Capstone Project - 2

NYC Taxi Trip Time Prediction

Team Members

Amir Khan
Saurabh Daund
Het Kothari
Kamya Malhotra
Mouleena Jaiswal

Table of contents:

Introduction

New York City is one of the highly advanced cities of the world with extensive use of taxi services. The city taxi rides constitutes the core of the traffic in the city of New York.

The rides taken everyday by many New Yorkers in the lively city can give us a good grasp of traffic times, road blockages, and so on.

With ridesharing apps becoming more and more prevalent, it is increasingly significant for taxi companies to provide visibility to their estimated ride duration, since the competing apps bestow these metrics upfront.

Problem Statement

The main aim is to build a model that predicts the total ride duration of taxi trips in New York City. The primary dataset is one released by the NYC Taxi and Limousine Commission, which includes pickup time, geo-coordinates, number of passengers, and several other variables.

Data Overview

Data overview is nothing but understanding the data better.

The objectives of data understanding are as follows:

- id - A unique identifier for each trip.
- vendor_id - A code indicating the provider associated with the trip record.
- pickup_datetime - Date and time when the meter was engaged.
- dropoff_datetime - Date and time when the meter was disengaged.
- passenger_count - The number of passengers in the vehicle. (driver entered value)
- pickup_longitude - The longitude where the meter was engaged.
- pickup_latitude - The latitude where the meter was engaged.
- dropoff_longitude - The longitude where the meter was disengaged.
- dropoff_latitude - The latitude where the meter was disengaged.
- store_and_fwd_flag - This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip.

Continued...

- Summarize the data by identifying key characteristics, such as data volume and total number of variables in the data.

  Number of rows in our dataset are 1458644.

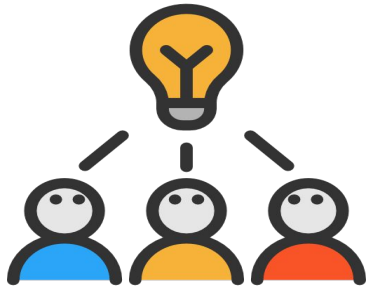
  Number of columns in our dataset are 11.

- Understand the problems with the data, such as missing values, inaccuracies, and outliers.
- There are no NAN/NULL values in our dataset.

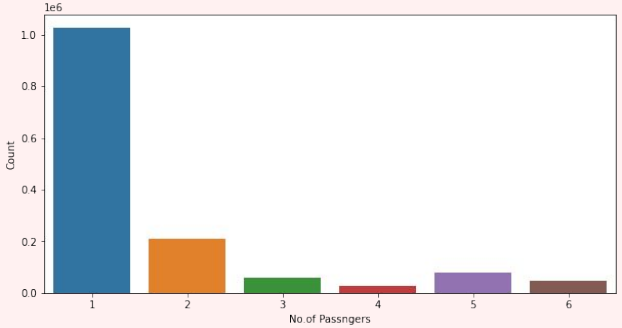
Feature Creation

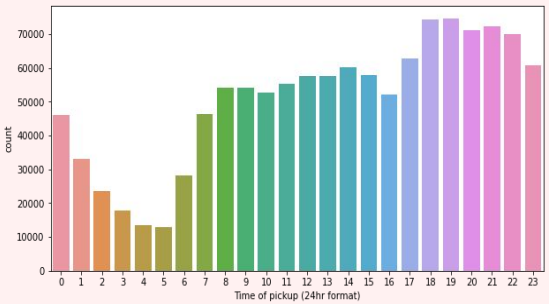
We have created the following features:

