Capstone Project

Supervised ML – Regression

New York Taxi Trip Duration Prediction

Contributor – Neha Pasi

New York City Taxi

New York City taxi rides form the core of the traffic in the city of New York. The many rides taken every day by New Yorkers in the busy city can give us a great idea of traffic times, road blockages, and so on. Predicting the duration of a taxi trip is very important since a user would always like to know precisely how much time it would require of him to travel from one place to another. Given the rising popularity of app-based taxi usage through common vendors like Ola and Uber, competitive pricing has to be offered to ensure users choose them. Prediction of duration and price of trips can help users to plan their trips properly, thus keeping potential margins for traffic congestions. It can also help drivers to determine the correct route which in-turn will take lesser time as accordingly. Moreover, the transparency about pricing and trip duration will help to attract users at times when popular taxi app-based vendor services apply surge fares.

Defaulters

1. Defining Problem Statement
2. EDA
3. Feature Selection
4. Preparing dataset for modeling
5. Applying model
6. Model selection

Data Summary

The data set is of shape(1458644, 11)

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

Data Summary

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 (Dependent Variable) - duration of the trip in seconds

Data Summary

Numeric – 1. passenger_count
2. pickup_longitude
3. pickup_latitude
4. dropoff_latitude
5. dropoff_longitude
6. trip_duration

Categorical – 1. Vendor_id
2. store_and_fwd_flag

Unique - 1. id

String – 1.pickup_datetime
2.dropoff_datetime

Futuristic feature

trip_duration_hour – trip duration in hours
pickup_day - values for each day for pickup
dropoff_day - values for each day for dropoff
pickup_timezone – pickup in 4 timezones
dropoff_timezones – dropoff in 4 timezones
pickup_hour – pickups in hours
dropoff_hour – dropoffs in hours
pickup_month – pickups in months
dropoff_month – dropoffs in months
log_distance – log of distance
distance – Calculating distance by using pickup_latitude, pickup_longitude, dropoff_latitude, dropoff_longitude with haversine formula.

Speed – calculating speed with distance and trip_duration_hour by using speed formula

Data Pipeline

Data Processing

1. First we have observe that the no of records with passenger count 0 ,9 and 7 are very small compared to the entire data set. hence, we will drop the values

2. We have checked for null values in columns and We don’t have any null values in any column.

3. The 2 columns pickup_datetime and dropoff_datetime is object type. we have converted them in datetime format.

4. The time part is represented by hours,minutes and seconds so, we have divide the times into 4 time zones: morning (4 hrs. to 10 hrs.) , midday (10 hrs. to 16 hrs.) , evening (16 hrs. to 22 hrs.) and late night (22 hrs. to 4 hrs.)

Data Pipeline

5. In trip_duration (dependent variable) 4 observations where far away from our other observations. So, we have removed them.

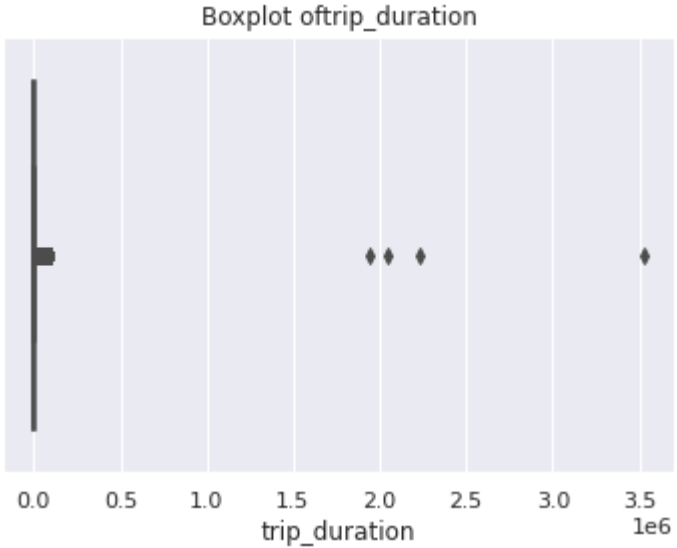
EDA – In this part we have done some exploratory data analysis on the features and data we have selected.

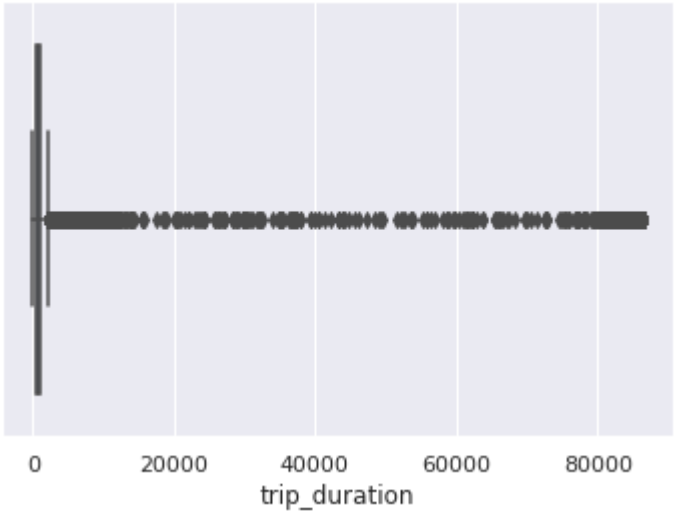
Fitting the different models – We have fitted four models

