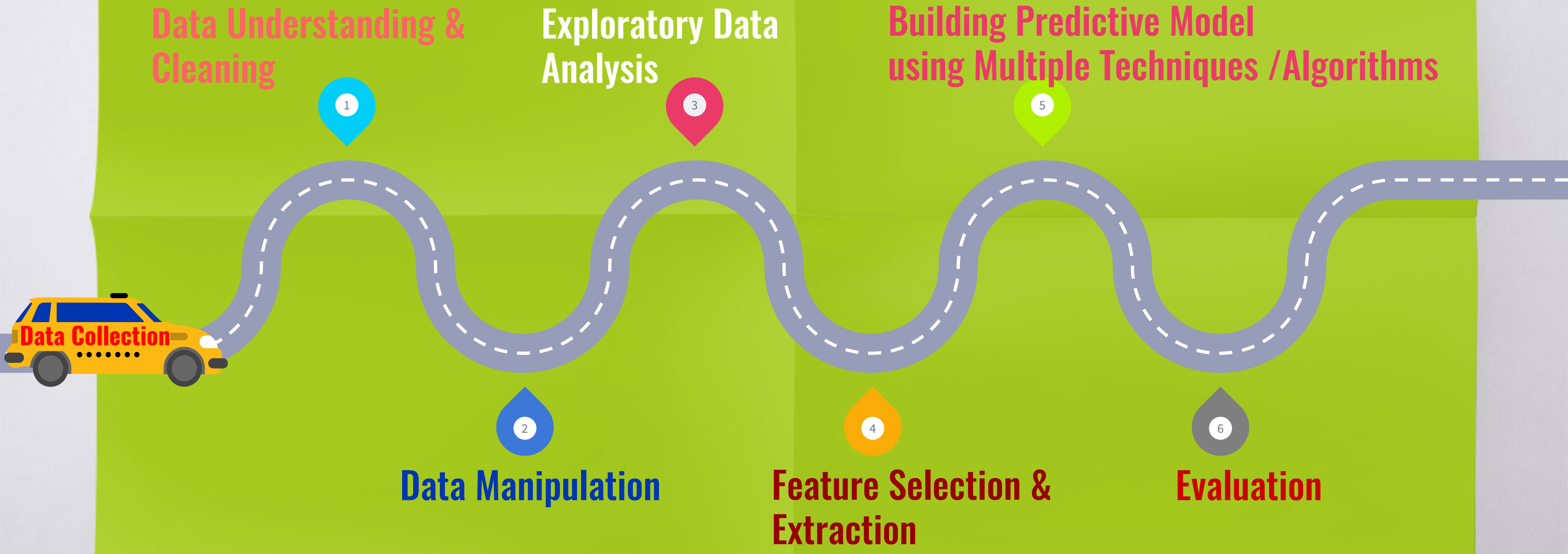
Capstone Project - 2

NYC Taxi Trip Duration Prediction

Team Members

- Mritunjay Sarkar
- Akash Choudary
- Kajal Mahajan
- Adarsh Gaurav
- Vivek Raikwar

Roadway


Data Understanding & Cleaning
Exploratory Data Analysis
Building Predictive Model using Multiple Techniques /Algorithms
1
3
5
Data Collection
2
4
6
Data Manipulation
Feature Selection & Extraction
Evaluation

DATA

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. Based on individual trip attributes, you should predict the duration of each trip in the test set.

This dataset contains around 1458644 observations distributed among 11 columns

Data Features

- 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.