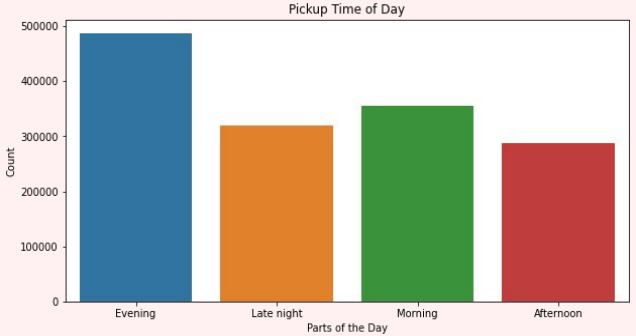
- pickup_weekday which contains the name of the day on which the ride was taken.
- pickup_weekday_num which contains the day number instead of characters with Monday = 0 and Sunday = 6.
- pickup_hour with an hour of the day in the 24 - hour format.
- pickup_month with month number as January = 1 and December = 12.
- Distance from geographical coordinates.
- Speed in km/h.
- Time of the day the ride was taken . **Morning** (from 6:00 am to 11:59 pm), **Afternoon** (from 12 noon to 3:59 pm), **Evening** (from 4:00 pm to 9:59 pm) and **Late Night** (from 10:00 pm to 5:59 am).

