Capstone Project-2

NYC TAXI TRIP TIME PREDICTION

BY

KISHAN KUMAR SINGH

CONTENTS:

- Introduction
- Defining problem statement
- Data summary
- Feature creation
- Exploratory data analysis
- Feature Engineering
- Model creation
- Model evaluation

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

Task is to build a model that predicts the total ride duration of taxi trips in New York City. The 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 summary

- 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 the NYC Taxi and Limousine Commission (TLC). The data was sampled and cleaned for the purposes of this project.
- Dataset has 1458644 rows and 11 columns.
- Data has no null values.

Continued…..

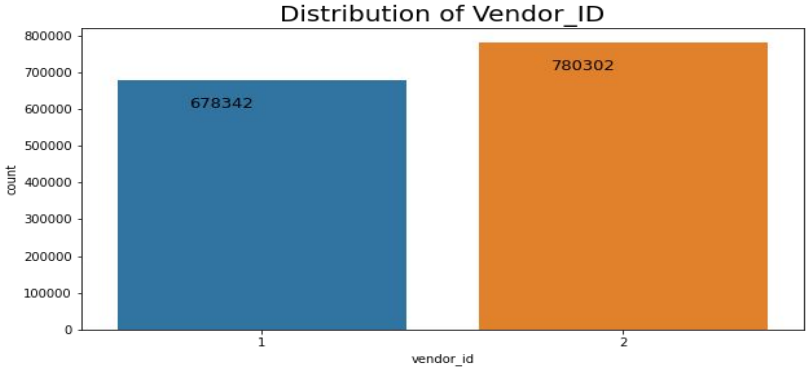
- id - A unique identifier for each trip.
- vendor_id - A code indicating the provider associated with the trip record.
- pickup_datetime - Date and time when the meter was engaged.
- dropoff_datetime - Date and time when the meter was disengaged.
- passenger_count - The number of passengers in the vehicle. (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.
- trip_duration - duration of the trip in seconds

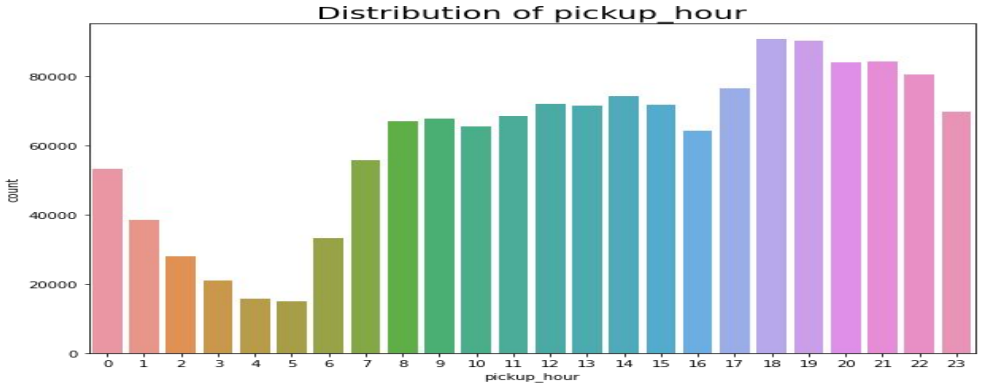
Feature creation

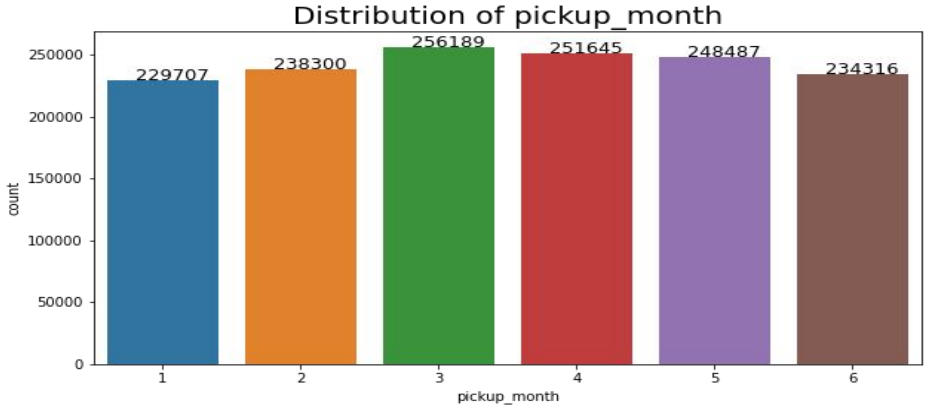
We have created the following features:

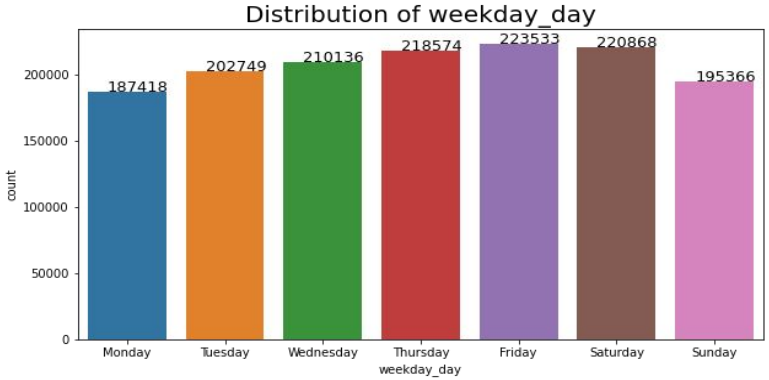
- Distance feature is created by the help of pickup(latitude & longitude) And dropoff(latitude & longitude).
- Speed feature is created by the help of distance and trip duration features.
- Features like pickup hours, pickup month, pickup weekday is extracted from pickup datetime feature.
- pickup_hour with an hour of the day in the 24 - hour format.
- pickup_month with month number as January = 1 and December = 12.
- Speed in km/h.