1. Linear regression
2. Random Forest Algorithm
3. GradiantBoosting Algorithm
4. XGBoost Algorithm

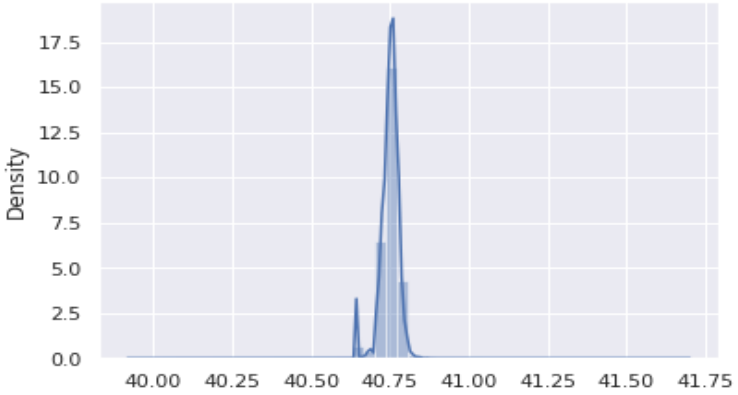
Selecting the best model for our data

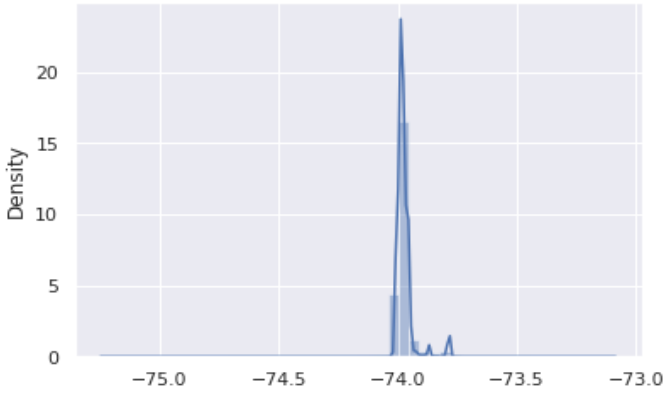
EDA

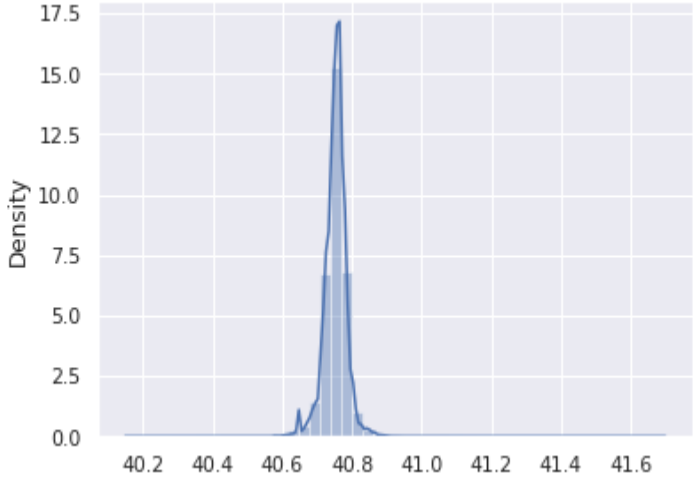

Boxplot oftrip_duration
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
1e6
trip_duration

