|  |
| --- |
|  |
| NYC Taxi Trip Time Prediction  Supervised Machine Learning: - Regression |
| |  |  |  | | --- | --- | --- | | Md Nawab Ali | 11/29/22 | Almabetter- ML | |

# NYC Taxi Trip Time Prediction

# Mr. Md Nawab Ali, Ms. Janhavi Jaolekar and Mr. Kaustubh Amare

# Data science trainees, Alma better, Bangalore

­­­­­­­­­­­­­­­­­­­­­­\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Abstract

We were provided with the dataset of NYC Taxi and limousine commission. Which included with the data set of 6 months with features like, ID, vendor id, pickup and drop coordinates, date, time required etc. For the better customer accessibility, we should be able to predict the distance required to complete the trip as this feature will help customer and help us to stay up in competition.

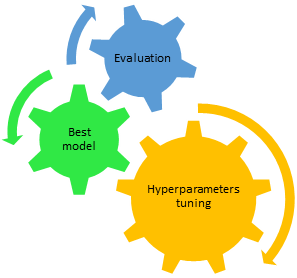
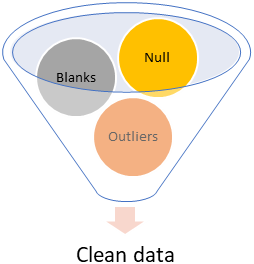
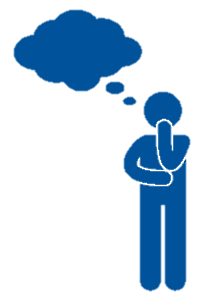
# Introduction

The New York City Taxi and Limousine Commission (NYC TLC) is a department of the city government of New York that issues license and manages the for-hire car and medallion taxi sectors, In New York City many of people commute to different regions of city via taxi. A lot of streets and roads in New York city are quite busy due to traffic jams, construction, or road blockage etc. Therefore, it is very important to predict the trip duration of taxi so that the user will know how much time it will take to commute from one place to other. Also, due to the increasing popularity of app-based taxi decisions has to be taken by the user for opting which one to choose based on trip pricing and duration.

Problem statement: -

We need to find a machine learning algorithm which can predict the time required to reach at the destination selected by user.





**Conclusion**

# Approach

1. Understand the DataFrame
   1. Shape and info in Dataframe.
   2. Problem statement
   3. Feature information.
   4. Plan of execution.
2. Data cleaning
   1. Process null values
   2. Blanks
3. Exploratory Data Analysis
   1. Extracting insights
   2. Studies trends from data
4. Multicollinearity studies
   1. Correlation between independent

variables.

* 1. Effects of independent variable

on dependent variable

* 1. Implementation of VIF.

1. Dataset preparation for modeling.
   1. Feature engineering.
      1. Adding some features
      2. Transforming some features.
   2. Scaling
   3. Train test split using sklearn.
2. Machine learning
   1. Training and testing data on multiple

algorithms

* 1. Finding best Model for given data.
  2. Evaluate

1. Conclusion

Distribution of vendor ids reveals Vendor 2 gets a greater quantity of reservations. The bulk of the time, according to the data from Store and\_ fwd flag Count, the taxi driver has not logged into the vendor's systems. Daily distribution of pick-ups and drop-offs indicates that, as compared to other days, weekend taxi booking rates are greater (4- Friday and 5-Saturday). This implies that people once went out on the weekends for festivities, parties, or perhaps other professional activities. Initially, we added a speed column to the feature engineering section, but because to its strong link with the distance and trip time columns, we were forced to remove it. The accuracy of the models was around 99% (false results/predictions) as a result of this high collinearity. So we took it out. There were many outliers in our variables, and when we tried to remove them, we discovered that doing so resulted in significant data loss. We used different strategies to train our model, and the accuracy we obtained was 71%.

