Capstone Project-2
Supervised ML - Regression
NYC Taxi Trip Time Prediction

Team Members-

Kaustubh M.Amare
Jahnavi Jaolekar
Md. Nawab Ali

OVERVIEW OF THE PROJECT

- Introduction.
- Problem Statement.
- Importing and taking an overview of the data.
- Preprocessing and feature engineering.
- Exploratory data analysis.
- Multicollinearity and correlation check.
- Processing outliers and checking distribution.
- Supervised Machine Learning of NYC Taxi Trip Time.
- Comparing evaluation metrics of different models.
- Conclusion.

INTRODUCTION

- There are many methods to travel within a city, but no other mode of transportation in cities has as many applications as the taxi ride. Analyzing and predicting journey duration between two sites in the city becomes crucial when the essential set of parameters that affect trip duration are available. In order to offer a quality taxi service and connect it with the existing transportation system, the project acts as an adequate means to comprehend New York City's traffic system. The factors pick-up latitude, pick-up longitude, drop-off latitude, drop-off longitude, etc. are taken into consideration when making predictions about required travel time.

PROBLEM STATEMENT

- The task is to build a machine learning model that predicts the duration of an NYC taxi trip using the dataset which includes pickup time, geo-coordinates, the number of passengers, and several other variables.

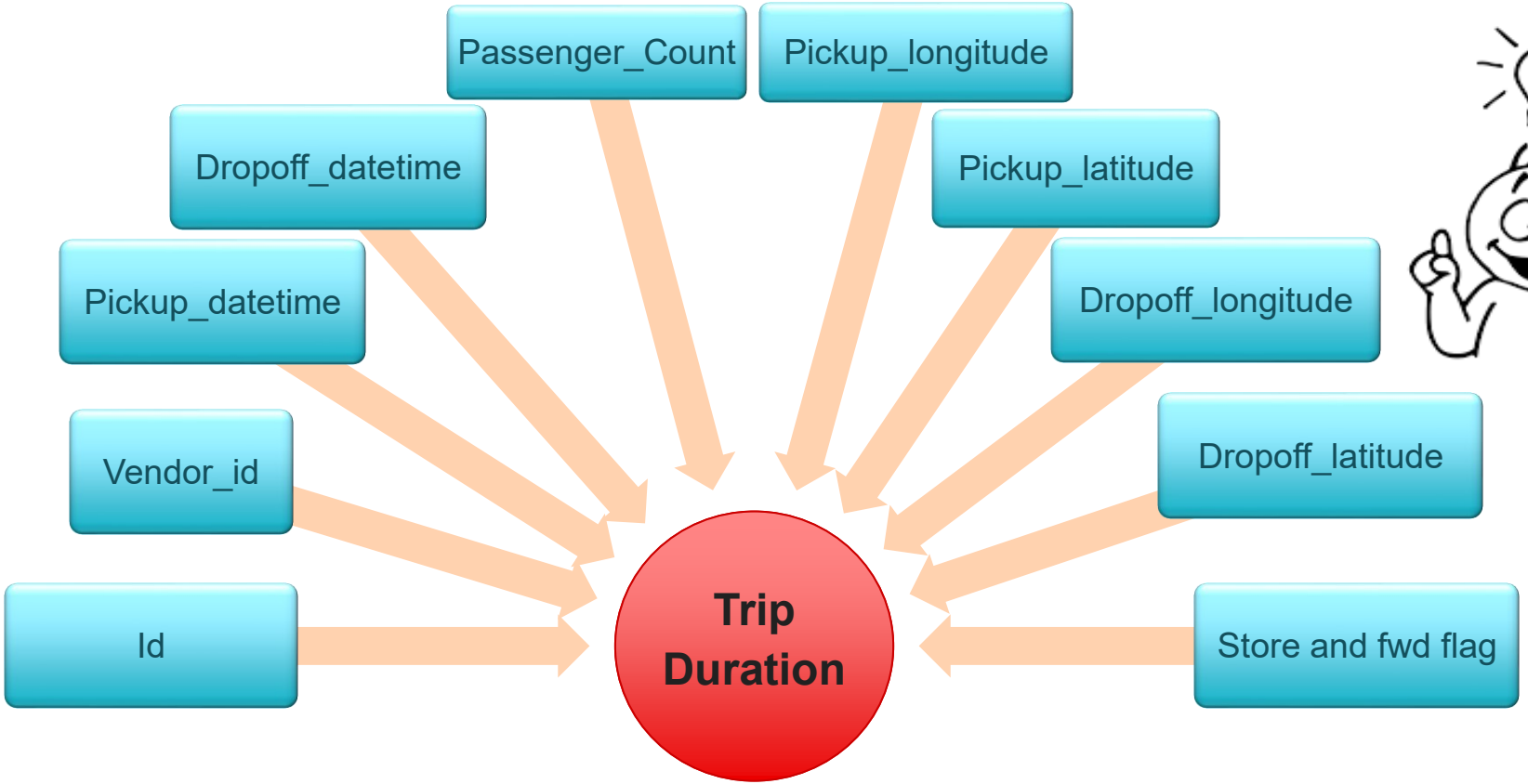
Importing and taking an overview of the data.

Dataset file format	CSV (Comma Separated) file is used
Name of the Data-source file	NYC Taxi Data.csv
Number of columns	11
Number of rows	1458644
Number of Numerical columns are	7
Number of Categorical columns are	2

We can see that we do not have any null values in the dataset

	Null Count	Dtype	unique_count
id	0	object	1458644
vendor_id	0	int64	2
pickup_datetime	0	object	1380222
dropoff_datetime	0	object	1380377
passenger_count	0	int64	10
pickup_longitude	0	float64	23047
pickup_latitude	0	float64	45245
dropoff_longitude	0	float64	33821
dropoff_latitude	0	float64	62519
store_and_fwd_flag	0	object	2
trip_duration	0	int64	7417

Overview of Data



Preprocessing and feature Engineering

Preprocessing

- In the given dataset pickup_datetime and dropoff_datetime are not in proper date time format and we need to convert them into datetime format.

```
# Converting timestamp to datetime format
taxi_df['pickup_datetime']=pd.to_datetime(taxi_df['pickup_datetime'])
taxi_df['dropoff_datetime']=pd.to_datetime(taxi_df['dropoff_datetime'])
taxi_df.head()
```

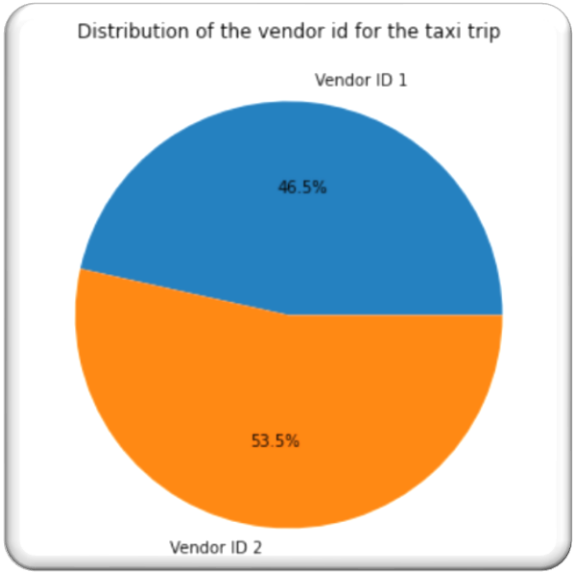
Feature Engineering

Geodesic Distance-
It is the length of the shortest path between 2 points on any surface. In our case, the surface is the earth. Below program illustrates how to calculate geodesic distance from latitude-longitude data.