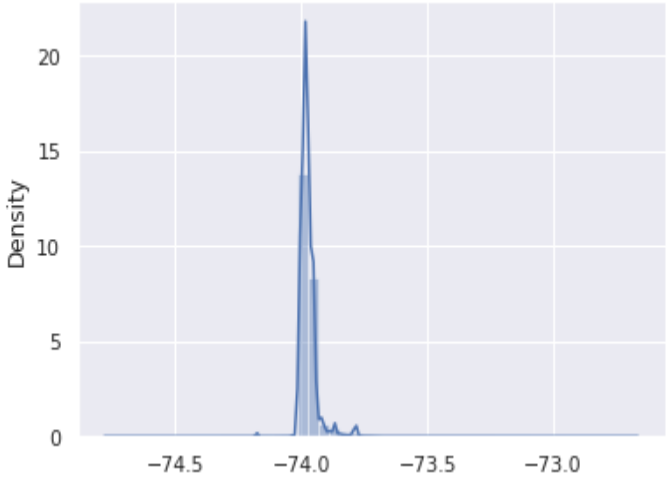

0
20000
40000
60000
80000
trip_duration

EDA

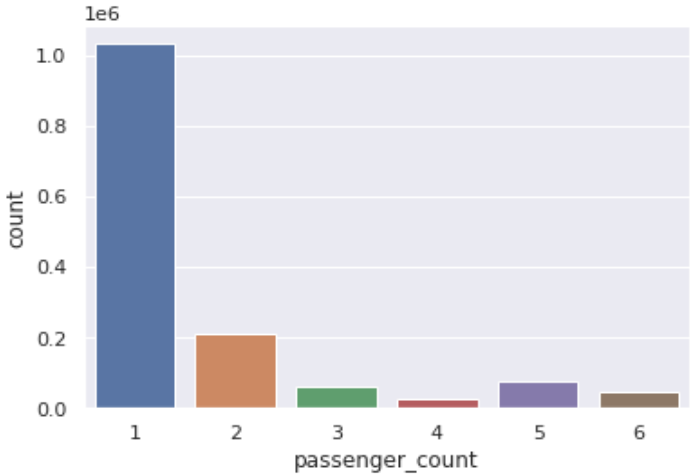

Density
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
40.00
40.25
40.50
40.75
41.00
41.25
41.50
41.75

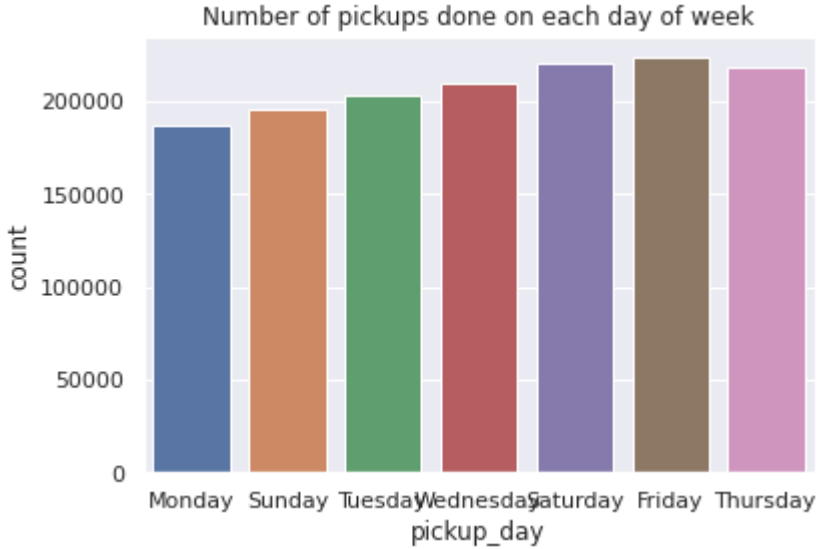

Density
20
15
10
5
0
−75.0
−74.5
−74.0
−73.5
−73.0

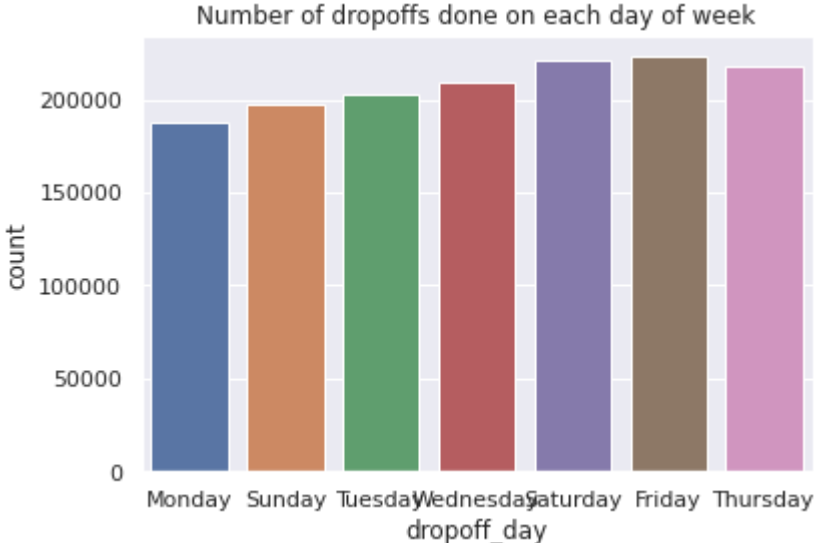

Density
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
40.2
40.4
40.6
40.8
41.0
41.2
41.4
41.6