# Data understanding and Overview.

Shape: - There were 1458644 rows and 11 columns in given data sets.

brief(user defined function):- This function gives information, about Null value counts, datatypes and unique values counts.

|  |  |  |  |
| --- | --- | --- | --- |
| **feature** | **null values** | **data type** | **unique count** |
| dropoff\_datetime | 0 | object | 1380377 |
| dropoff\_latitude | 0 | float64 | 62519 |
| dropoff\_longitude | 0 | float64 | 33821 |
| id | 0 | object | 1458644 |
| passenger\_count | 0 | int64 | 10 |
| pickup\_datetime | 0 | object | 1380222 |
| pickup\_latitude | 0 | float64 | 45245 |
| pickup\_longitude | 0 | float64 | 23047 |
| store\_and\_fwd\_flag | 0 | object | 2 |
| trip\_duration | 0 | int64 | 7417 |
| vendor\_id | 0 | int64 | 2 |

# Data cleaning: -

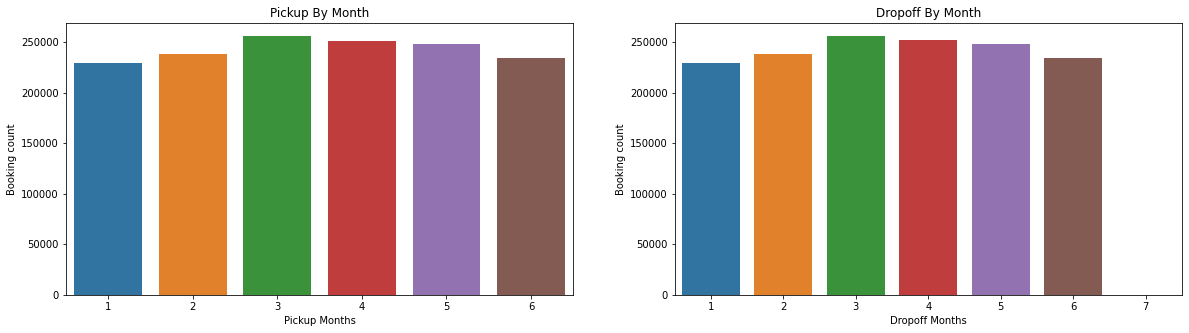
There were no null values in the data. But there were some data types in appropriate for analysis, thus converted them to appropriate dtype , eg. Objects converted to datetime.

There were lots of outliers and inappropriate data in distance and passenger count and Trip duration column.

# EDA Insights generation: -

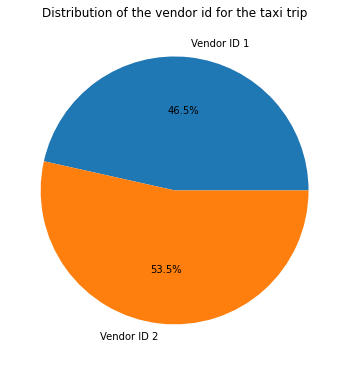
* Various information was generated using date time variable. Like day of month, day of week, month, year etc.
* We used geodesic function from geopy library. To calculated distance between pickup and drop-off coordinates.

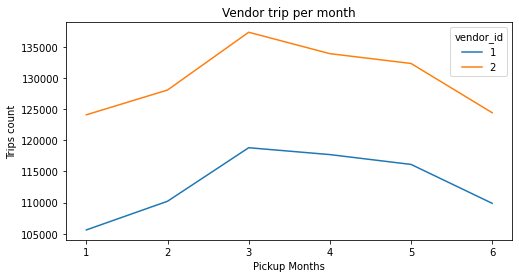
### Pickup and drop-off comparison w.r.t months



As we can observe from the chart that pickups are almost similar in the period of 6 months but still it peaks in month of March and April and is slightly less of other months and least for January.