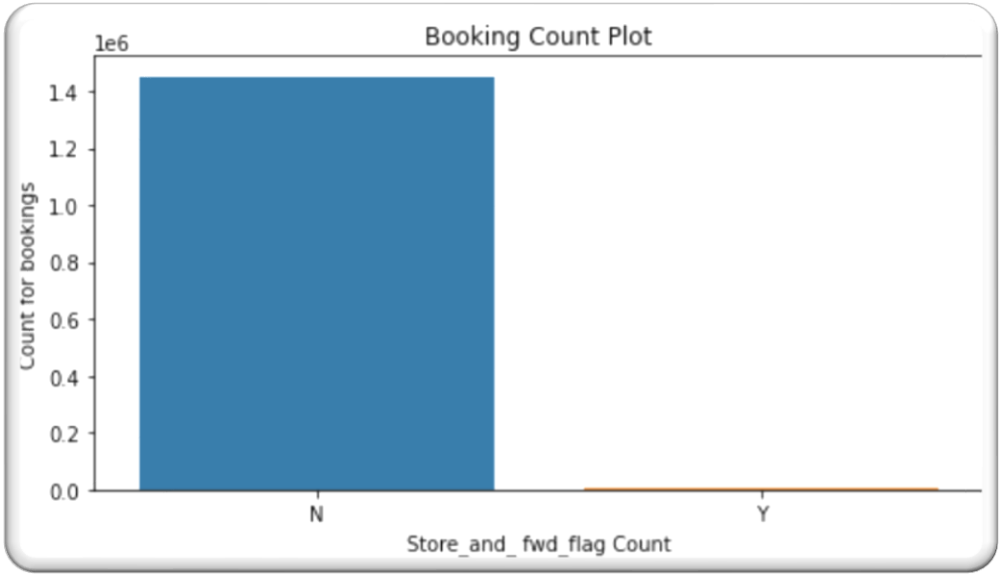
```
# Creating the function for calculating distance between pickup and dropoff
from geopy.distance import geodesic
def distance_cal(pickup_lat,pickup_long,dropoff_lat,dropoff_long):
  start_coordinates=(pickup_lat,pickup_long)
  stop_coordinates=(dropoff_lat,dropoff_long)
  return geodesic(start_coordinates,stop_coordinates).km


# Applying the function to our dataset and creating the feature 'distance'.
taxi_df['distance'] = taxi_df.apply(lambda x: distance_cal(x['pickup_latitude'],
  x['pickup_longitude'],x['dropoff_latitude'],x['dropoff_longitude'] ), axis=1)
```

EXPLORATORY DATA ANALYSIS

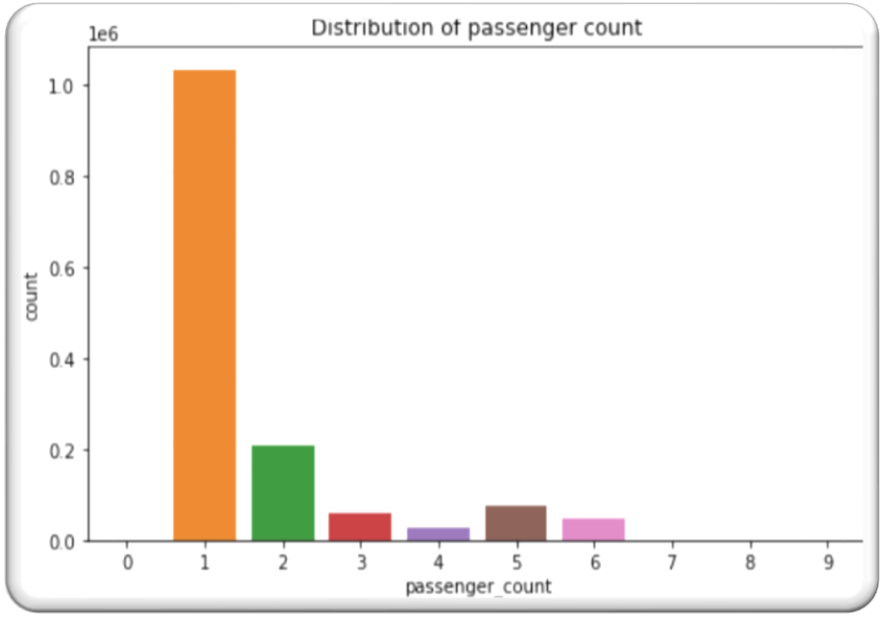


We can see that for vendor 2 there are more number of bookings which is of 54 %

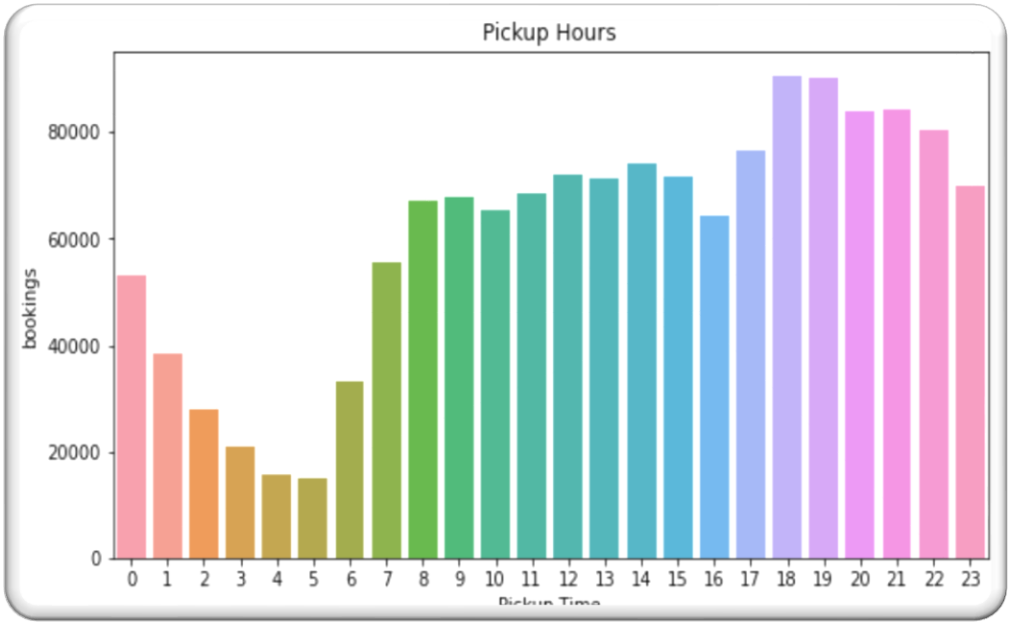


This demonstrates that when the store and fwd flag is marked, there is a very low count of 'yes' and the majority of the time the taxi driver hasn't logged onto the vendor's systems.

EXPLORATORY DATA ANALYSIS

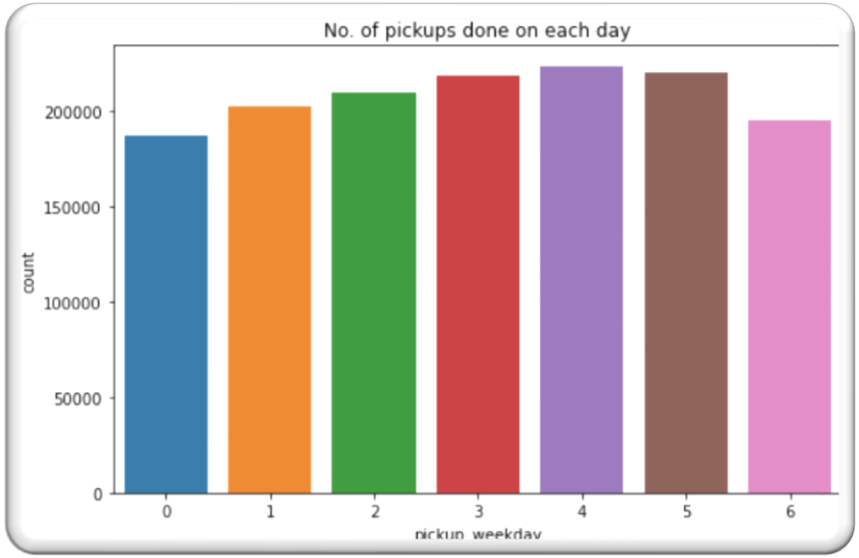


We can notice that most of the bookings are made by solo travelers, which means less number of people prefer car pool or may be less number of groups book car...people prefer to ride solo

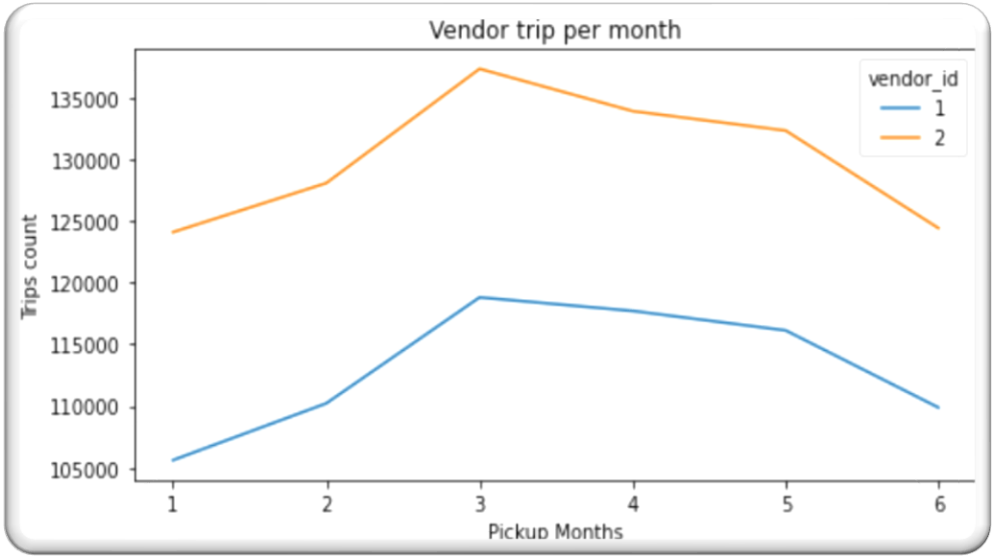


Distribution of pickup hours shows that most of the pickups are in the evening. We can see that people often use taxi services to get to their workplaces in the mornings after 10:00. Additionally, the demand for taxis tends to surge in the late evening after six o'clock.