EDA Univariate analysis

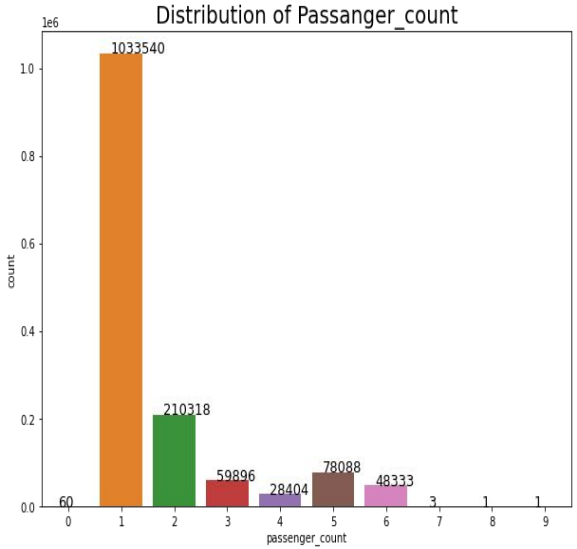


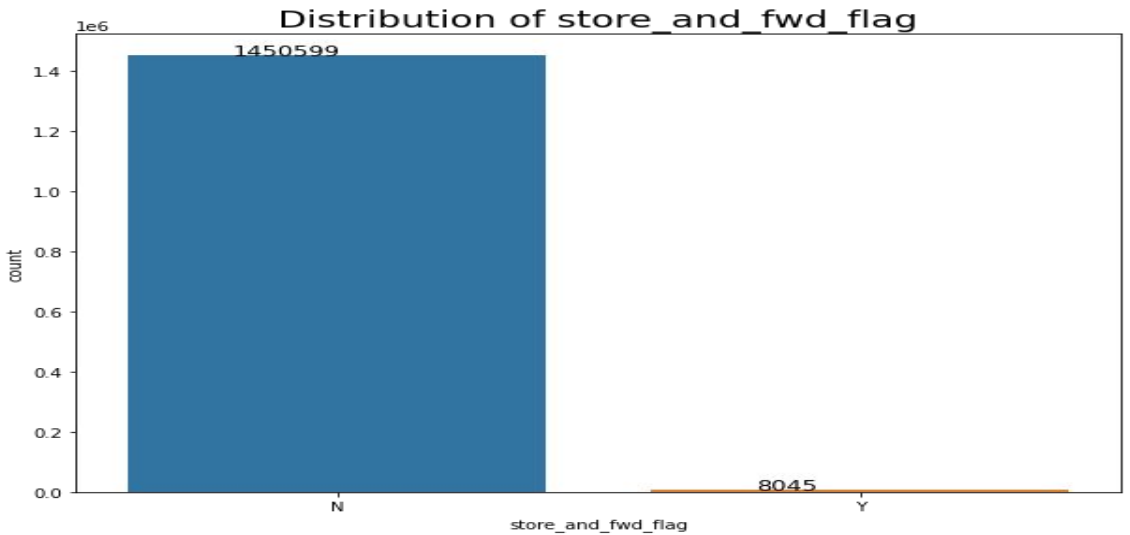


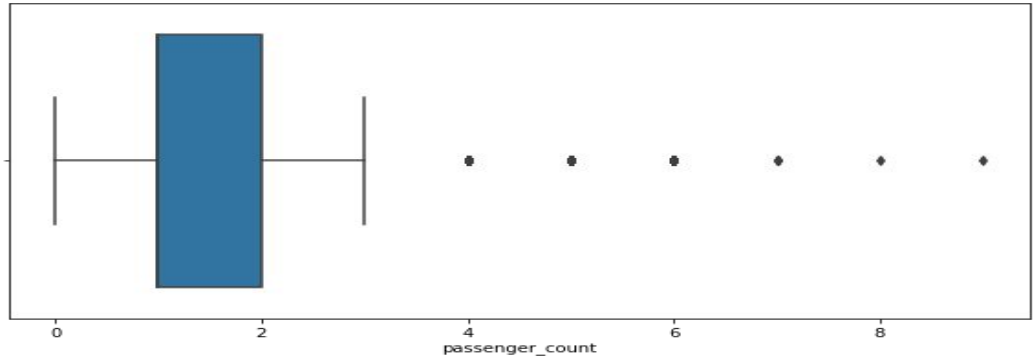




continued….


Distribution of Passanger_count
1e6
1033540
210318
59896
28404
78088
48333
60
3
1
1
count
passenger_count