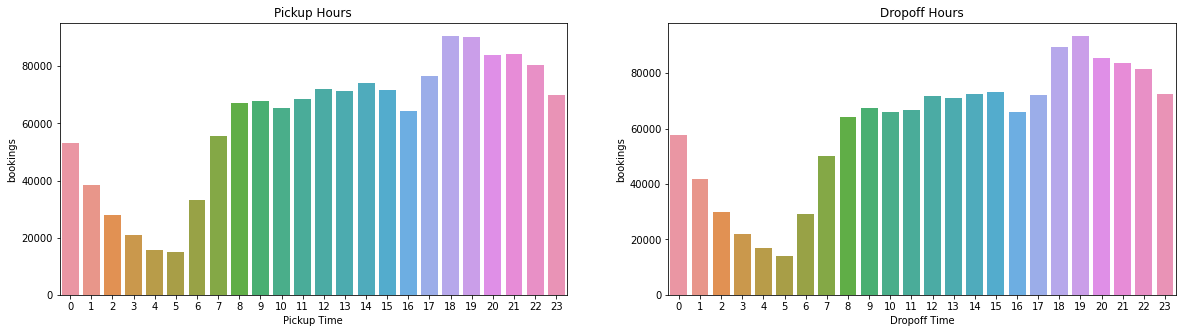
### Monthly trend of Vendor 1 and Vendor 2





As we can observe from the plot above both of the vendor follow the similar trend through the span of 6 months, but vendor 2 is surpassing vendor 1 in number of trips with a big margin.

1. Volume of trips throughout a day.



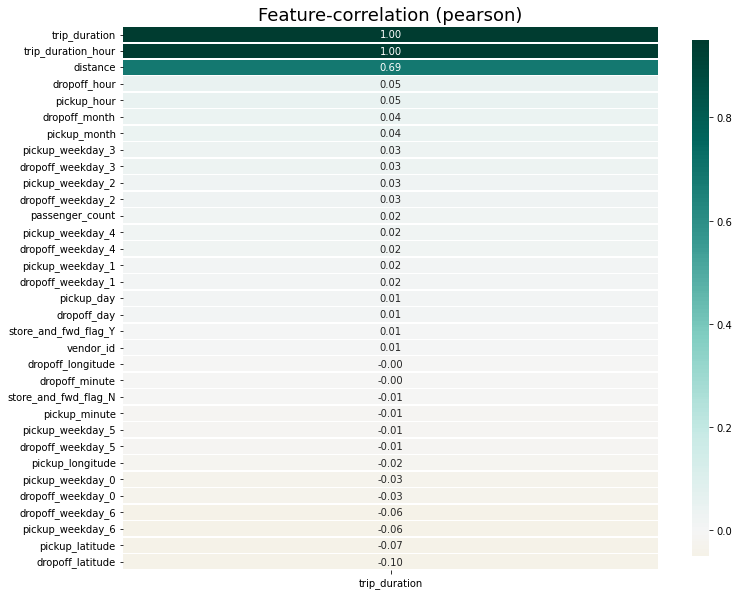
* Distribution of pickup and drop-off hours follows same pattern,
* it shows that most of the pickups and drop-offs are in the evening (x axis represents time in hrs., and peak is >15hr).
* We can see that people often use taxi services to get to their workplaces in the mornings after 10:00. Additionally, the demand for taxis tends to surge in the late evening after six o'clock. Might because of people going back to home after completing their job.

## Multicollinearity studies: -

* Study of collinearity among independent variable.



We can observe that and with the help of a function we found out that, 'dropoff\_day', 'dropoff\_hour', 'dropoff\_month', 'dropoff\_weekday', 'trip\_duration this are the highly correlated columns.



* Study of collinearity between

dependent and independent

variable.We can observe that

distance is having most effect

on dependent variable.

* We Implemented VIF to

find out the least

dependent columns.

# Machine learning: Supervised Regression

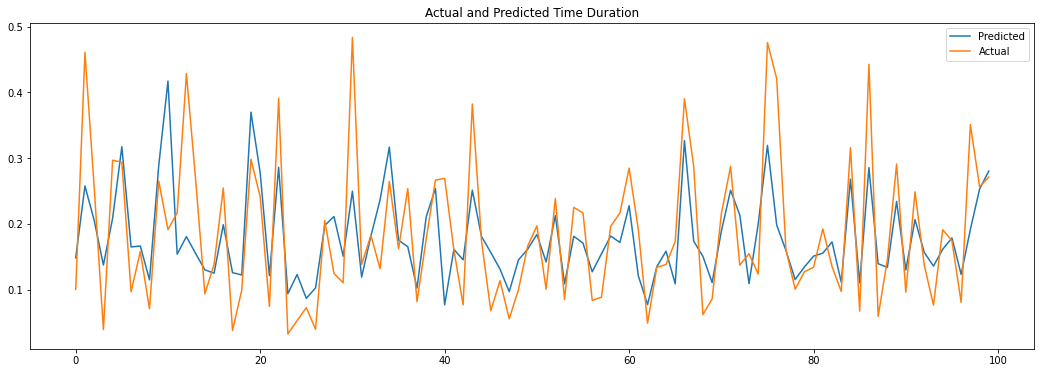
We created dummy variables of Selected columns for machine learning, i.e 'store\_and\_fwd\_flag', 'pickup\_weekday', 'dropoff\_weekday' .