EXPLORATORY DATA ANALYSIS

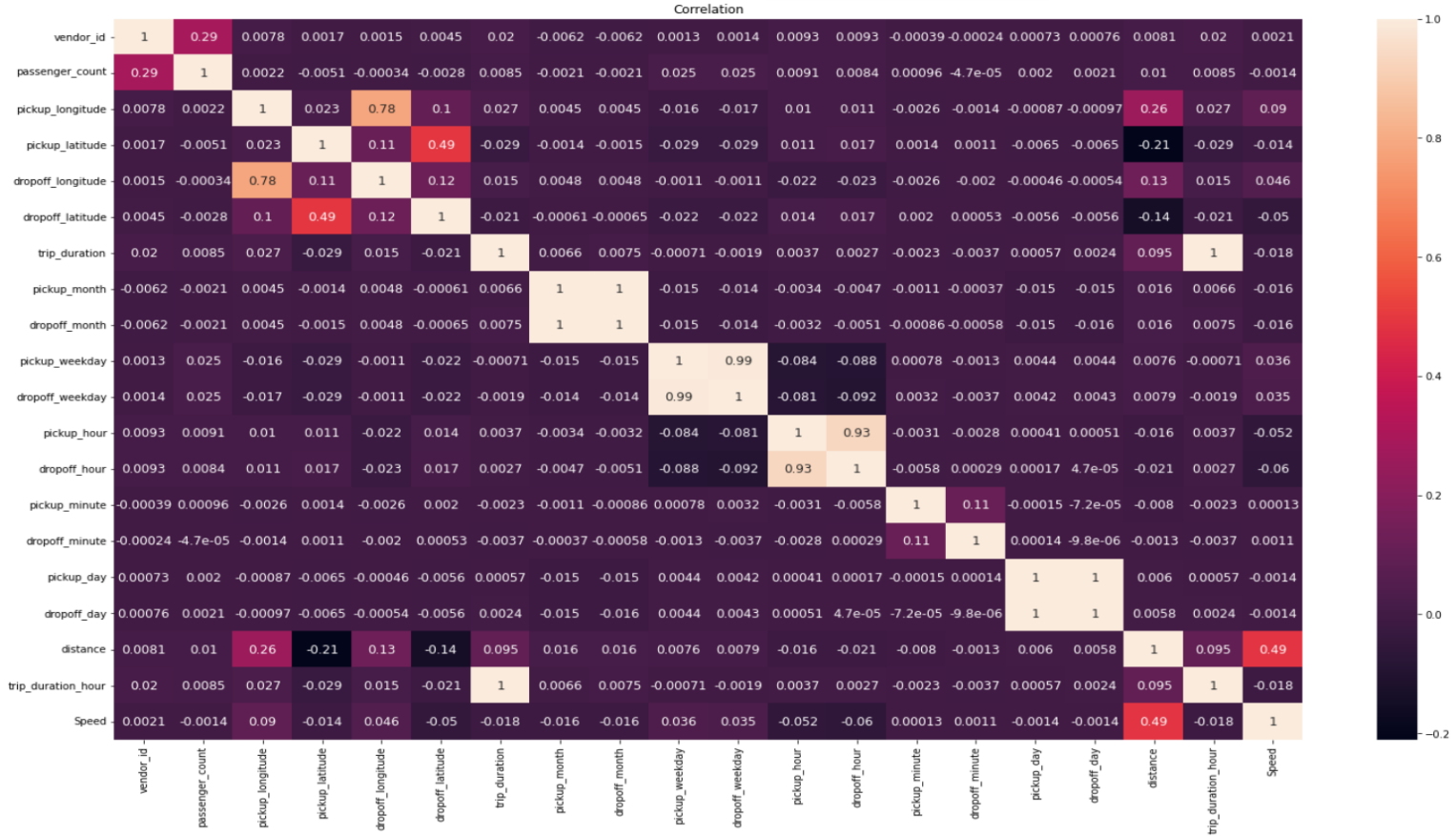


We can see that compared to other days, taxi booking rates are higher on the weekends (4-Friday and 5-Saturday). This suggests that individuals used to go out on weekends for their celebrations, parties, or even other personal work.



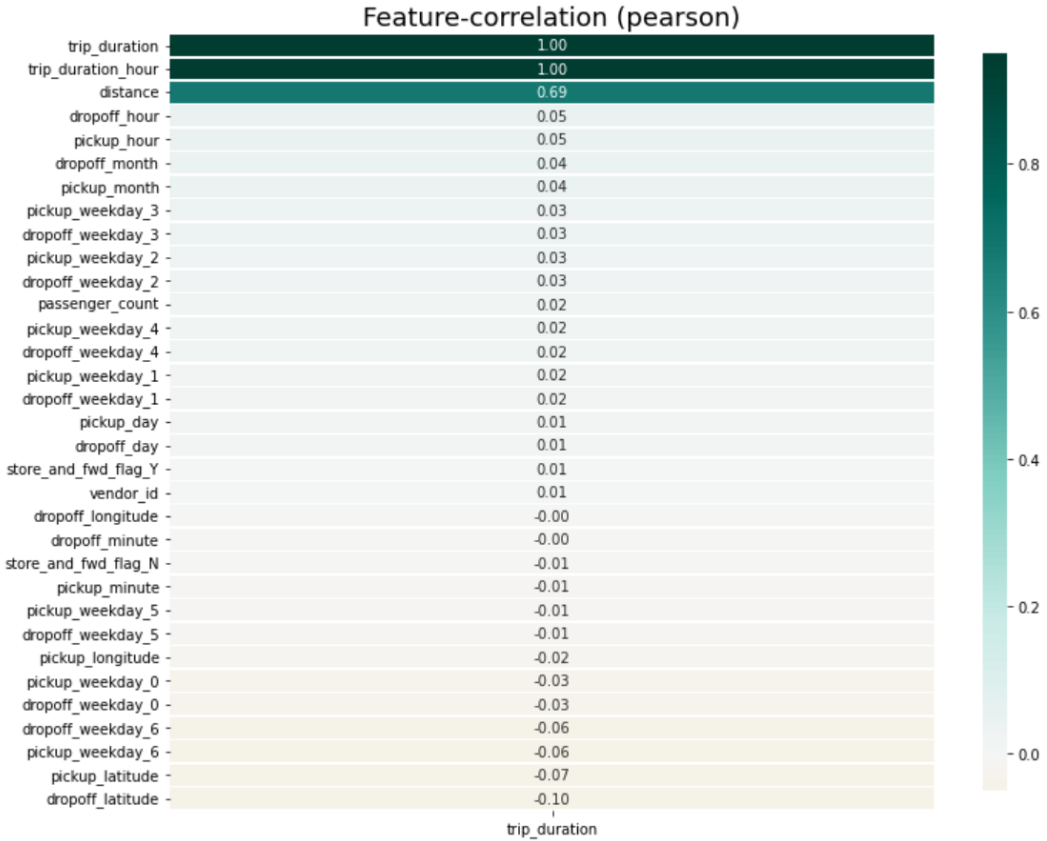
We can see that both vendors' trips are at their maximum in the month of March and their lowest in the month of January, February, and after June.

Multicollinearity and correlation check.



Above heatmap clearly shows that 'dropoff_day', 'dropoff_hour', 'dropoff_month', 'dropoff_weekday' are highly correlated.

Multicollinearity and correlation check.



As we can observe from the plot, ‘distance’ column highly affects our dependent variable ‘trip duration’.