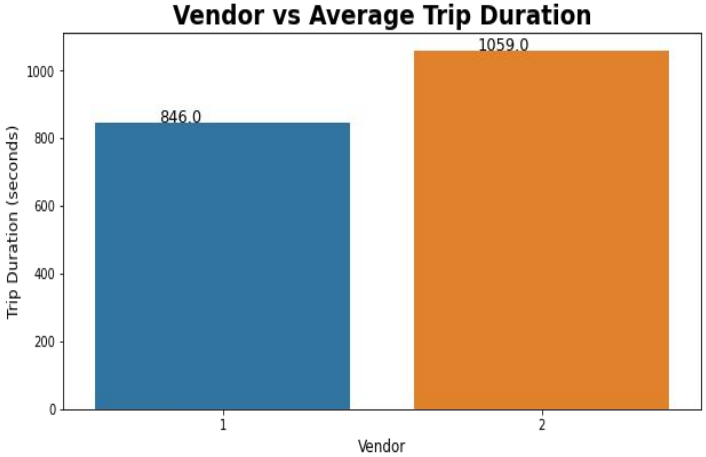

Distribution of store_and_fwd_flag
1e6
1450599
8045
count
N
Y
store_and_fwd_flag

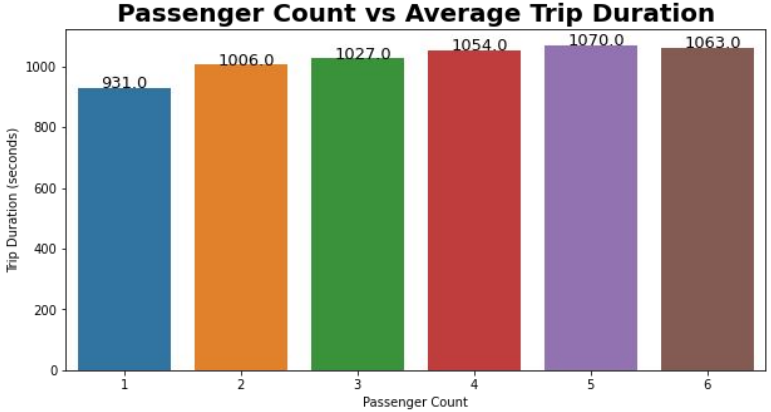

passenger_count

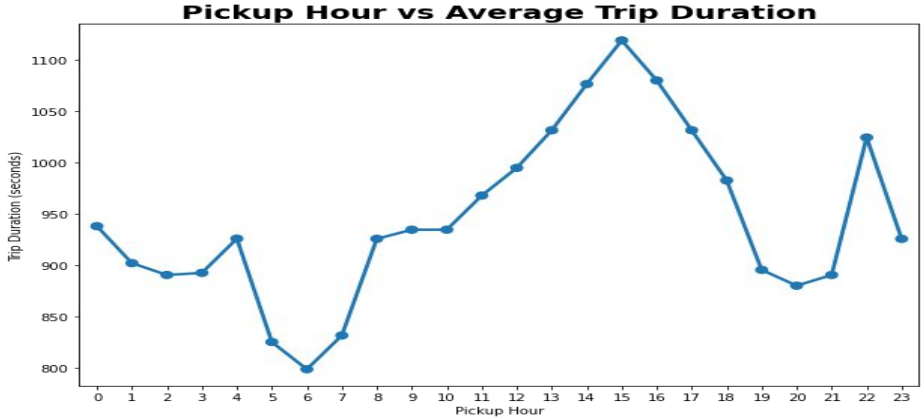
Insights from Univariate Analysis:

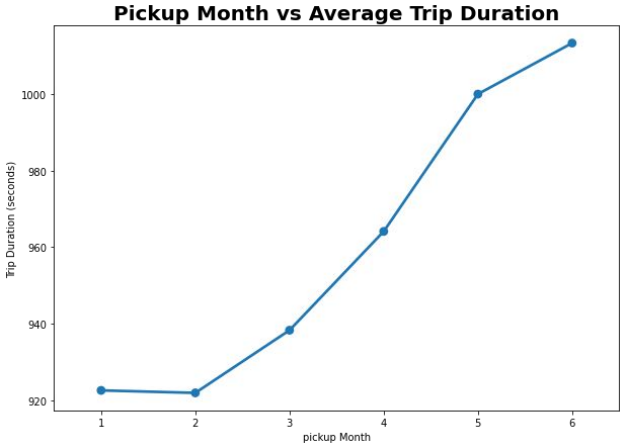
- vender with id 2 has large market share than vender with id 1
- only about 1% of the trip details were stored in the vehicle first before sending it to the server.
- pickup hour is on its peak at 18 to 22 hour (evening) ,followed by 12 to 15 (Afternoon). and pickup hour is on its lowest at 4 to 5 hour (early morning).
- dropoff_datetime has a extra month compared to pickup_datetime column with only 127 observations when dropoff_datetime month is 7 , this is because people have taken ride at late night of last date of 6th month. Trip across the month is nearly balanced
- an increasing trend of taxi pickups starting from Monday till Friday. The trend starts declining from saturday till monday which is normal where some office going people likes to stay at home for rest on the weekends.
- most of the trip is taken by the single passenger and there are very fewer no. of trip is taken by more than 6 passengers and they might be outliers because fitting more then is not possible. 6 people is possible which contain 1 child and driver. 0 is outlier too.
- there are some trips with 0 passenger count, few trips consist of 7,8,9 passenger , they are outliers, most of the trips consist of 1 or 2 passengers.