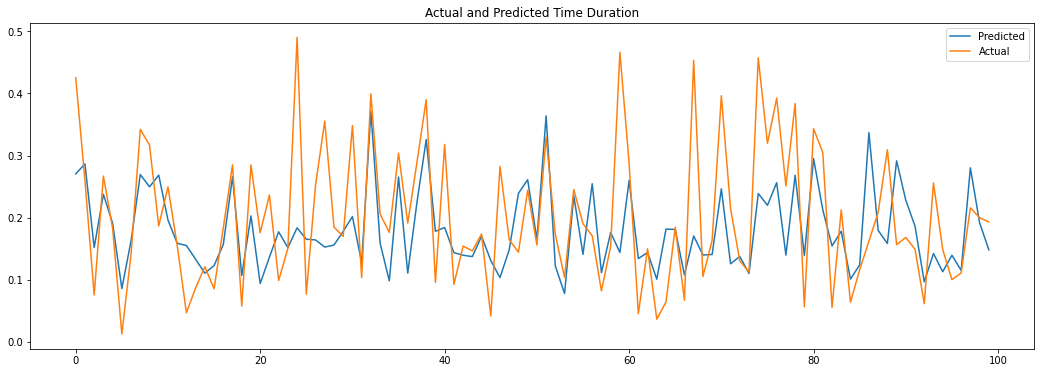
## Model selected for ML: -

* Linear Regression
* DecisionTree Regression
* XGBoost Regression.
* Gradient Boost Regression
* Random Forest Regression.

### Linear regression: -

* Training:
  + Mean Squared Error: 0.0056
  + Root Mean Squared Error: 0.07483314773547883
  + R2 Score : 0.4928134668452804
  + Adjusted R2 Score : 0.49271200797864023

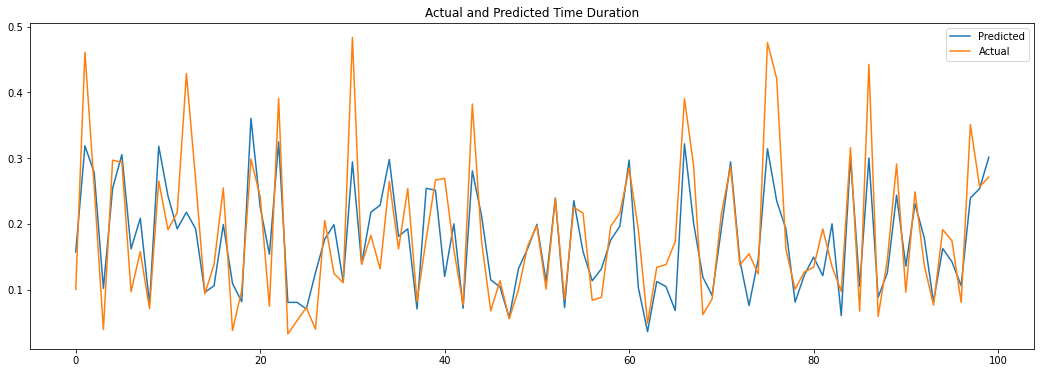


* Testing:
  + Mean Squared Error: 0.0055
  + Root Mean Squared Error: 0.07416198487095663
  + R2 Score: 0.4947690526368055
  + Adjusted R2 Score: 0.49436452402959885