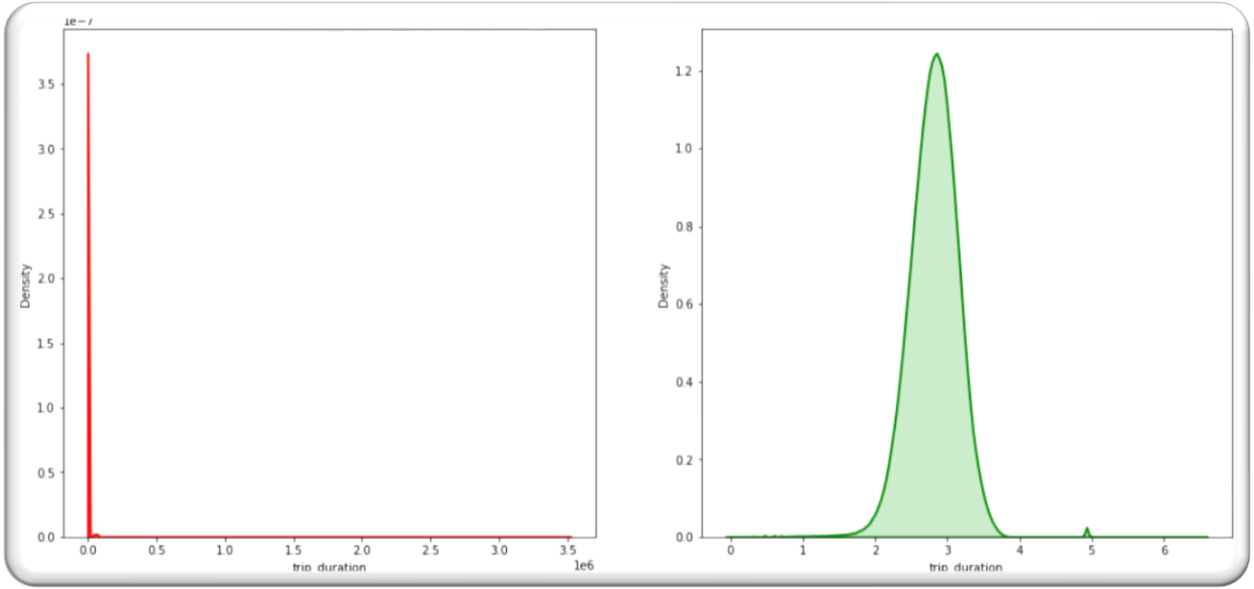
Multicollinearity and correlation check.

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables.

We know that the acceptable range of VIF is <5, pickup day and dropoff day have values around 8 so we need to remove one of them.

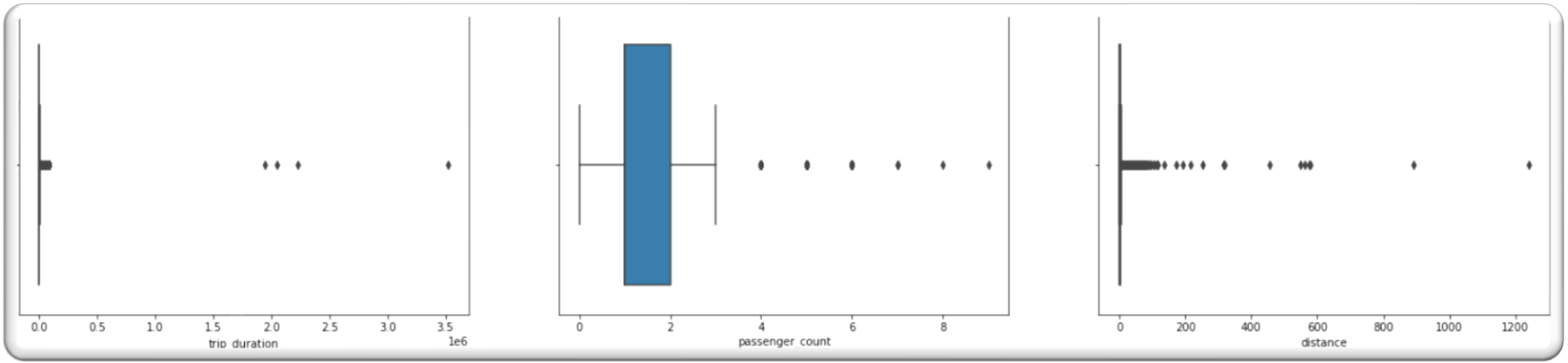
	variables	VIF
0	vendor_id	1.141995e+01
1	passenger_count	2.841012e+00
2	pickup_longitude	2.933240e+06
3	pickup_latitude	1.682321e+06
4	dropoff_longitude	2.736948e+06
5	dropoff_latitude	1.477376e+06
6	pickup_month	3.572189e+06
7	dropoff_month	3.572873e+06
8	pickup_weekday	2.868816e+02
9	dropoff_weekday	2.871991e+02
10	pickup_hour	1.455222e+02
11	dropoff_hour	1.448103e+02
12	pickup_minute	4.115507e+00
13	dropoff_minute	4.097740e+00
14	pickup_day	8.133573e+04
15	dropoff_day	8.133847e+04
16	distance	1.725101e+00

Processing outliers and checking distribution



We checked skewness of our dependent variable, and found that, the graph is positively skewed so it needs to be regularized. We used log transformation to convert this into a normal distribution.

Processing outliers and checking distribution



Trip duration, distance, passenger count had a lot of outliers, as for removal of outliers we used quartile method to remove outliers where we calculated 3 quarters, IQR based on that we calculated upper and lower limits and removed outliers that were beyond these limits.

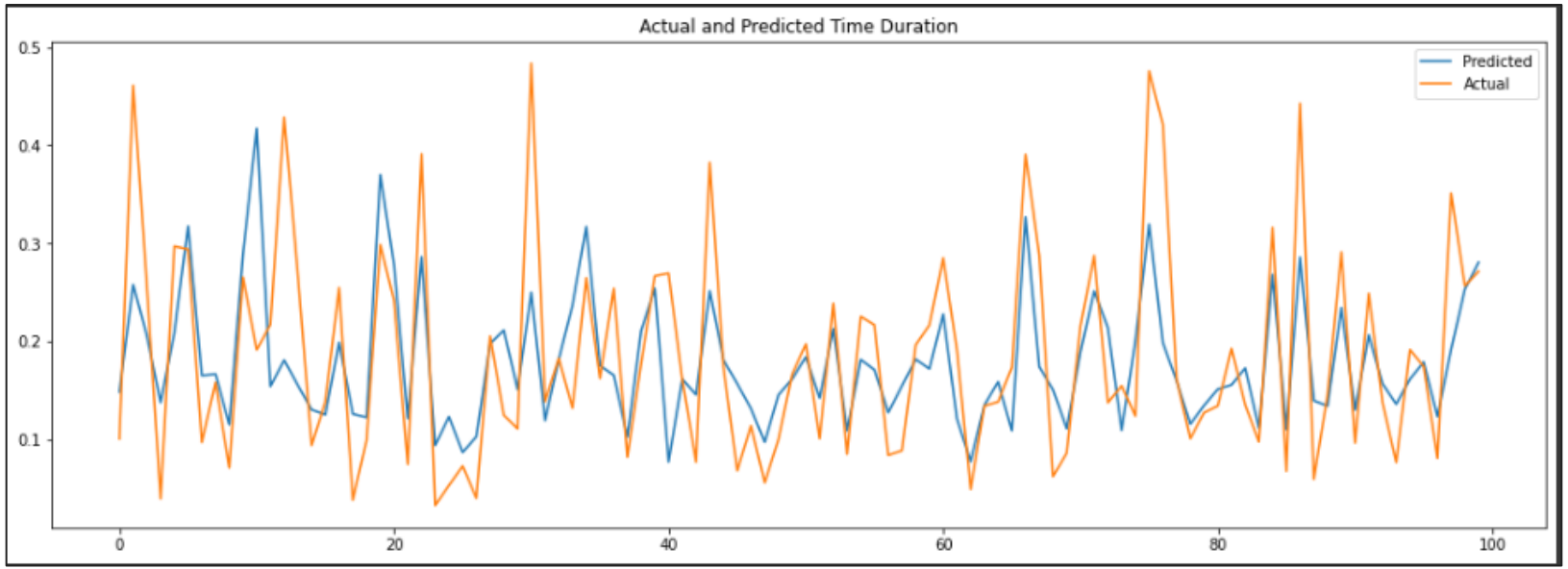
SUPERVISED MACHINE LEARNING MODELS

- We started this process by converting our textual information containing columns into dummy variables.
- Columns containing categorical data : Store_and_fwd_flag and week_days for pickup and dropoff.
- As we did not drop the columns which were collinear, we created a separate list of features/variables which were not correlated, so as to remove collinear data(which would have given us pseudo good results).
- Lastly we split the data into train and test sets for further calculation of evaluation metrics for various algorithms.

