mean Squared Error : 0.3839
- R2 value: 0.7006
- Adjusted R2: 7030



The plot shows predicted and actual values of the model.

Boosting Models:

AdaBoost model gave the following outcomes

- Accuracy : 0.6
- Root Mean Squared error : 0.4429
- R2 Value : 0.601
- Adjusted R2: 0.75

Gradient Boosting gave the following outcomes

- Accuracy : 0.66
- Root Mean Squared Error : 0.405
- Adjusted R2 : 0.75
- R2 Value : 0.66

XgBoost with GridSearch model gave following outcomes

- Accuracy: 0.77
- Mean Squared Error : 0.111
- Root Mean Squared Error : 0.334
- R2 value : 0.77
- Adjusted R2 : 0.77

CONCLUSION:

The taxi trip data has been successfully visualized and analyzed on top of that we also added few machine learning models on the dataset after carefully fine tuning it. The models we have applied are Linear Regression, Decision trees, Random forests and their respective boosting methods such as Gradient Boosting, ADA Boosting and XG boosting. The Results of our models are as follows:

- Linear Regression : [score: 0.6725, R2_Score : 0.6789, Adj_r2: 0.6788]
- Decision Trees : [score : 0.6994, RMSE : 0.3839, R2 : 0.7006, Adj_r2 : 0.7030]
- Random Forests : [score : 0.942, RMSE : 0.156, R2 : 0.6828, ADJ_R2 : 0.6830]

Boosting Methods:

- ADA Boost : [score: 0.526, RMSE : 0.189, R2 : 0.534, ADJ_R2 : 0.682]
- Gradient Boosting : [Score: 0.579, RMSE : 0.18, R2 : 0.598, ADJ_R2 : 0.682]
- XG Boosting : [Score : 0.706, RMSE : 0.15 , R2 : 708 , ADJ_R2 : 0.737]

After Closely Watching the models and their boosting methods we can clearly see that Random Forest gave the best scores and the results are on-par when we tested the Test values with the actual data set.

Thank You