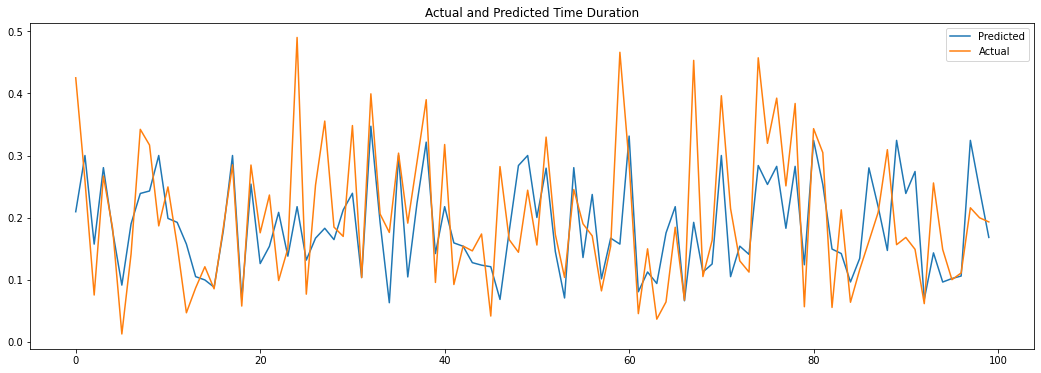
### Decision Tree Regressor: -

We used GridsearchCV to tune hyperparameters.

* Training:
  + Mean Squared Error: 0.0046
  + Root Mean Squared Error: 0.06782329983125268
  + R2 Score: 0.583740252123725
  + Adjusted R2 Score: 0.5836569824793504

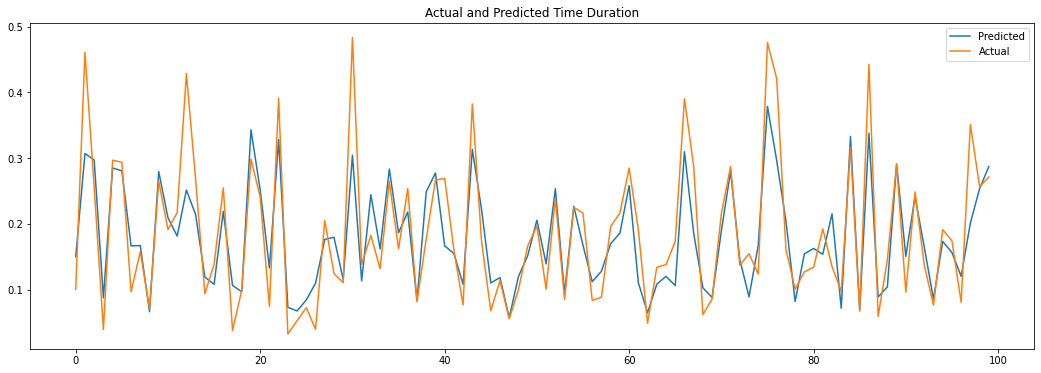


* Testing:
  + Mean Squared Error: 0.0049 Root
  + Mean Squared Error: 0.06999999999999999
  + R2 Score : 0.5471424276844363
  + Adjusted R2 Score : 0.5467798334214604

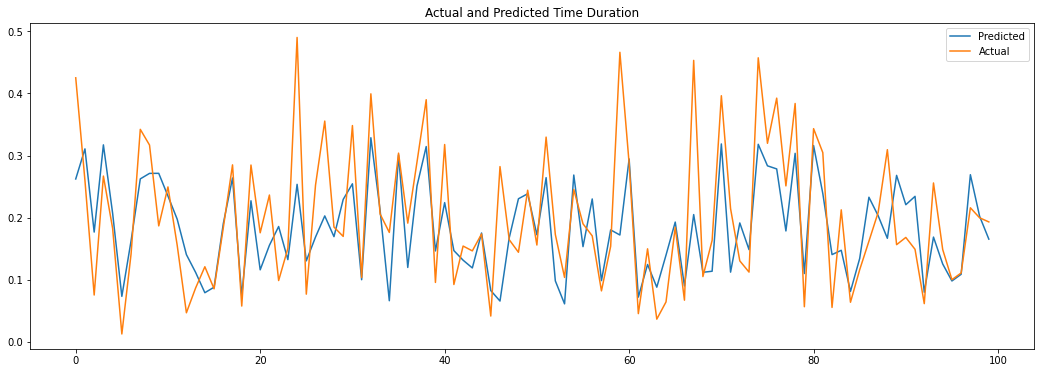


### XGBoost: -

* Training:
  + Mean Squared Error: 0.0031
  + Root Mean Squared Error: 0.055677643628300216
  + R2 Score: 0.7191718076114904
  + Adjusted R2 Score: 0.7191156300352777