Model 1 – Linear Regression(Train set)

MSE - 0.0056 RMSE - 0.074833 R2 - 0.492813 AdR2 - 0.492712

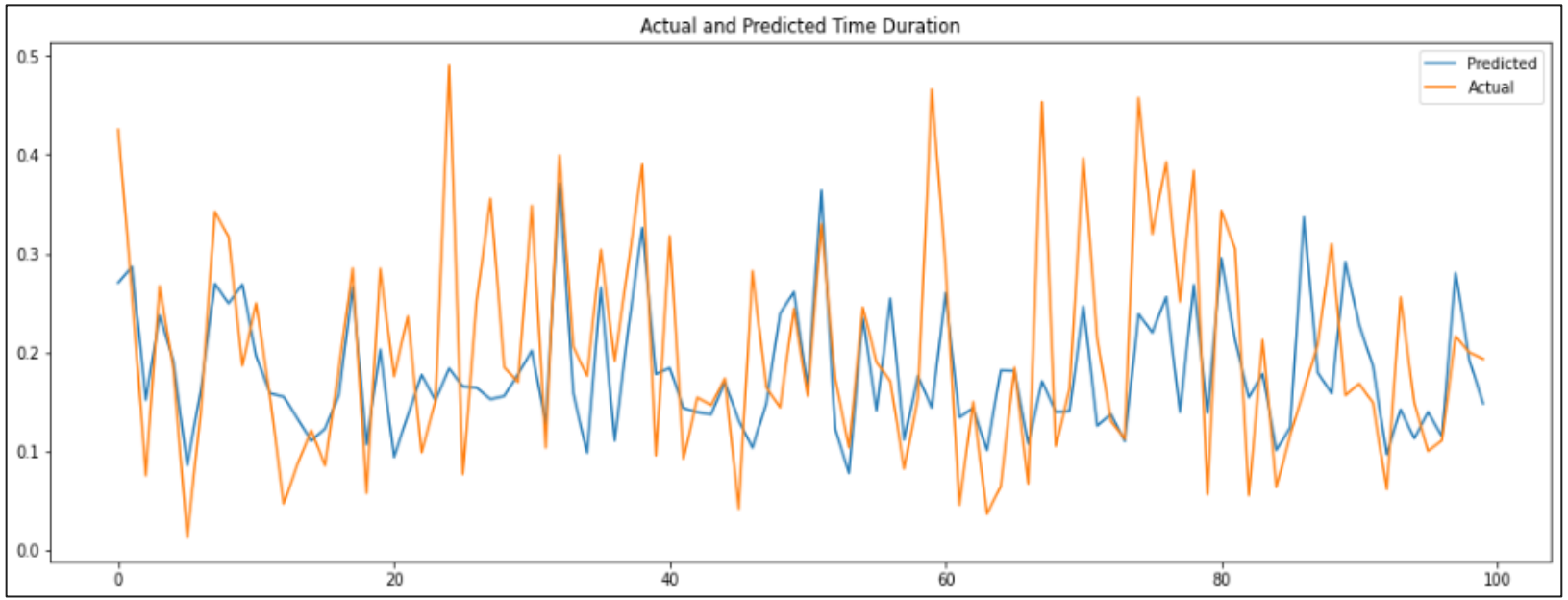
Since this algorithm has a very low R2 and a very high MSE, so this algorithm is not suitable for training our model. These are training set's results



Model 1 – Linear Regression(Test set)

MSE - 0.0055 RMSE - 0.074161 R2 - 0.494769 AdR2 - 0.494364

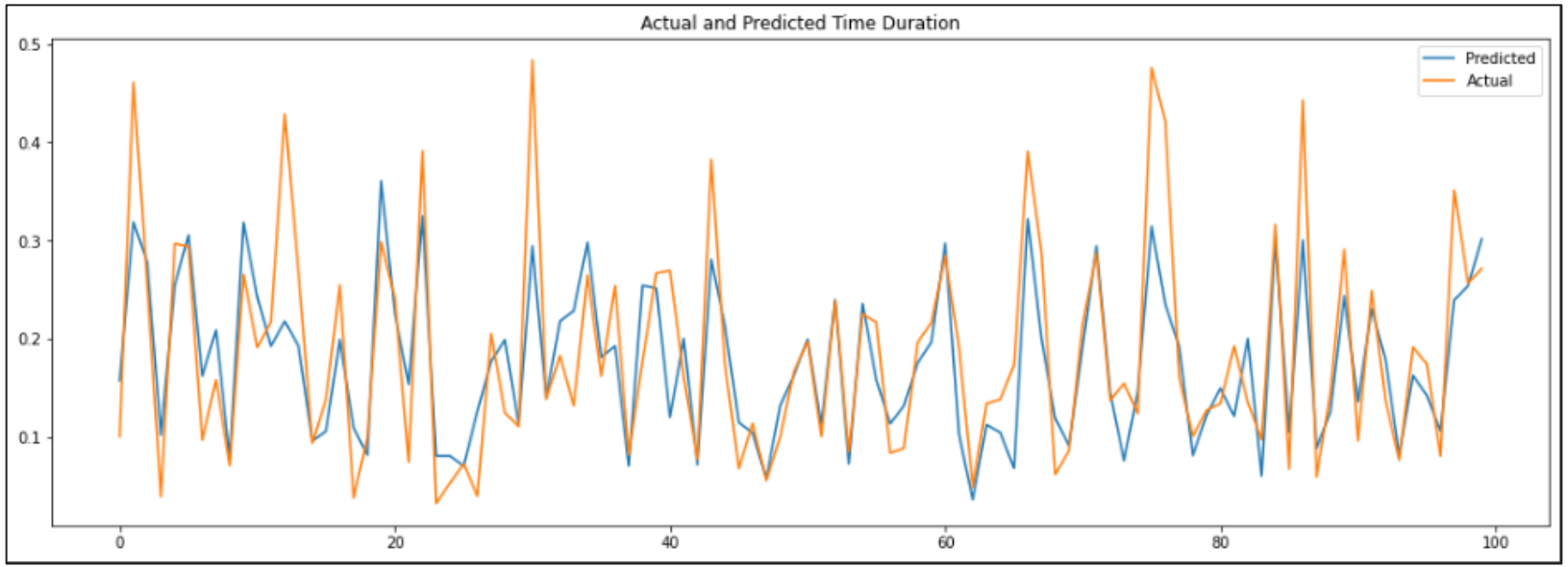
Test set results are also similar to the train set, therefore this algorithm is not suitable.



Model 2 – Decision Tree(Train set)

MSE - 0.0046 RMSE - 0.067823 R2 - 0.583740 AdR2 - 0.583656

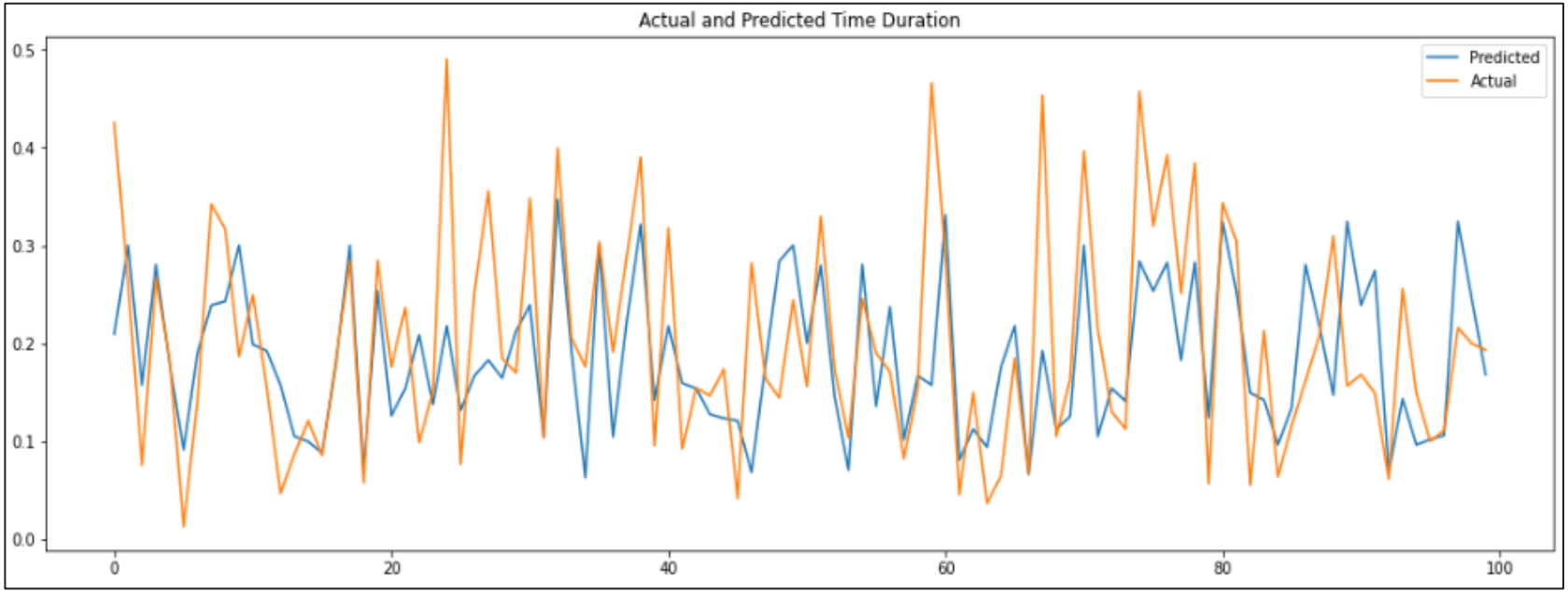
This algorithm is better than the previous one (Linear Regression) as it has an accuracy score of 58%, but still not good enough to train the model.