Density
20
15
10
5
0
−74.5
−74.0
−73.5
−73.0

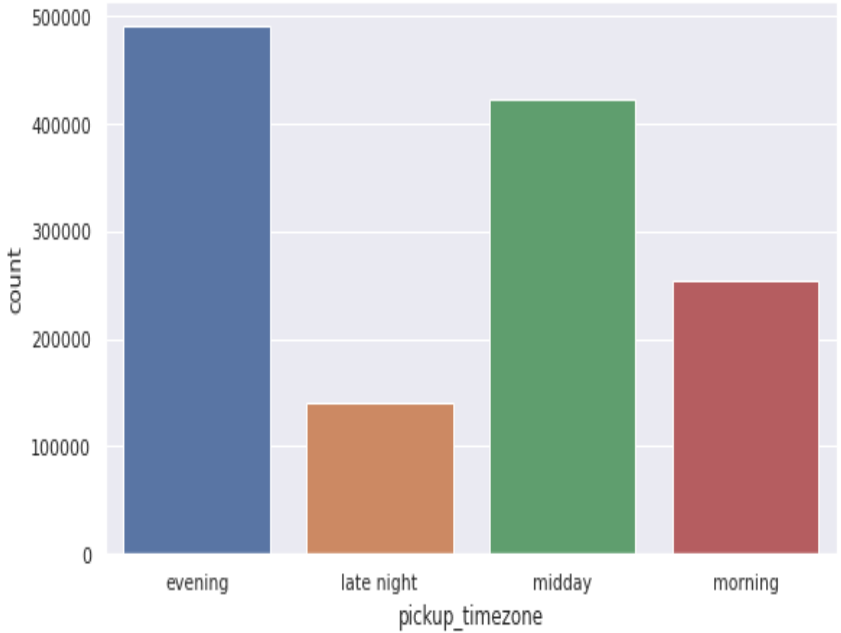
EDA

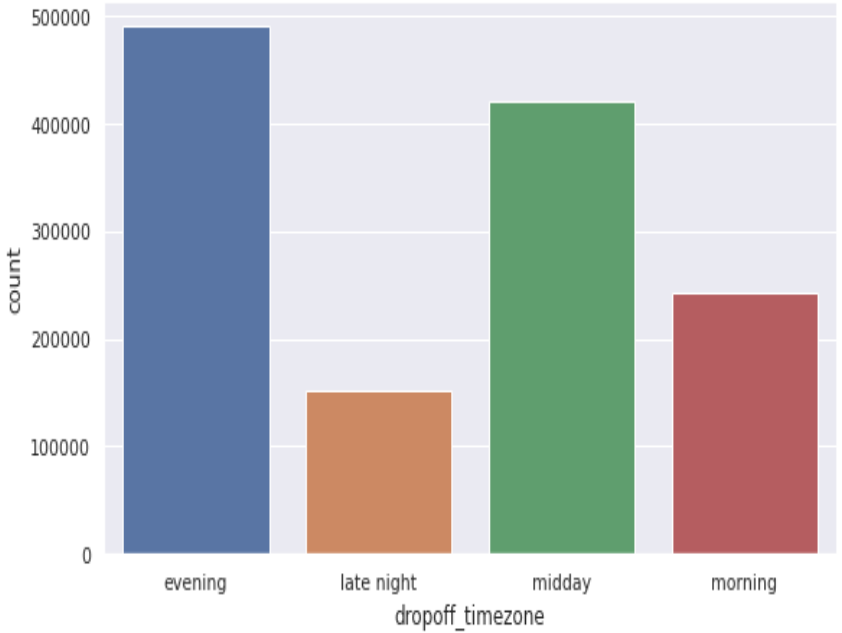

1e6
1.0
0.8
0.6
0.4
0.2
0.0
count
1
2
3
4
5
6
passenger_count


Number of pickups done on each day of week
200000
150000
100000
50000
0
count
Monday
Sunday
Tuesday
Wednesday
Saturday
Friday
Thursday
pickup_day