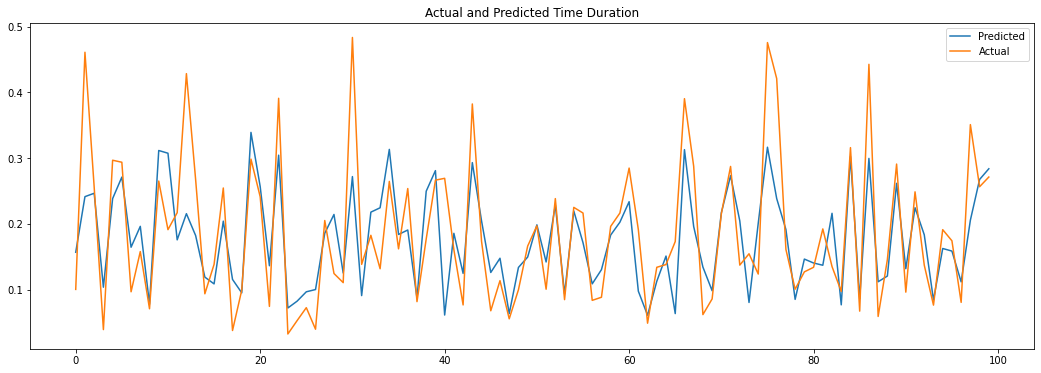


* Testing:
  + Mean Squared Error: 0.0043
  + Root Mean Squared Error: 0.06557438524302
  + R2 Score: 0.6057517406261809
  + Adjusted R2 Score: 0.6054360737017961

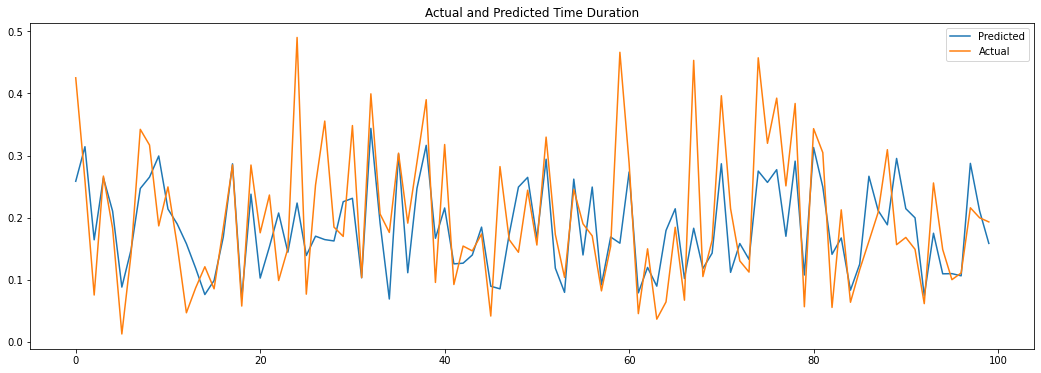


### Gradient Boost Regressor:

* Training:
  + Mean Squared Error: 0.0047
  + Root Mean Squared Error: 0.06855654600401044
  + R2 Score: 0.5719978570834632
  + Adjusted R2 Score: 0.5719122384609225