Model 2 – Decision Tree (Test set)

MSE - 0.0049 RMSE - 0.069999 R2 - 0.547142 AdR2 - 0.546779

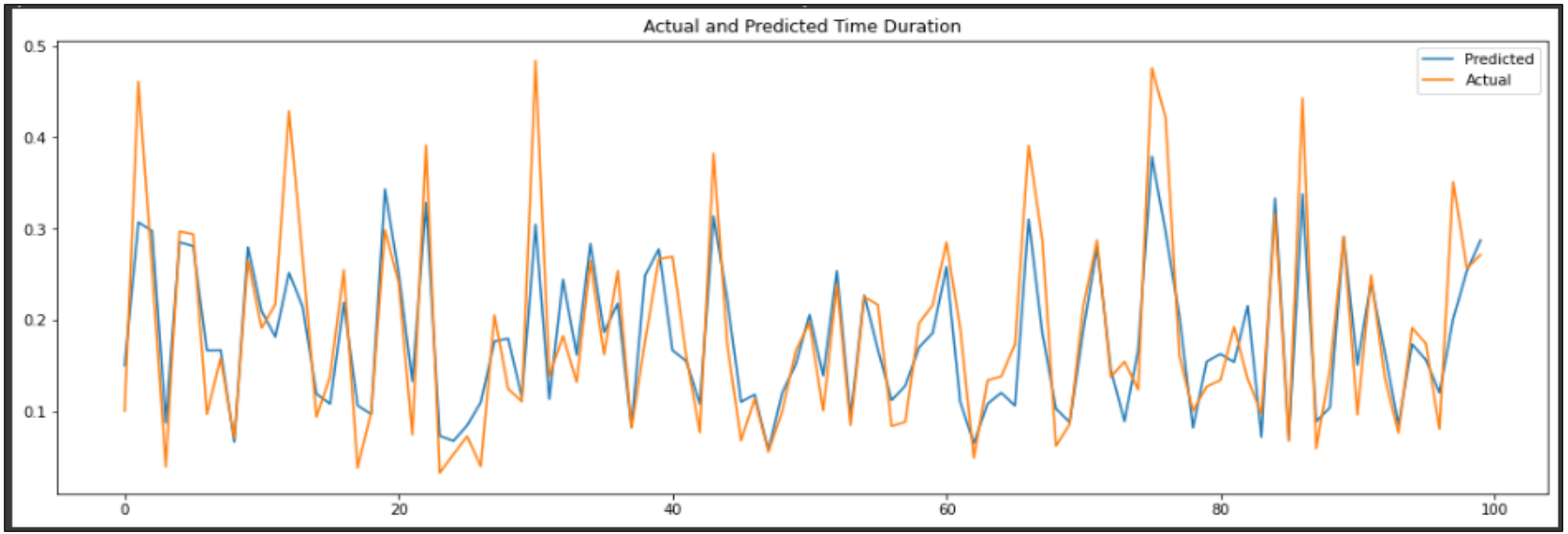
This algorithm also gives similar results for train and test



Model 3 – XG Boost (Train set)

MSE - 0.0031 RMSE - 0.055677 R2 - 0.719171 AdR2 - 0.719115

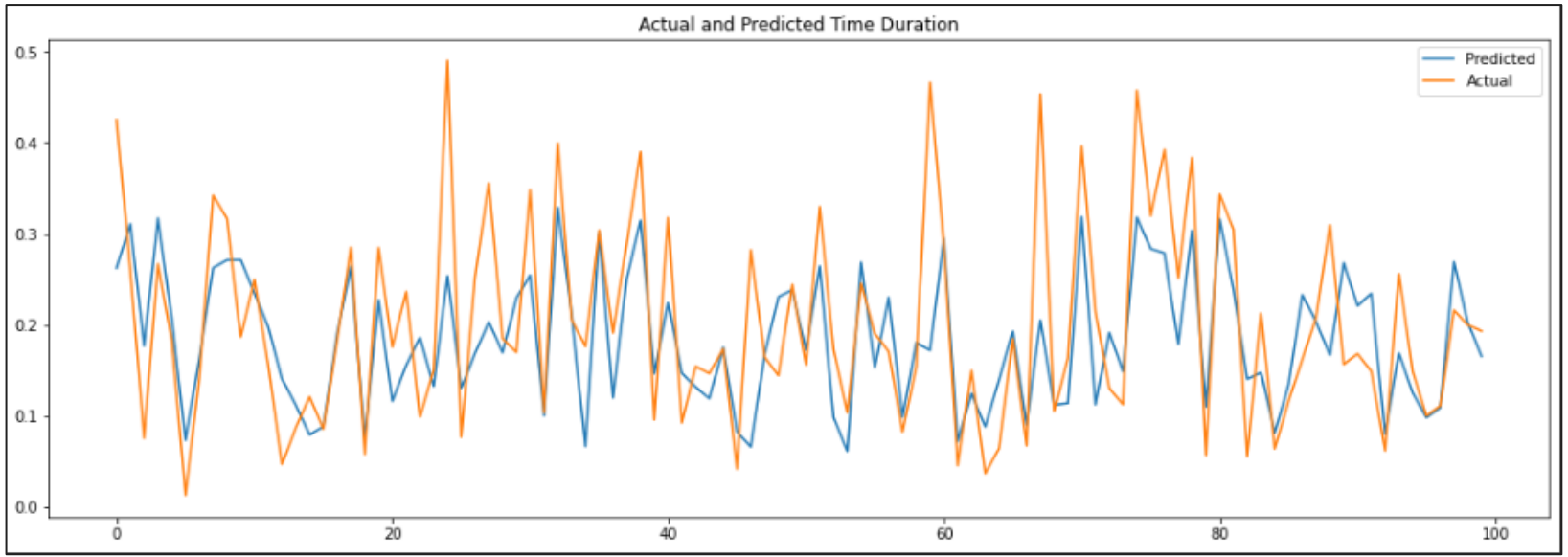
As we can see from the graph that actual and predicted values are near to each other(lines coinciding), this algorithm has performed well as compared to other algorithms and has an accuracy score of 71%.



Model 3 – XG Boost (Test set)

MSE - 0.0043 RMSE - 0.065574 R2 - 0.605751 AdR2 - 0.605436

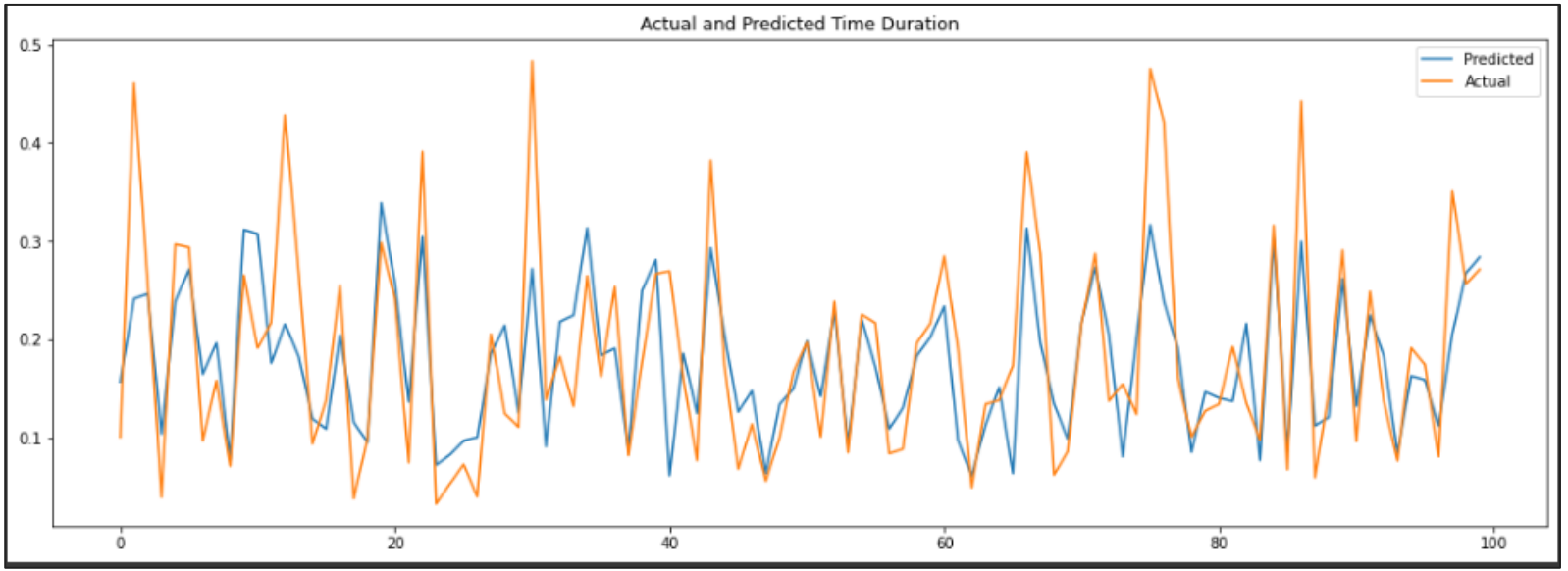
Test set also gave better results than the other algorithms but a little less than the train set. Accuracy score for test set was found to be 60%. There is are coincinding points as compared to train set.