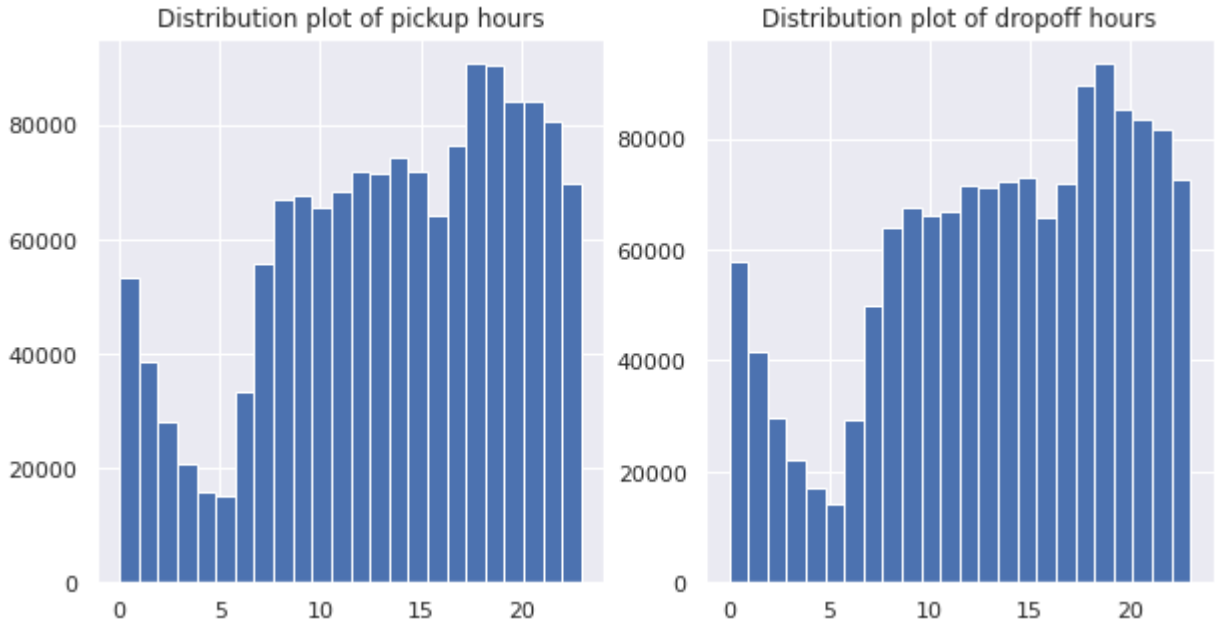

Number of dropoffs done on each day of week
200000
150000
100000
50000
0
count
Monday
Sunday
Tuesday
Wednesday
Saturday
Friday
Thursday
dropoff_day

EDA

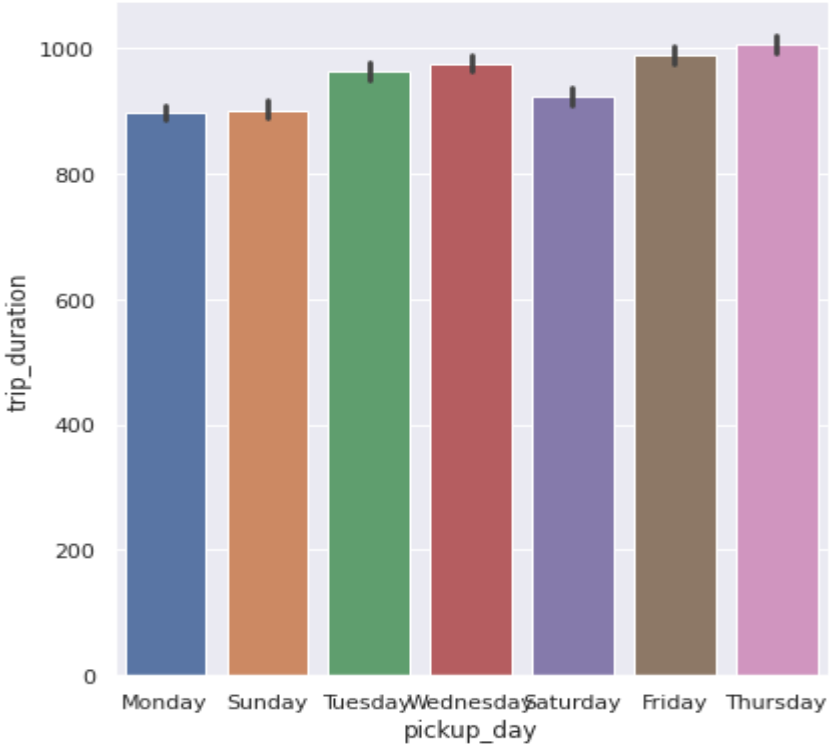

500000
400000
300000
200000
100000
0
count
evening
late night
midday
morning
pickup_timezone