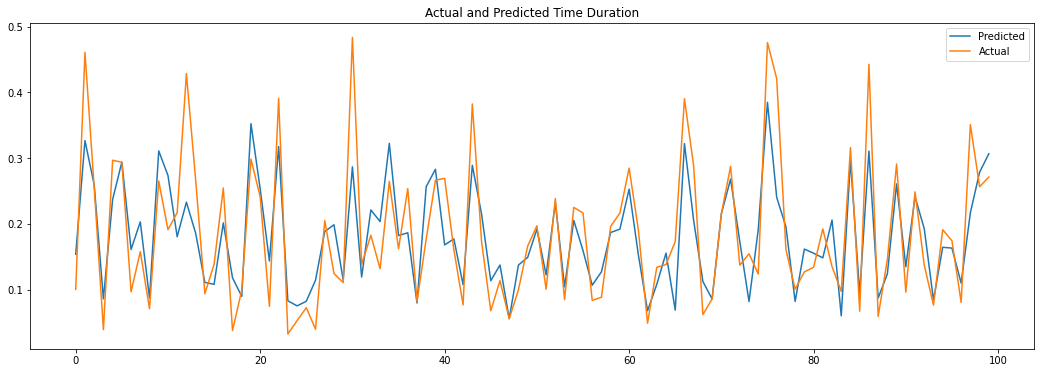


* Testing:
  + Mean Squared Error: 0.0047
  + Root Mean Squared Error: 0.06855654600401044
  + R2 Score: 0.5636976705125473
  + Adjusted R2 Score: 0.563348331710976

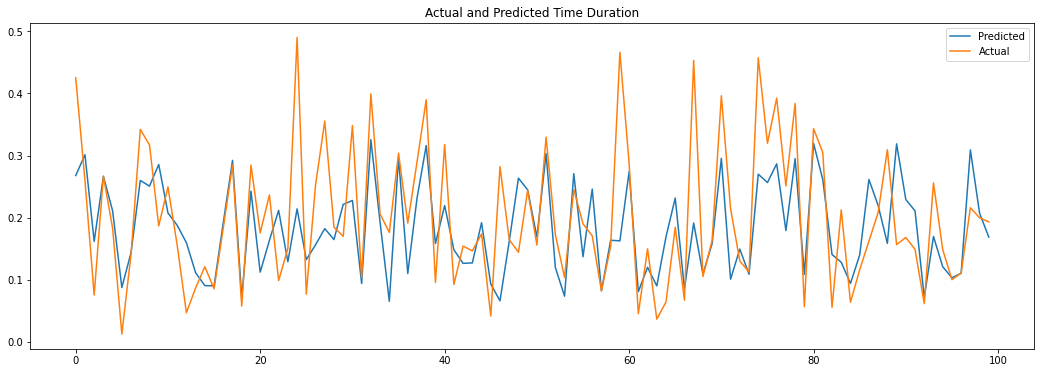


### Random Forest regressor:

* Training:
  + Mean Squared Error: 0.0043
  + Root Mean Squared Error: 0.06557438524302
  + R2 Score : 0.6123412519345506
  + Adjusted R2 Score : 0.6122637037059389



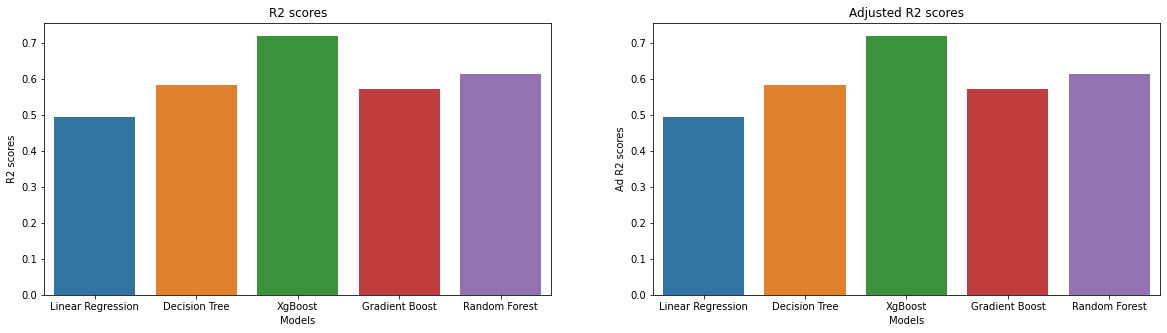
* Testing:
  + Mean Squared Error: 0.0047
  + Root Mean Squared Error: 0.06855654600401044
  + R2 Score: 0.5663396796597682
  + Adjusted R2 Score: 0.5659924562636093