Data Wrangling

Data wrangling is a process of cleaning and unifying messy and complex data sets for easy access analysis

01 → 02 → 03 → 04

Handling Missing Value

Quite surprising to see that NO NULL Values

Removing Duplicate Data

There are 1458644 unique ids means no duplicate data

Converting columns to proper dtype format

Converted timestamp into Datetime format

Adding And Removing Column

Assigned new columns such as weekday,month and pickup hour ETC

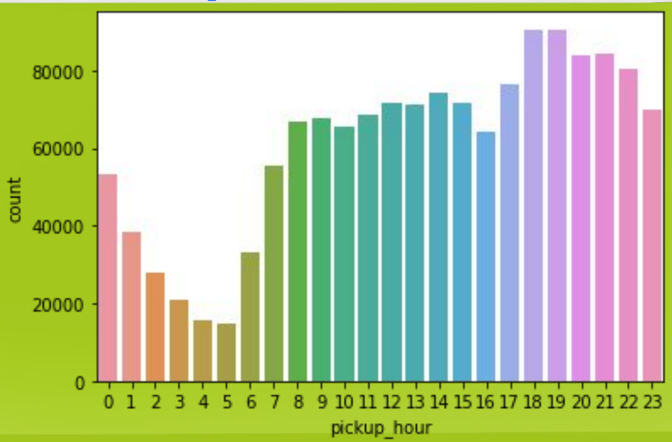
Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

1. Total Trips Per Hour
2. Total Trips Per Weekday
3. Trip Duration Per Hour
4. Trip Duration Per Weekdays
5. Trip Duration Per Month

6. Trip Duration Per Vendor
7. Distance Per Hour
8. Distance Per Weekdays
9. Average Speed Per Hour
10. Average Speed Per Weekday

1.Total Trips Per Hour

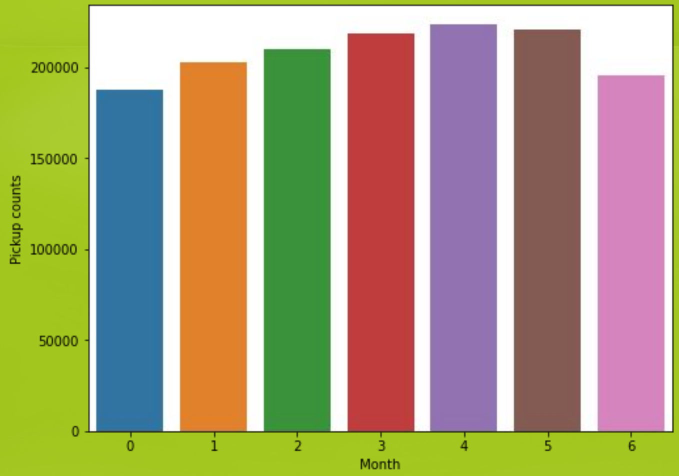


We have plotted a countplot on the distribution of pickup across 24 hour time scale

Conclusion

- Taxi pickup starts increasing at 6 and goes max at 18 &19

2.Total Trips Per Weekday

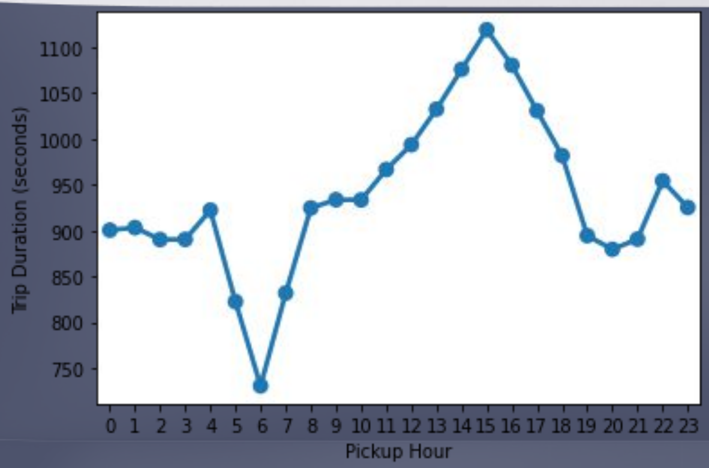


Conclusion

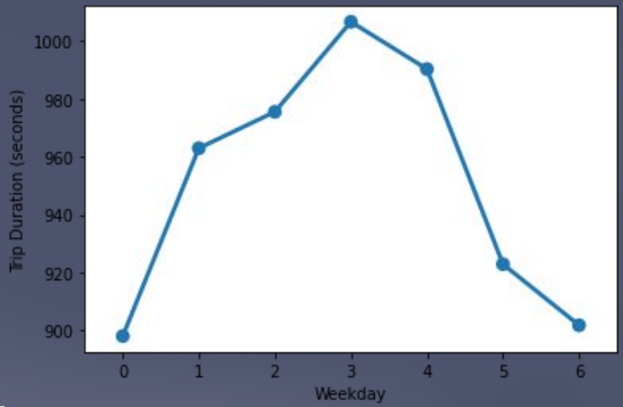
- Taxi trips are max at friday followed by saturday which may be due to weekend