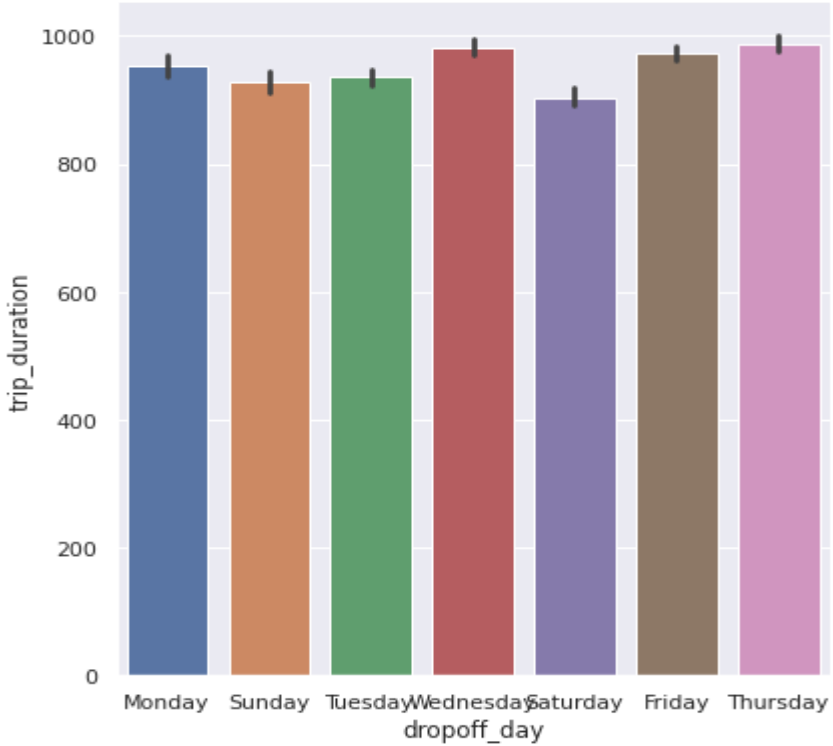

500000
400000
300000
200000
100000
0
count
evening
late night
midday
morning
dropoff_timezone

EDA

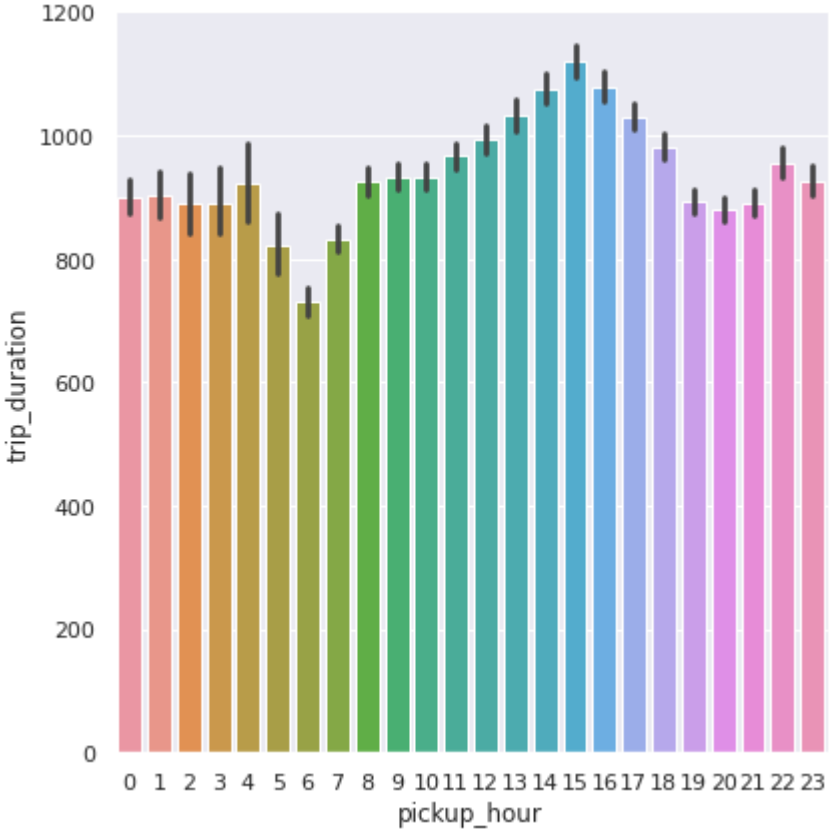

Distribution plot of pickup hours
80000
60000
40000
20000
0
0
5
10
15
20
Distribution plot of dropoff hours
80000
60000
40000
20000
0
0
5
10
15
20