EDA Bivariate analysis

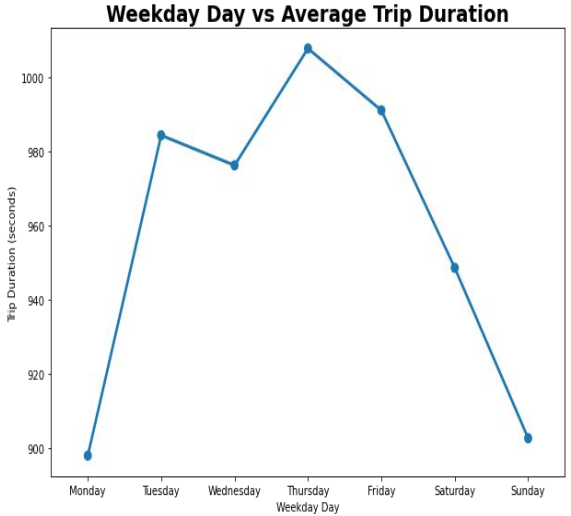


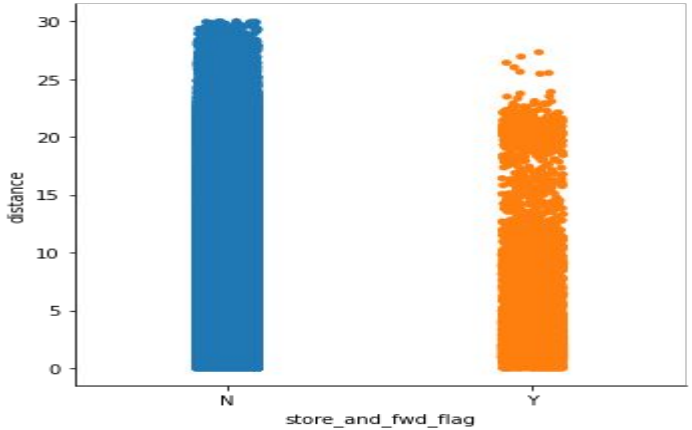


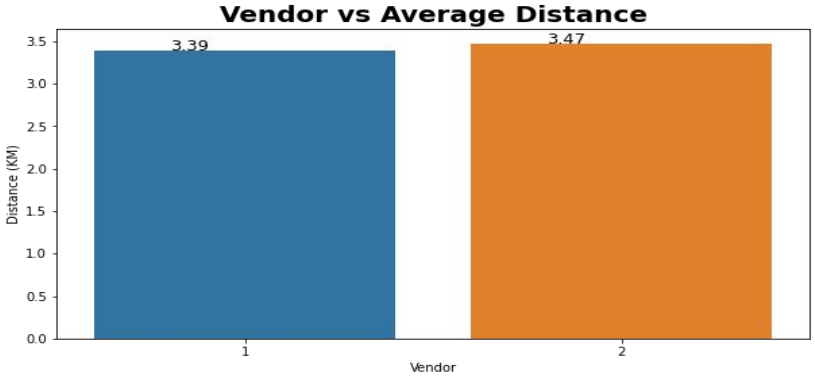


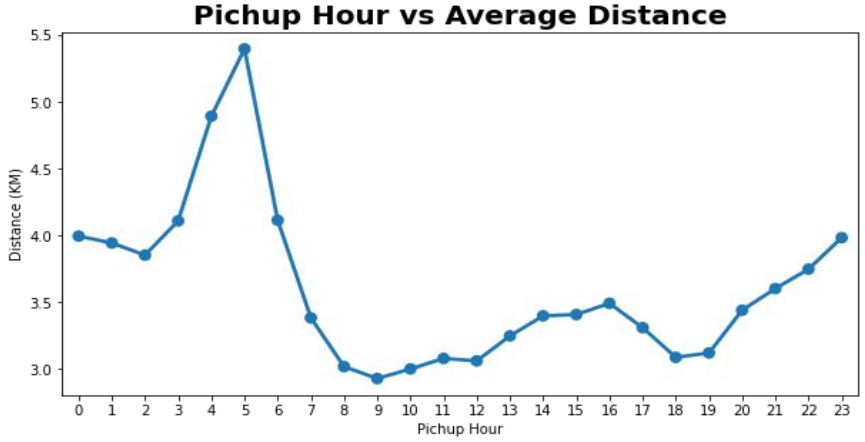


continued…..

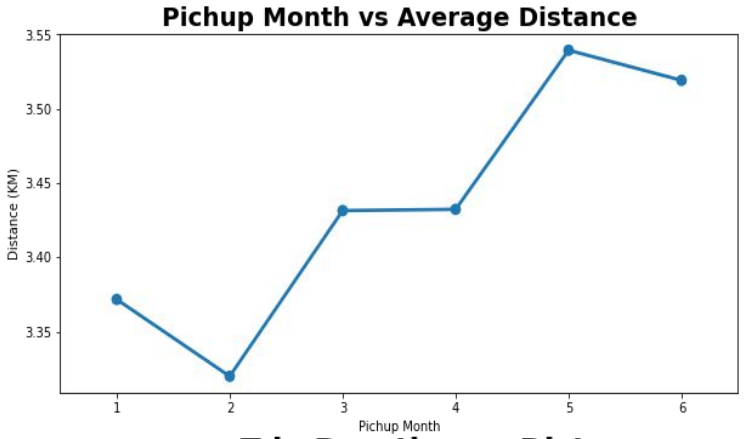

Weekday Day vs Average Trip Duration
Trip Duration (seconds)
1000
980
960
940
920
900
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Weekday Day