Model 4 – Gradient Boost (Train set)

MSE - 0.0047 RMSE - 0.068556 R2 - 0.571997 AdR2 - 0.571912

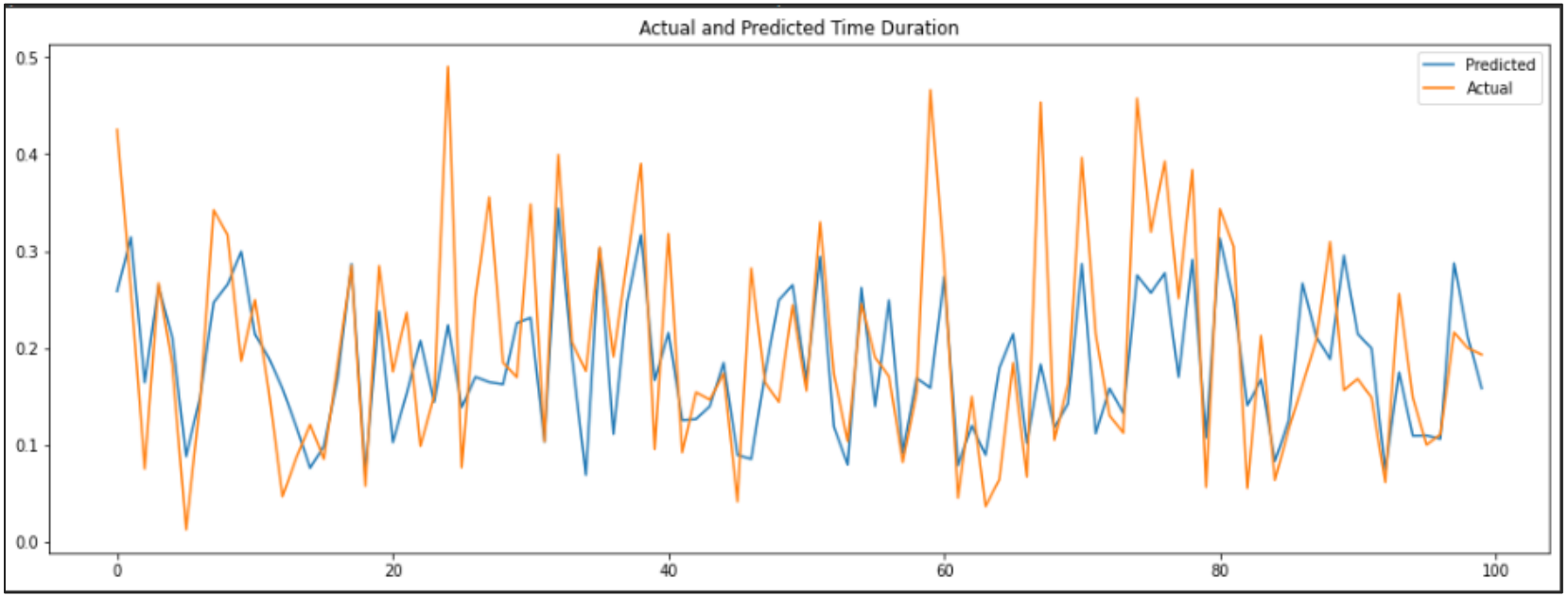
This algorithm is better than 1st one but has an accuracy score (57%) less than XG Boost (till now best score).



Model 4 – Gradient Boost (Test set)

MSE - 0.0047 RMSE - 0.068556 R2 - 0.563697 AdR2 - 0.563348

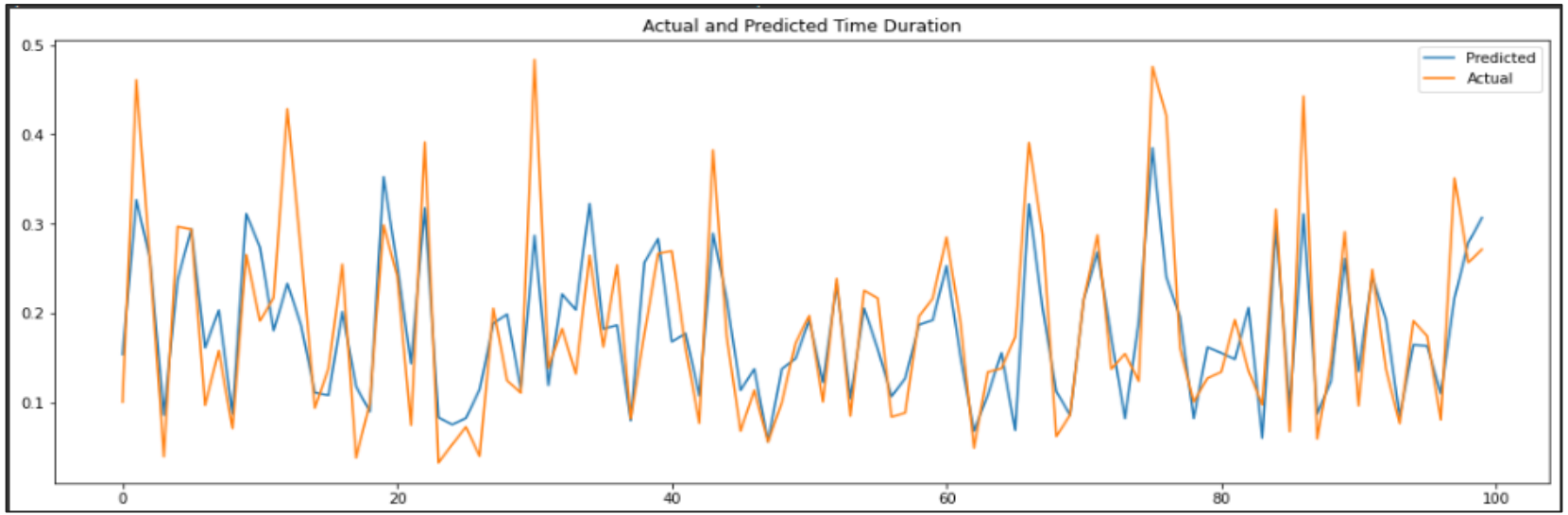
Test set also gave similar results as train set.



Model 5 – Random Forest (Train set)