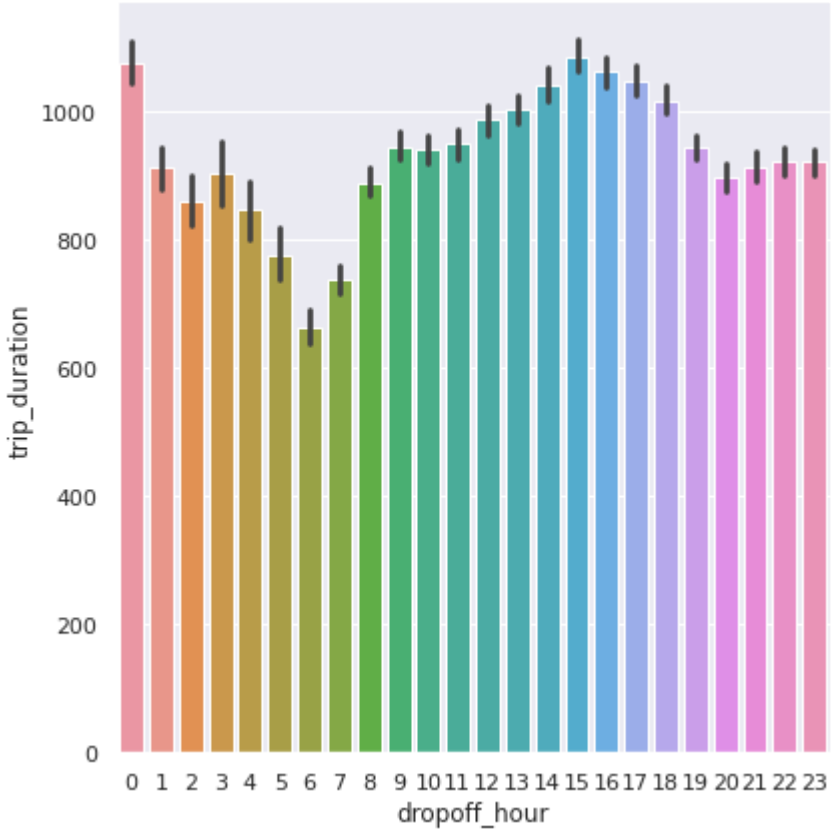
EDA


trip_duration
0
200
400
600
800
1000
Monday
Sunday
Tuesday
Wednesday
Saturday
Friday
Thursday
pickup_day

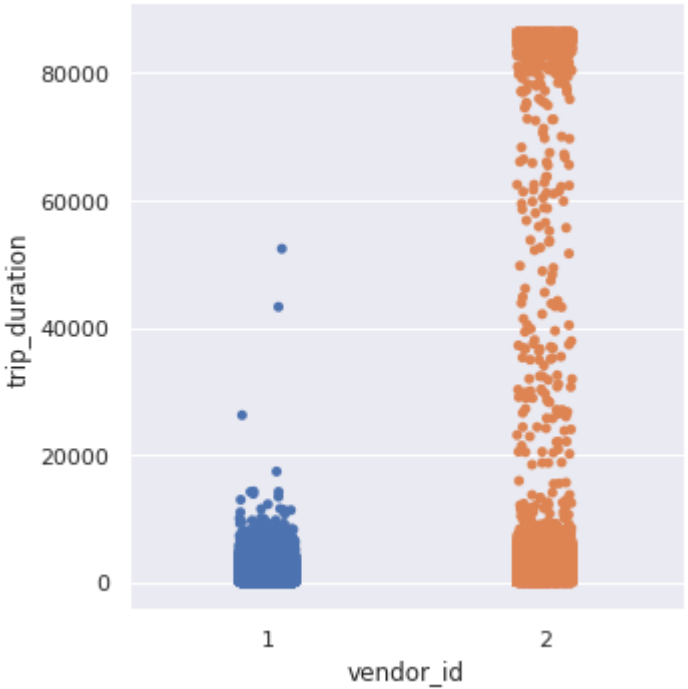

trip_duration
0
200
400
600
800
1000
Monday
Sunday
Tuesday
Wednesday
Saturday
Friday
Thursday
dropoff_day

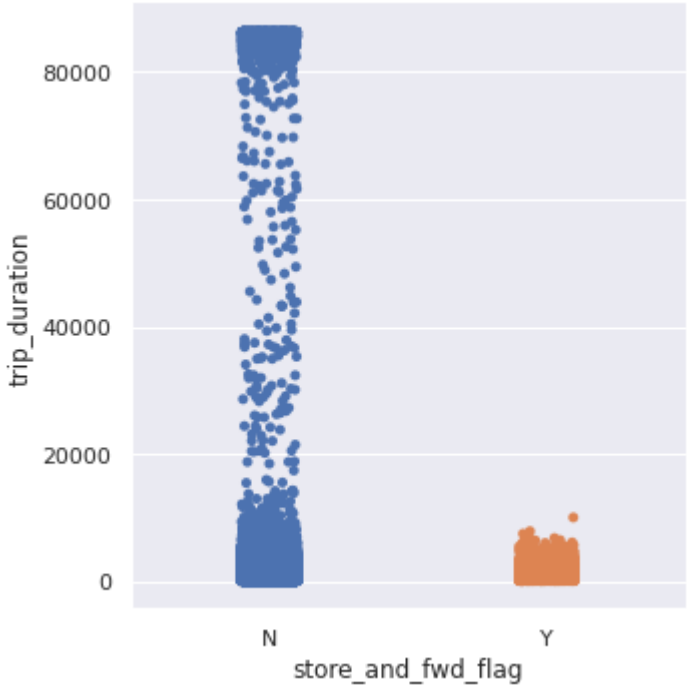
EDA


1200
1000
800
600
400
200
0
trip_duration
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
pickup_hour


1000
800
600
400
200
0
trip_duration
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
dropoff_hour