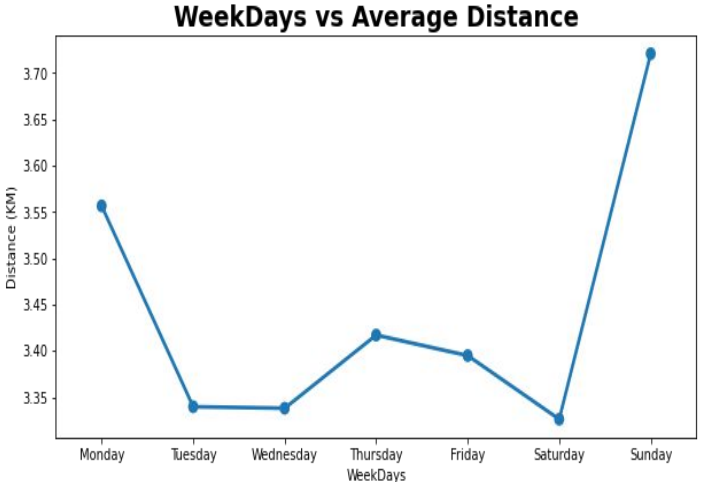

distance
30
25
20
15
10
5
0
N
Y
store_and_fwd_flag

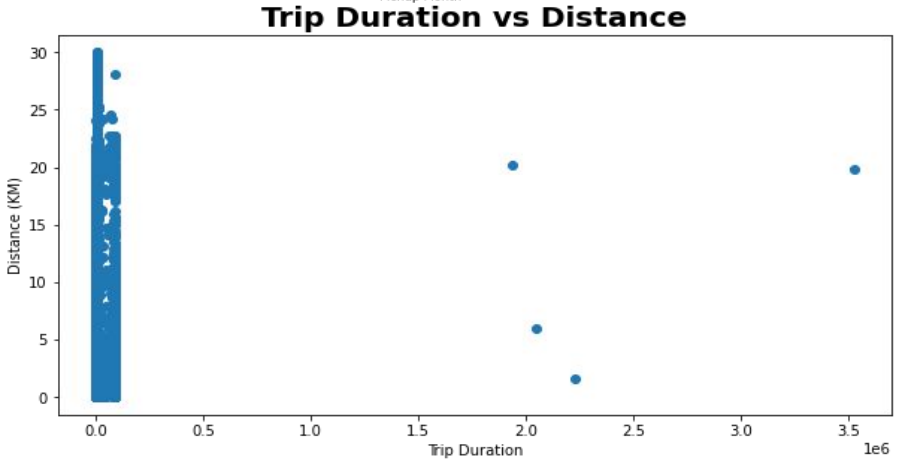

Vendor vs Average Distance
Distance (KM)
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
3.39
3.47
1
2
Vendor

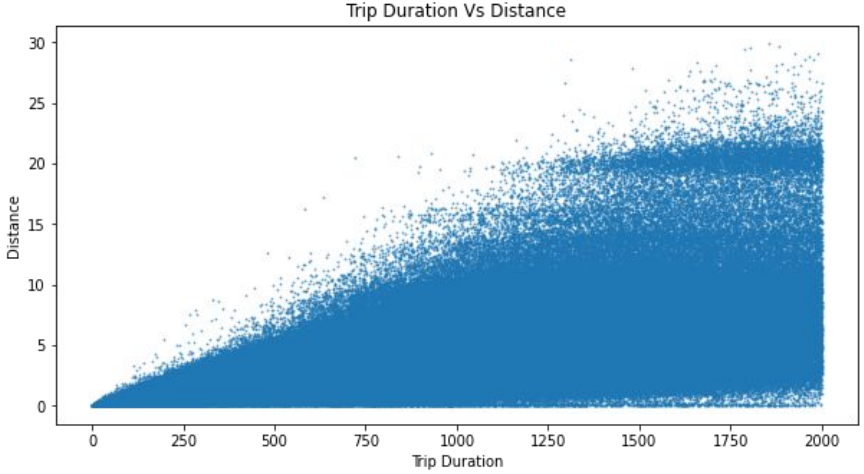

Pichup Hour vs Average Distance
Distance (KM)
5.5
5.0
4.5
4.0
3.5
3.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Pichup Hour

continued…..

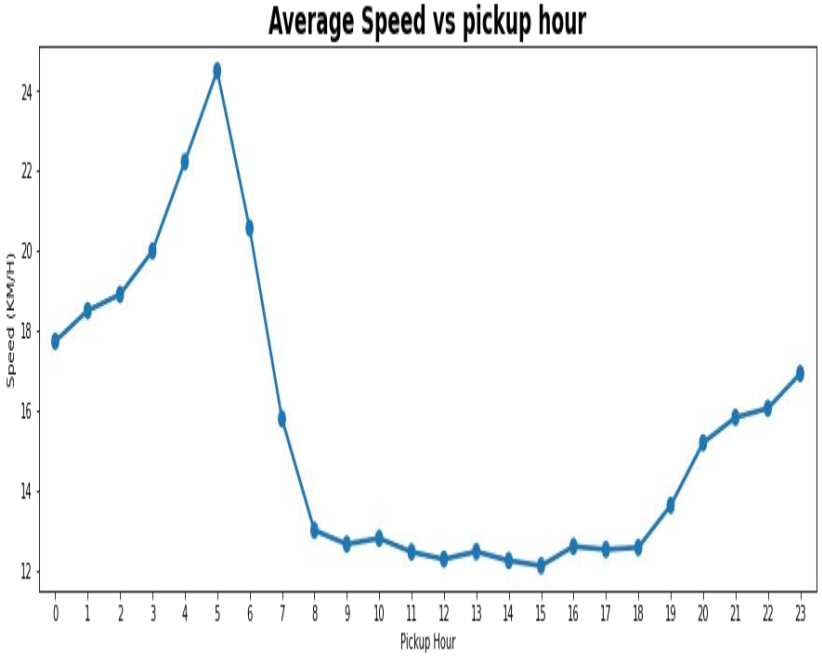

Pichup Month vs Average Distance
Distance (KM)
Pichup Month
3.55
3.50
3.45
3.40
3.35
1
2
3
4
5
6