MSE - 0.0043 RMSE - 0.065574 R2 - 0.612341 AdR2 - 0.612263

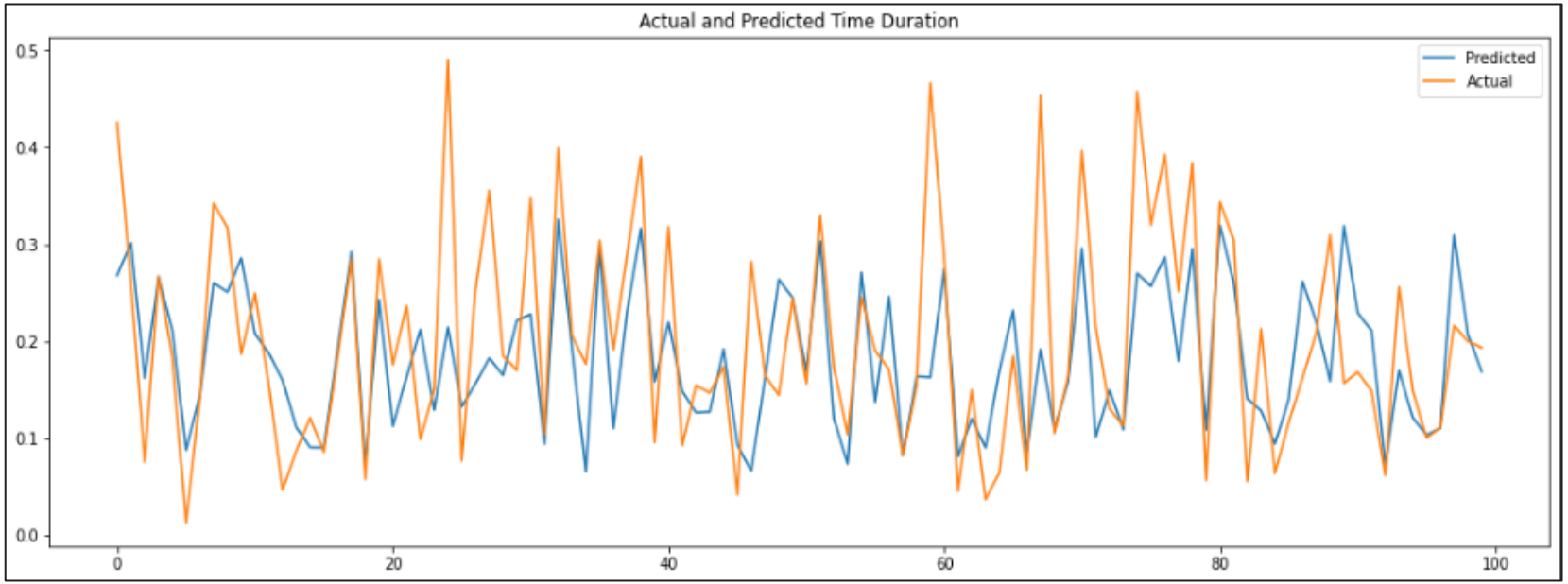
This algorithm has better accuracy score (61%) than our rest of the algorithms but still not better than XG Boost (still with highest score). We can see the lines merging but not as frequent as XG Boost



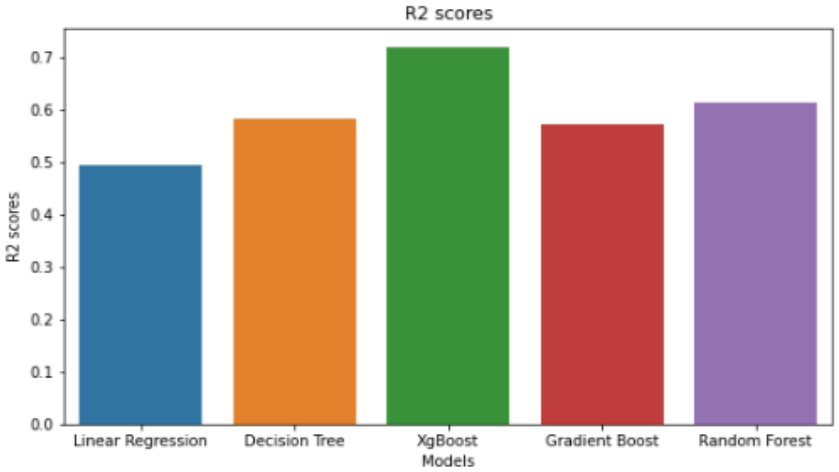
Model 5 – Random Forest(Test set)

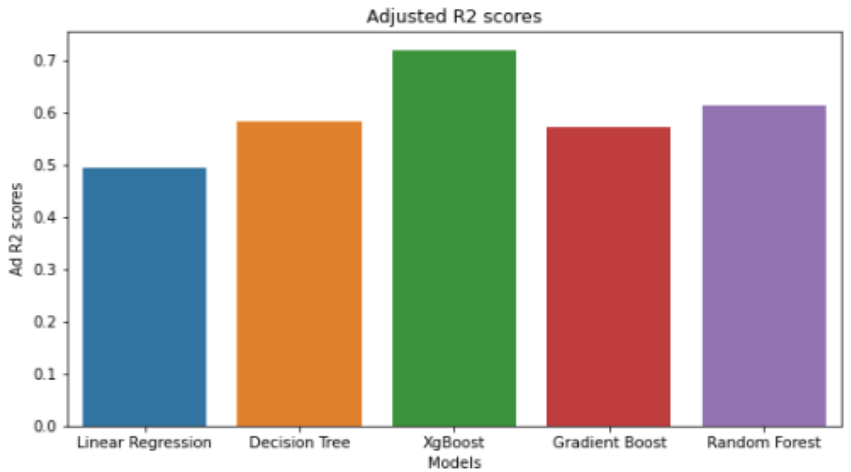
MSE - 0.0047 RMSE - 0.068556 R2 - 0.566339 AdR2 - 0.565992

Test gives nearby similar scores as that obtained from train set.



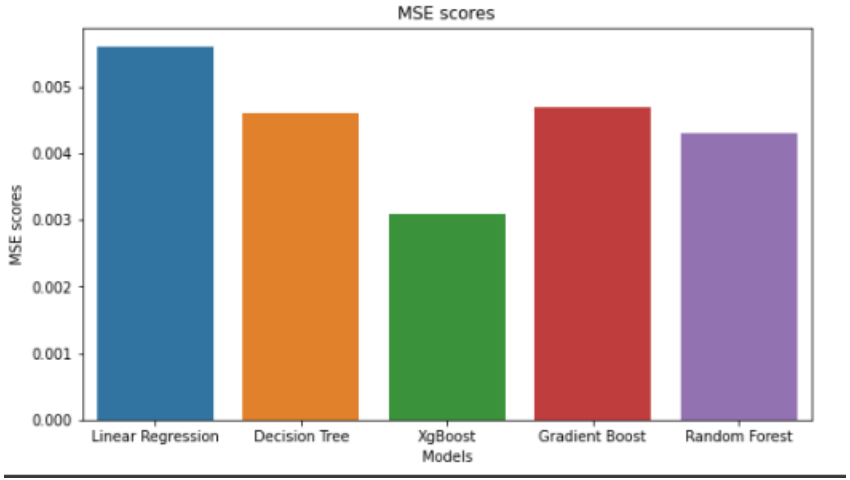
COMPARING EVALUATION METRICS OF DEIFFERENT MODELS

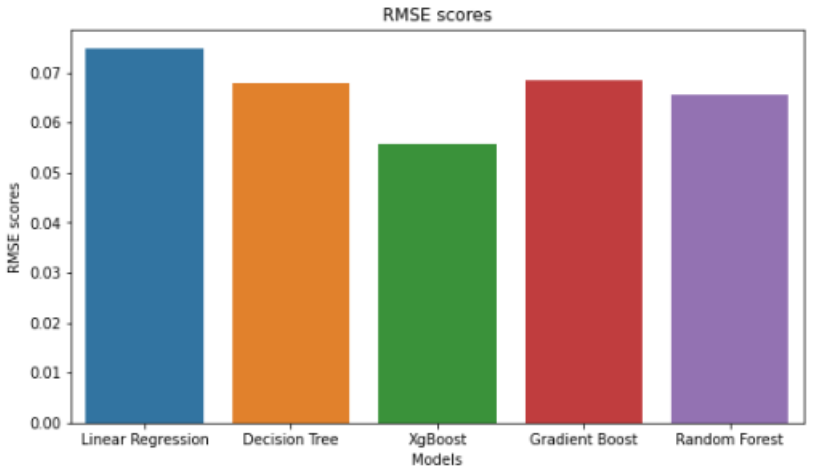




Graphs clearly show that XG Boost has better scores than other algorithms

COMPARING EVALUATION METRICS OF DEIFFERENT MODELS





Good model should have low MSE score, XG Boost is the only model with low MSE scores.

CONCLUSION (EDA)

- Vendor 2 has number of bookings.
- Majority of the time taxi drivers have not logged into the system.
- It was observed that taxi booking rates were higher on the weekends, which clearly shows that individuals use to go out for celebrations, parties, or may for some work.
- Taxi reservations are higher in the month of March and April.
- Taxi services are used at peak in the morning hours and in the late evening which depicts quite a picture that people use this service for going to work ang coming back from work.
- Most of the bookings are made by solo travelers, people don’t prefer car pool.

CONCLUSION (Model Training)

- As we know a good model has low MSE scores and high R2 scores.
- In our case we tried 5 algorithms: Linear Regression, Decision Tree, XG Boost, Gradient Boost, Random Forest.
- Out of these 5 XG Boost gave the best accuracy score of 71% (high R2 and low MSE)
- During the feature engineering part we tried to add a new column ‘speed’ but due its high correlation with ‘distance’ column it gave false predictions and gave an accuracy score of 99% in 3 of the algorithms.
- As we got similar results for both train and test sets of data for different algorithms we can say that model was not overfit.
- We can come to conclusion that XG Boost has the best accuracy score, and we tried taking an optimum parameter so that our model doesn’t overfit.

THANK YOU...