EDA

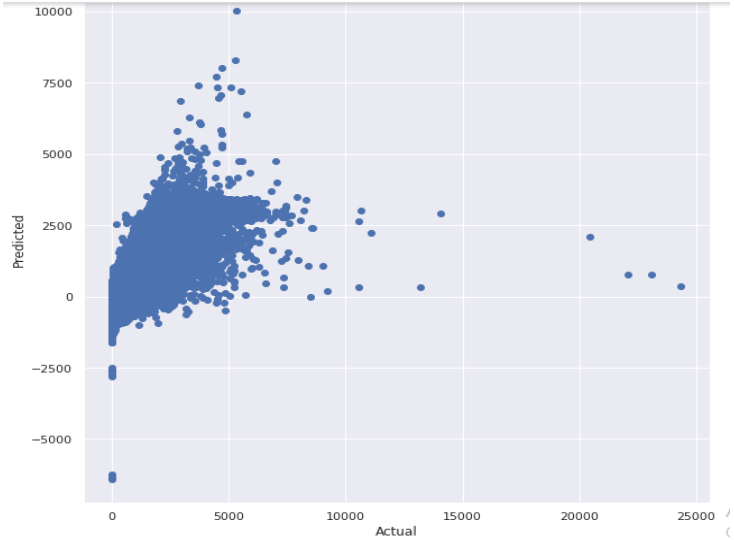

80000
60000
40000
20000
0
trip_duration
1
2
vendor_id


80000
60000
40000
20000
0
trip_duration
N
Y
store_and_fwd_flag

Model Fitting

Linear Regression

R squared	MSE	RMSE	MAE
0.6385635577343	159021.5646764	398.7708031022	260.090872112



	Actual value	Predicted value	Differance
1003422	1235	1025.256279	209.743721
1274202	538	550.871339	-12.871339
1333814	1402	1756.425057	-354.425057
804650	594	784.676423	-190.676423
248070	404	594.746798	-190.746798

Model fitting

Lasso regression

R squared	MSE	RMSE	MAE
0.63862949364	158992.554787	398.7387049027	260.0593561578

Ridge Regression

R squared	MSE	RMSE	MAE
0.638618077762	158997.5774431	398.7450030322	260.0536273471

Model Fitting

RandomForest

R squared	MSE	RMSE	MAE
0.913257059520	38164.3810822	195.3570604874	98.90383585599

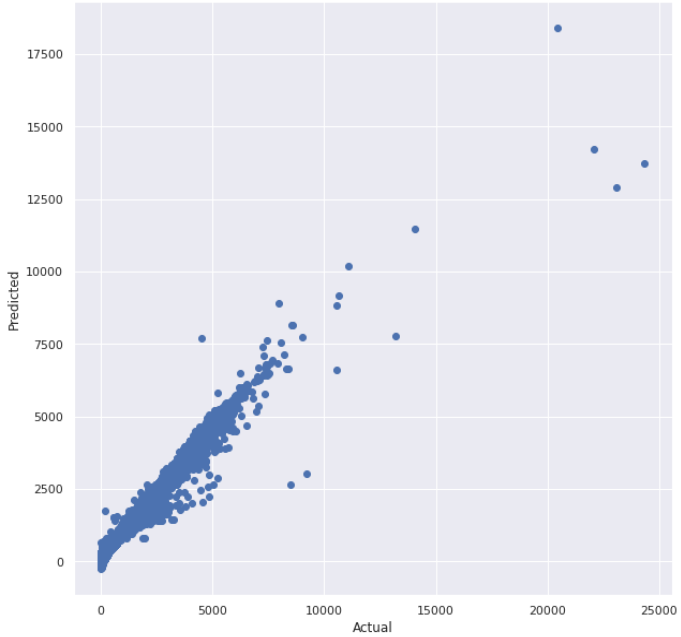
	Actual value	Predicted value	Differance
1003422	1235	1133.381255	101.618745
1274202	538	518.201017	19.798983
1333814	1402	1469.211898	-67.211898
804650	594	667.707571	-73.707571
248070	404	452.218303	-48.218303

...

Model Fitting

GradiantBoosting

R squared	MSE	RMSE	MAE
0.98927274172	4719.68290095	68.6999483330	29.9932568377



	Actual value	Predicted value	Differance
1003422	1235	1163.701614	71.298386
1274202	538	544.457972	-6.457972
1333814	1402	1339.074013	62.925987
804650	594	571.796290	22.203710
248070	404	435.136654	-31.136654

Model fitting

XGBoost

R squared	MSE	RMSE	MAE
0.996452702842	1577.77994761	39.72127827269	7.13538918665

	Actual value	Predicted value	Differance
181906	879	876.438721	2.561279
1446754	470	468.423096	1.576904
491700	284	284.980316	-0.980316
83520	417	419.794617	-2.794617
1108802	762	748.812500	13.187500

Conclusion

Trip Duration varies a lot ranging from few seconds to more than 20 hours

Most trips are taken on Friday , Saturday and Thursday.

The average duration of a trip is most on Thursday and Friday.

The average duration of trips started in between 14 hours and 17 hours is the largest
.
Vendor 2 mostly provides the longer trips. The flag was stored only for short duration trips and for long duration trips the flag was less stored.

When we compare the root mean squared error and mean absolute error of all the models, the XGBoost model has less root mean squared error and mean absolute error, ending with the accuracy of 99% . so, finally this model is best for predicting the trip duration count on daily basis.