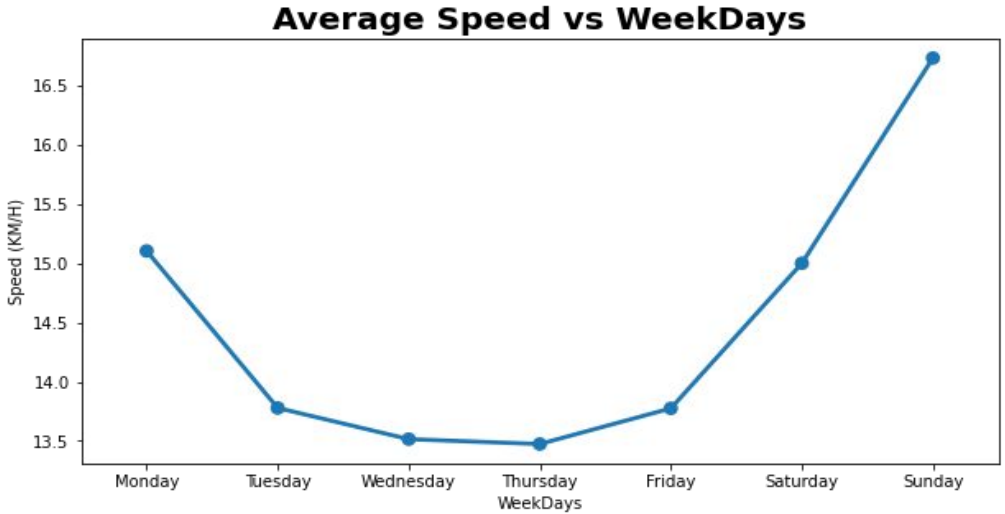

WeekDays vs Average Distance
Distance (KM)
WeekDays
3.70
3.65
3.60
3.55
3.50
3.45
3.40
3.35
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday

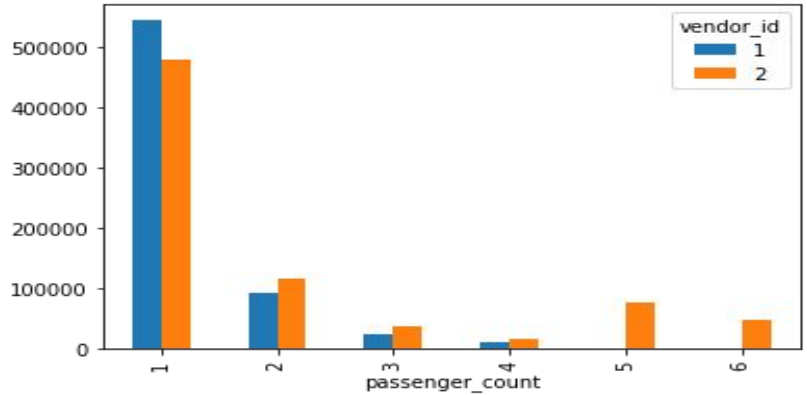

Trip Duration vs Distance
Distance (KM)
Trip Duration
1e6
30
25
20
15
10
5
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5


Trip Duration Vs Distance
Distance
Trip Duration
30
25
20
15
10
5
0
0
250
500
750
1000
1250
1500
1750
2000

continued…..


Average Speed vs pickup hour
Speed (KM/H)
Pickup Hour


Average Speed vs WeekDays
Speed (KM/H)
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
WeekDays


vendor_id
1
2
passenger_count

Insights from Bivariate Analysis:

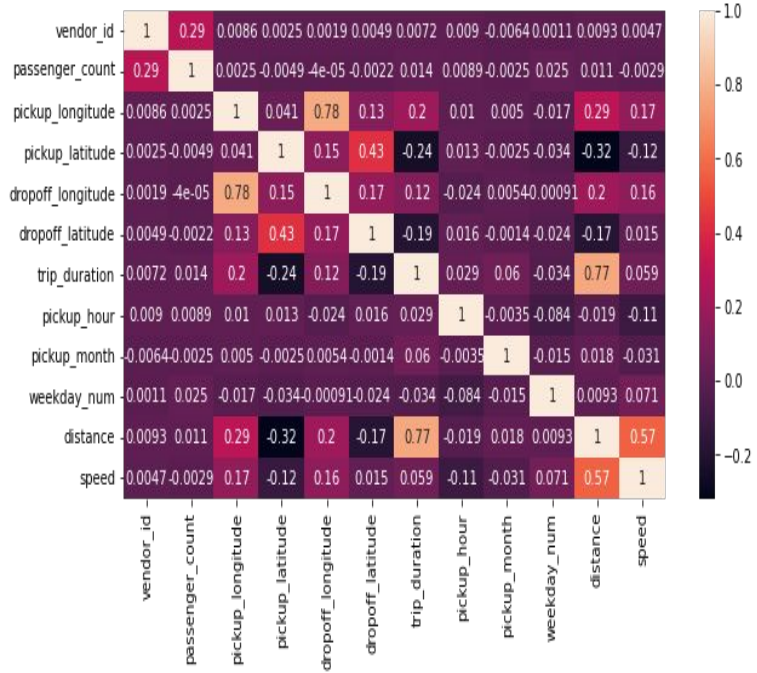
- Average trip duration of Vendor 2 is higher than Vendor 1 by 200 seconds approx
- Trip duration for all the passenger counts are approx similar. except passenger count 1(average trip duration of passenger count 1 is lower than others).
- Average trip duration is lower at 6am in the morning due to minimal traffic on the road. and higher at 15 hour(3pm) because streets are busy at this time.
- Average Trip duration is increasing along with each subsequent month,
- During winter jan and feb average trip duration is less may be due to less traffic in winters but as summer approaches the average trip duration is started increasing.
- Average trip duration is almost same and lowest on monday and sunday and highest on thursday.
- Most of the trip is not held in vehicle memory and longer distance trip'S flag was not stored.
- Average Trip distance is highest in the early morning between 4 to 6 am, and kinda similar from morning to evening varying around 3 to 3.5 km, then it started increasing from evening throughout late night till 5 am in the morning.
- Average Trip distance is lowest in Feb and highest in may,Average Trip distance is varying around 3 to 3.5 km across the month except 2nd month(feb).
- there is non linear relationship between trip duration and distance may be because of traffics and all
- vender 2 has more market share than vendor 1, And vender 2 are carrying more number of people which means vendor 2 has more big cars or may be vendor 2 allow more passenger to sit, vendor 1 has large market share in term of single passenger may be because vendor 1 has more minicars than big cars.
- average speed is highest in the early morning between 4 to 6 am which is in range of 20 to 24 km/h, when traffic is less. and started decreasing when office hours approaches.