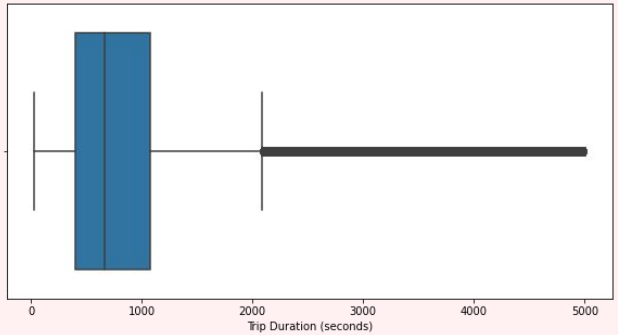


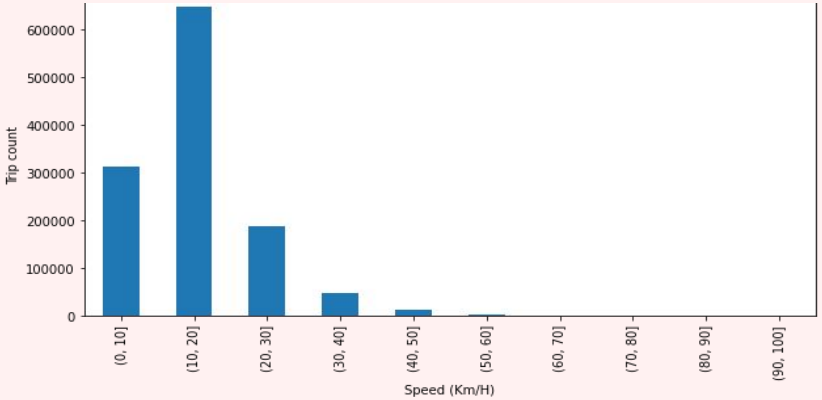
EDA Univariate analysis







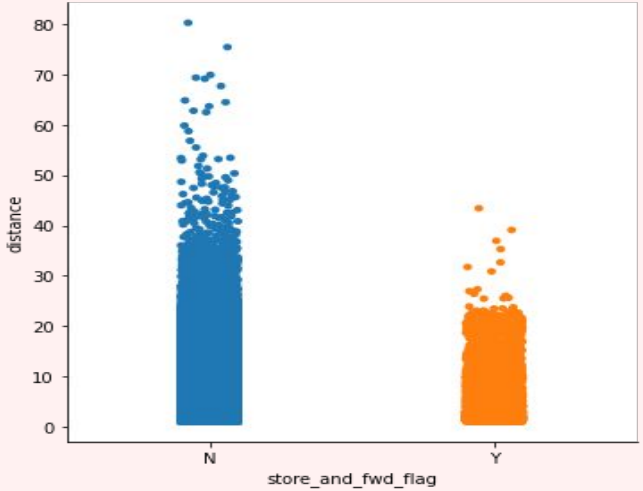


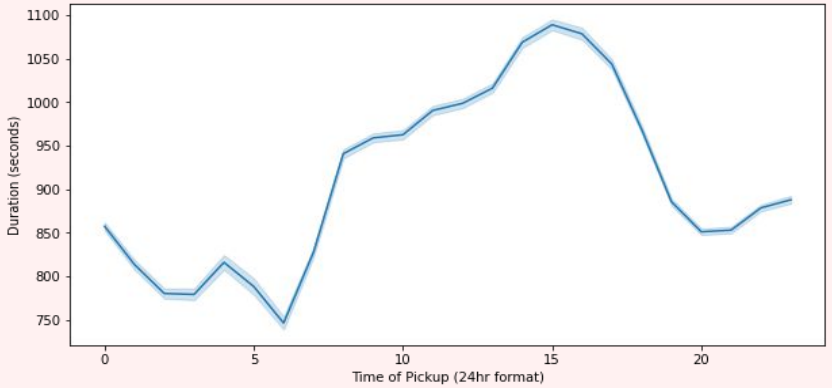


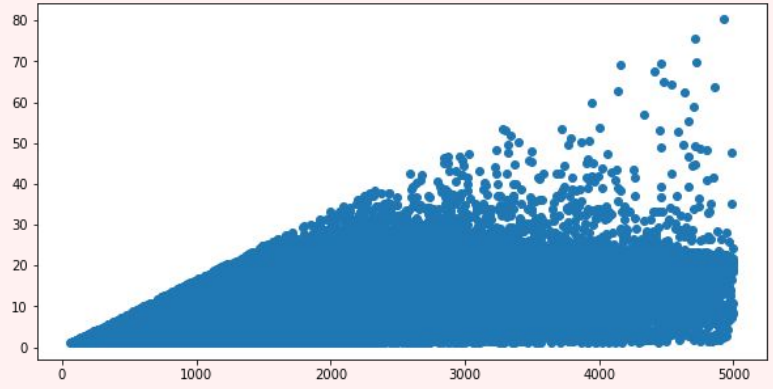
Insights from Univariate Analysis:

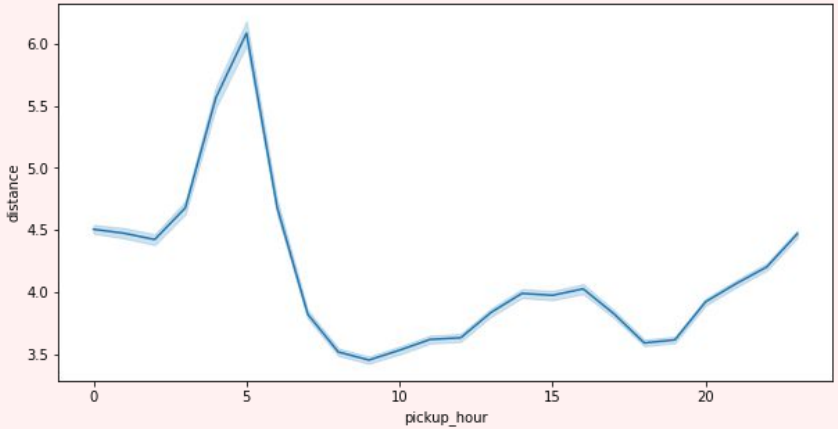
- Passenger : It is evident that most of the trips was taken by single passenger and that is inline with our day to day observations.
- Total trips Per Hour : The plot consist of the distribution of the pickups across the 24 hour time scale. It's inline with the general trend of taxi pickups which starts increasing from 6AM in the morning and then declines from late evening i.e. around 8 PM. There is no unusual behavior here.
- Trips per Time of Day : As we saw above, most of the taxi pickups are at evening followed by morning which could be because of school and jobs.
- Trip Duration : There are some durations with as low as 1 second which points towards trips with 0 km distance. Major trip durations took between 10-20 minutes to complete.
- Speed : Mostly trips are done at a speed range of 10-20 km/h.

EDA Bivariate analysis









Insights from Bivariate Analysis:

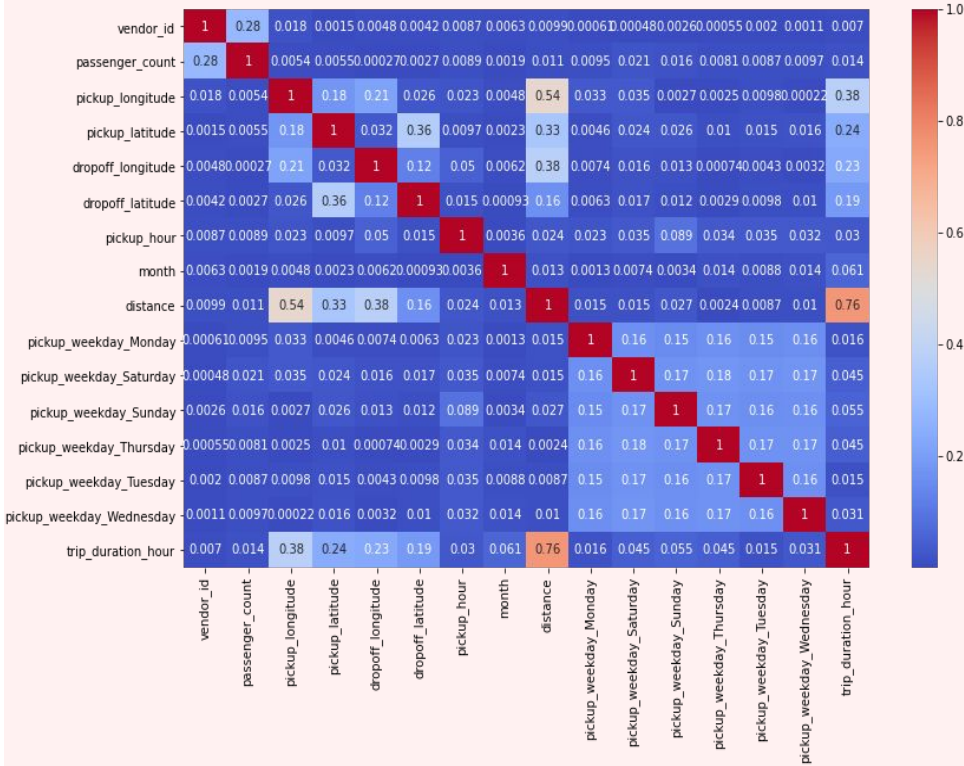
- Trip duration v/s Flag : Trip durations scale is less for the trips where the flag is set i.e. the trip details are stored before sending it to the server. Trip duration is longer for the trips where the flag is not set.
- Trip distance per hour : It is highest during early morning hours which can account for some things such as outstation trips taken during the weekends. Also because of longer trips towards the city airport which is located in the outskirts of the city.
- Trip duration per hour : Trip duration on an average is similar during early morning hours i.e. before 6 AM and late evening hours i.e. after 6 PM.
- Distance v/s Trip duration : There should have been a linear relationship between the distance covered and trip duration on an average but we can see dense collection of the trips in the lower right corner which showcase many trips with the inconsistent readings.

Feature Engineering

One Hot Encoding :
Dummify features like store_and_fwd_flag and pickup_weekday.

Feature Selection:
We remove columns which are not important for further analysis such as id, pickup_datetime,dropoff_datetime, store_and_fwd_flag, pickup_weekday, dropoff_weekday,pickup_weekday_num, pickup_timeofday, trip_duration,speed.

Correlation Analysis:
We draw heatmap to find correlation between different independent features and dependent feature. If correlation between independent features are high and has very less relation with dependent feature, remove them.



Model Creation

- **Linear Regression** : The linear regression model finds the set of θ coefficients that minimize the sum of squared errors.
- **Lasso Regression** : The lasso method was used to shrink coefficients. For duration prediction models, lasso was run using a range of values for the penalizing parameter, λ . Grid Search was used to find the lasso model with the lowest error and select the value of λ to use.
- **Ridge Regression** : To further confirm the best set of covariates to use, the regression method was used. It performs L2 regularization, i.e. adds penalty equivalent to square of the magnitude of coefficients.