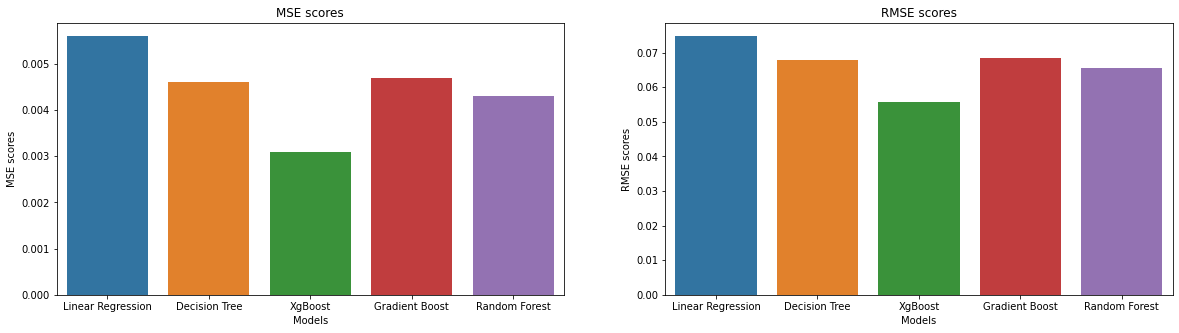
# Evaluation comparison matrix:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | MSE | RMSE | R2 score | Adjusted R2 score |
| Linear regression | 0.0056 | 0.074833 | 0.492813 | 0.492712 |
| Decision tree | 0.0046 | 0.067823 | 0.583740 | 0.583657 |
| XGBoost | 0.0031 | 0.055678 | 0.719172 | 0.719116 |
| Gradient Boost | 0.0047 | 0.068557 | 0.571998 | 0.571912 |
| Random Forest | 0.0043 | 0.065574 | 0.612341 | 0.612264 |

### R2 score vs Adjusted R2 score:



### Mean Square Error vs Root Mean Square Error



# Observation:

As we observed that the Yellow line (Actual) was matching more accurately to Blue line (Predicted) in XGBoost more accurately than any other model.

After comparing there MSE and RMSE scores we can conclude that XGBoost is more appropriate for this data set.

# Conclusion for EDA:

* Vendor id distribution shows Vendor 2 receives a greater number of bookings
* Store\_and\_ fwd\_flag Count shows that majority of the time the taxi driver hasn't logged onto the vendor's systems.
* Distribution of pickups and dropoffs on daily basis interprets that 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 personnel work.
* Distribution of pickups and dropoffs on monthly basis shows that taxi reservations were more in the month of March and April.
* Monthly trend for vendors tells us 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.
* Distribution of pickups and dropoffs on hourly basis gives us the insight 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.
* Passenger count distribution shows that most of the bookings are made by solo travelers, which means a smaller number of people prefer car pool or amy be a smaller number of groups book car...people prefer to ride solo

# Conclusion for Machine Learning:

* Initially in the feature engineering part we added a column called speed but unfortunately, we had to remove it because of its high correlation with distance column and trip duration column. Because of this high collinearity we got accuracy of the models around 99% (false results/predictions). so, we removed it.
* There were a lot of outliers in our variables some values were near to zero, we tried to remove those values but we found that we were losing a lot of data. we trained our model using various algorithms and we got an accuracy of 71%.
* we were curious whether the model was overfit or not, hopefully it was not, as it gave pretty much similar results for train and test data in all the algorithms tried.
* In all the above model's graph we saw that actual and predicted values are almost near to each other (lines coinciding) in only 2 models namely: XG Boost and Random Forest. R2 scores were also high for the above two models and MSE scores were also low in these models which satisfies the requirements of a good model.
* So we came to a conclusion that removing data removes a lot of information, new column if highly collinear can give pseudo good results, also we got our best R2 score from XG Boost model, we tried taking an optimum parameter so that our model doesn’t overfit.