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,trip_duration, weekday_num.

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.

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

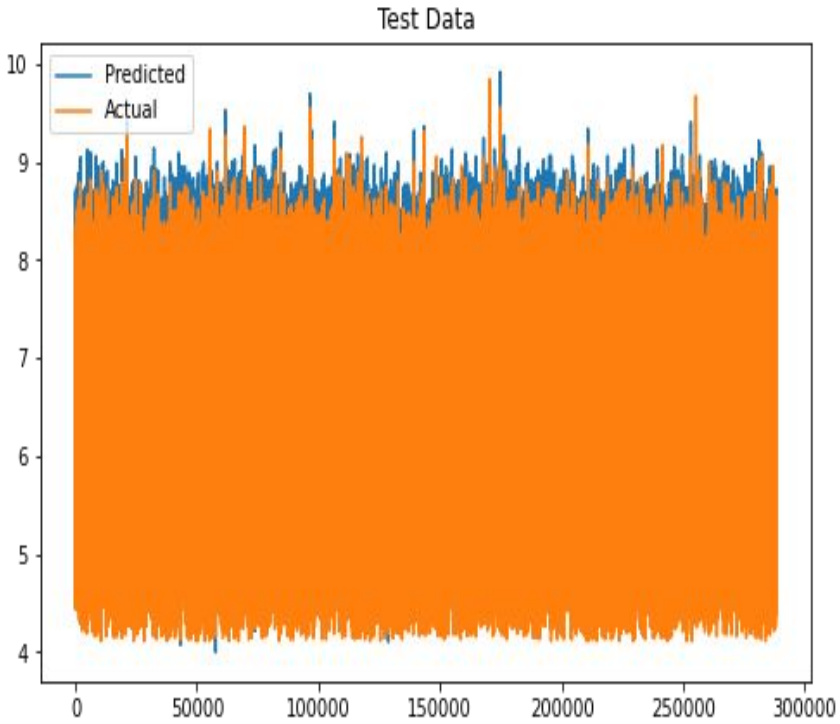
SL NO	MODEL_NAME	Train MSE	Train RMSE	Train R^2	Train Adjusted R^2
1	Linear Regression	0.02148703636629707	0.14658457069656775	0.959471484820573	0.9594710631935075
2	Lasso Regression	0.02148535175361829	0.14657882436975092	0.9594746623109146	0.9594742407169052
3	Ridge Regression	0.02148703636629707	0.14658457069656775	0.959471484820573	0.9594710631935075
4	DecisionTree Regressor	0.002677742254977419	0.0517469057526865	0.9949492840344577	0.9949492314907477
5	XGBRegressor	0.0015590983418606208	0.03948541935779106	0.9970592528566001	0.9970592222633602

SL NO	MODEL_NAME	Test MSE	Test RMSE	Test R^2	Test Adjusted R^2
1	Linear Regression	0.021207763545873486	0.14562885547127494	0.9599319775206714	0.9599303101213527
2	Lasso Regression	0.0212046626498388	0.14561820851060764	0.9599378360720342	0.9599361689165145
3	Ridge Regression	0.021207763545873486	0.14562885547127494	0.9599319775206714	0.9599303101213527
4	DecisionTree Regressor	0.002735699493674182	0.052303914706971814	0.9948314178167762	0.9948312027302838
5	XGBRegressor	0.001580998756573382	0.03976177506819058	0.9970130045263307	0.9970128802248576