Model Creation(continued)

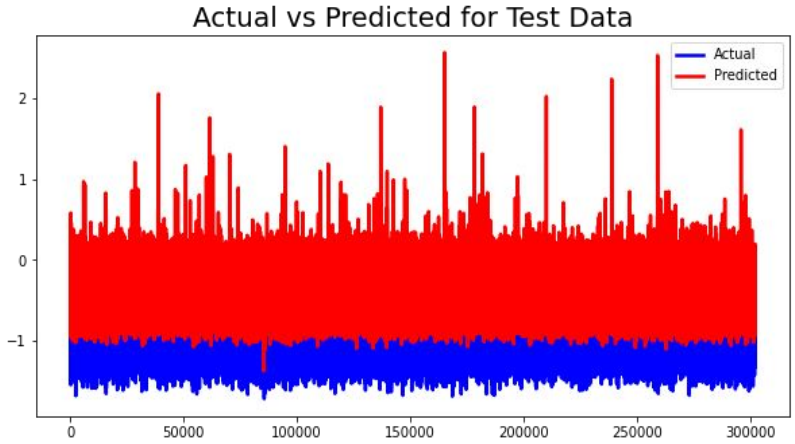
- **Decision Tree** :The decision trees was also built on the training data in order to improve prediction accuracy .We used GridSearch to tune the hyperparameters of Decision Tree to get the best possible test score.

- **XGBoost** was used for final prediction of the trip duration in the test dataset. The dataset was very large, as a result for this type of problem XGBoost was applied in which all the attributes were taken and parallel processing of boosting trees executed. Another aspect of XGBoost is that it keeps a nice check between bias and variance which helps in better prediction. The results were interpreted by using GridSearch, the XGBoost hyperparameters .

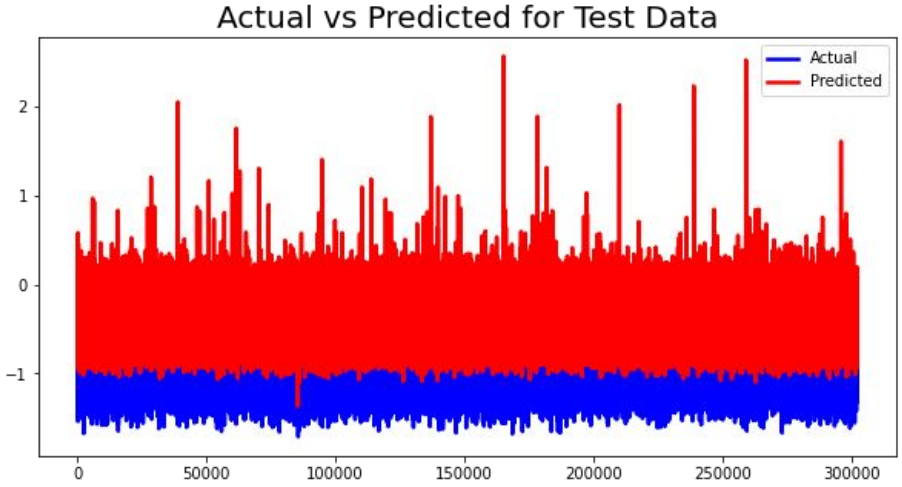
Model Evaluation

Training Model	Train MSE	Test MSE	Train RMSE	Test RMSE	Train R2	Test R2	Train Adjusted R2	Test Adjusted R2
Linear Regression	0.0389	0.039	0.197	0.197	0.475	0.475	0.475	0.474
Lasso Regression	0.038	0.039	0.197	0.197	0.475	0.475	0.475	0.475
Ridge Regression	0.038	0.039	0.197	0.197	0.475	0.475	0.475	0.475
Decision Tree	0.022	0.022	0.148	0.149	0.701	0.697	0.701	0.697
XGBoost	0.012	0.0139	0.112	0.117	0.831	0.813	0.831	0.813