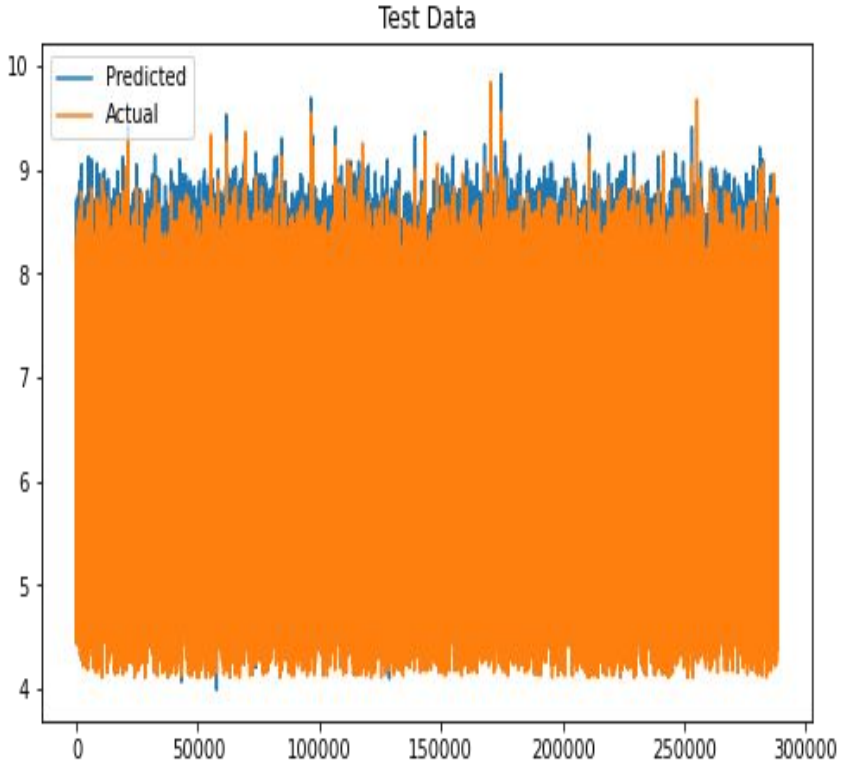
Actual v/s Predicted

Linear Regression

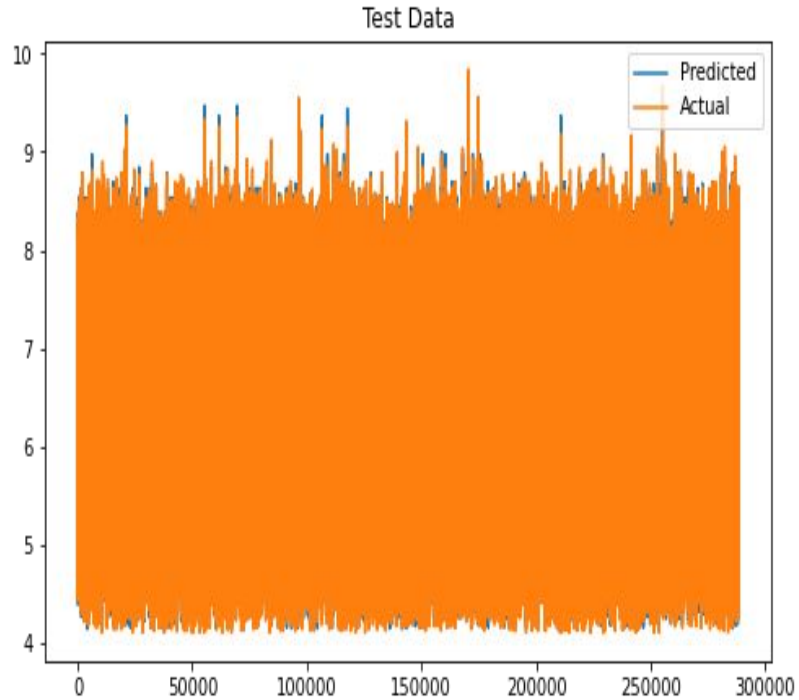
Ridge Regression



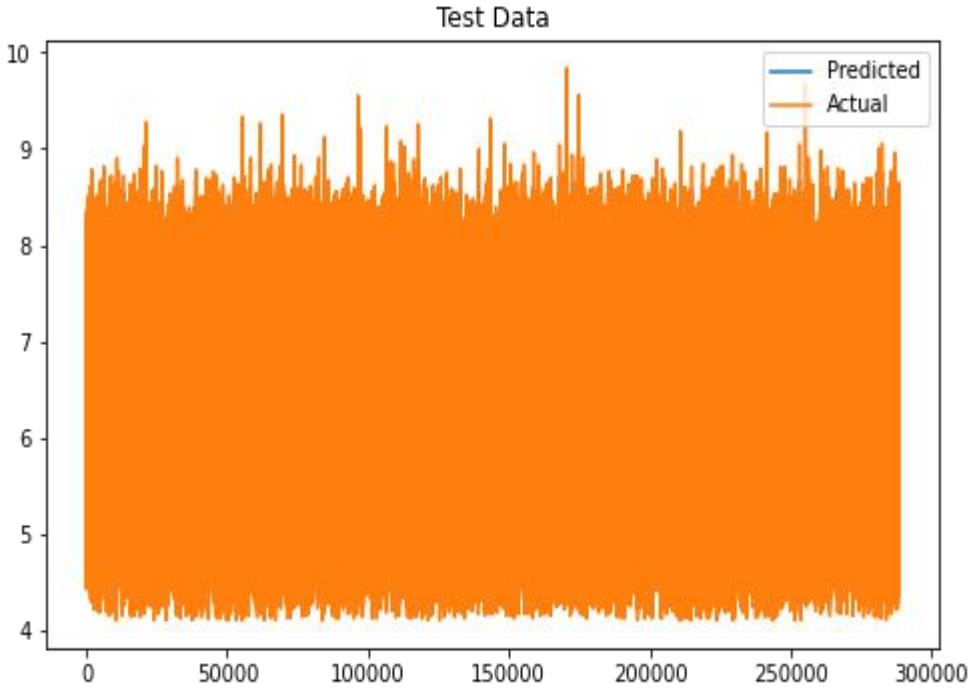
Lasso Regression



Decision Tree



XGBRegressor



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.

Conclusion

- Most passengers travel alone.
- only about 1% of the trip details were stored in the vehicle first before sending it to the server.
- pickup hour is on its peak at 18 to 22 hour (evening) ,followed by 12 to 15 (Afternoon). and pickup hour is on its lowest at 4 to 5 hour (early morning).
- During winter jan and feb average trip duration is less may be due to less traffic in winters but as summer approaches the average trip duration is started increasing.
- Average Trip distance is highest in the early morning between 4 to 6 am, and kinda similar from morning to evening varying around 3 to 3.5 km, then it started increasing from evening throughout late night till 5 am in the morning.
- vender 2 has more market share than vendor 1, And vender 2 are carrying more number of people which means vendor 2 has more big cars or may be vendor 2 allow more passenger to sit, vendor 1 has large market share in term of single passenger may be because vendor 1 has more minicars than big cars.
- XGboost and Decision tree regressor best fits our data with r2 score of 99.70 and 99.48 respectively.

Thank you