Actual v/s Predicted

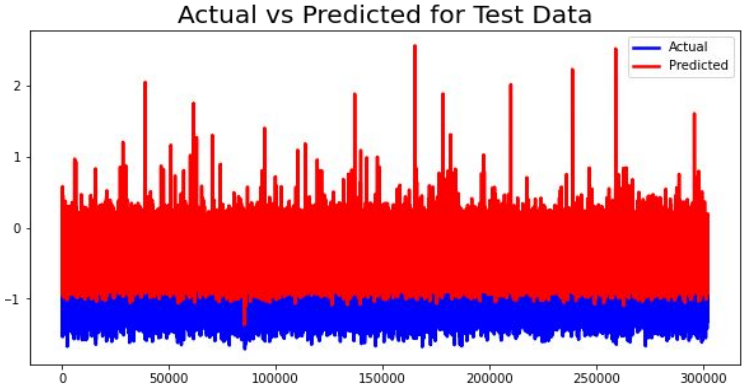


Linear Regression

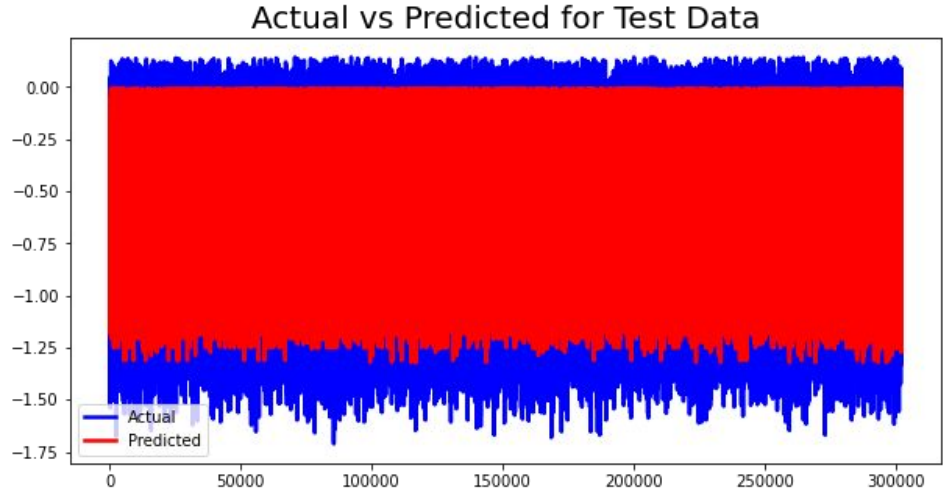


Lasso Regression

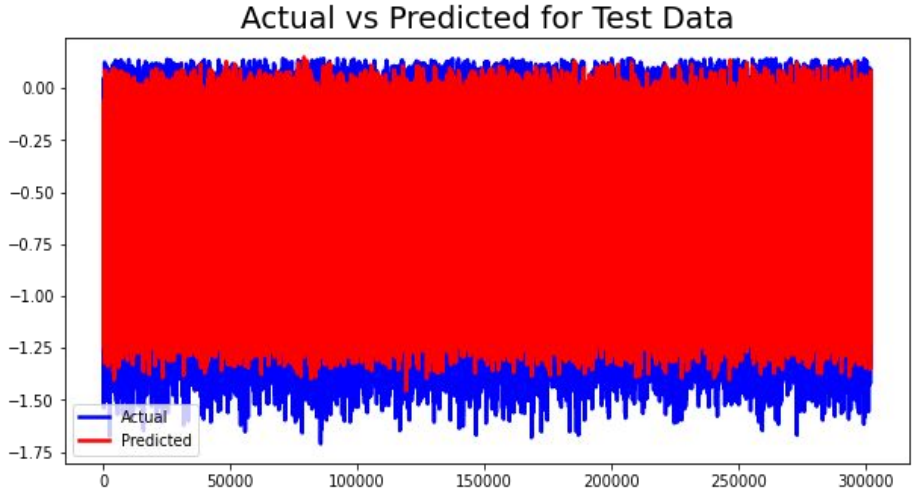
Ridge Regression



Actual v/s Predicted(continued)



Decision Tree



XG Boost

Conclusion

- For **Linear regression model,** MSE and RMSE for training and testing are similar but has very poor R2 for training and testing data.
- **Lasso regression** and **Ridge regression** R2 increases , but not with significant amount.
- We can see that MSE and RMSE of **Decision Tree** model are not varying much during training and testing time. Also the R2 is almost same for training and testing time.
- MSE and RMSE of **XGBoost** model are very similar and their R2 is 80.
- From above table, we can conclude **XGBoost** is best model for our dataset.

Challenges

- Large dataset to handle.
- Need to Remove outliers
- Carefully handled feature selection part as it affects the R2 score.
- Carefully tuned Hyperparameters as it affects the R2 score.

THANK YOU!!