3.Trip Duration Per Hour

- Average trip is lowest at 6 AM
- Average trip is highest at 3 PM
- Trip duration on an average is similar during early morning hours and late evening late hours

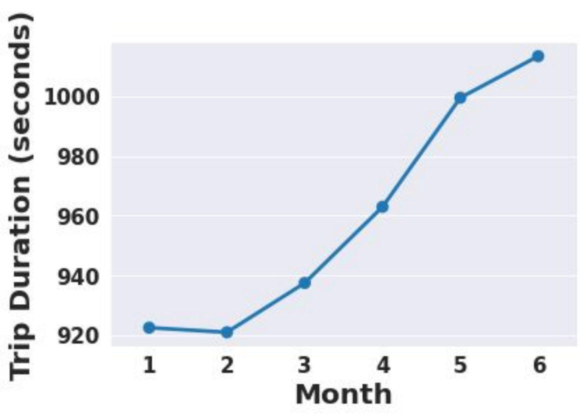


4.Trip Duration Per Weekday



- Trip duration on thursday is longest among all days

5.Total Duration Per Month



Conclusion

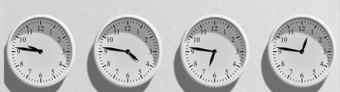
- It is lowest during february when winter starts declining
- There is an increasing trend in the average trip duration along with each subsequent month

6.Trip Duration Per Vendor

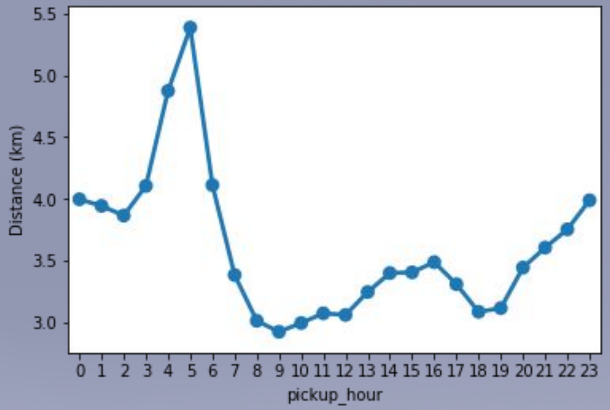
NYC Taxi Data has two vendors which are listed as 1 & 2 in the dataset

Conclusion

- Vendor 2 is higher than vendor 1 by approx 200 seconds



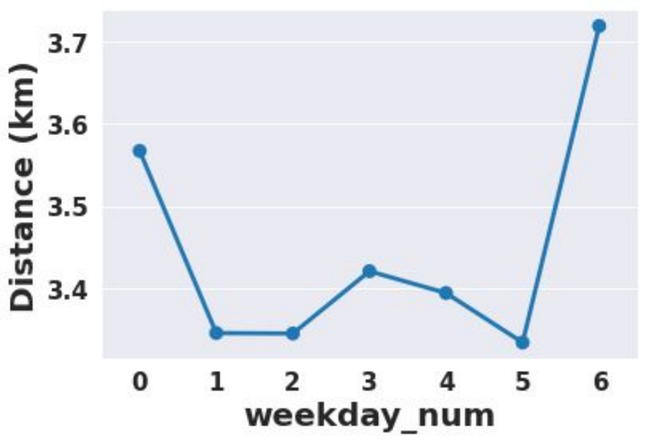
7.Distance Per Hour



Conclusion

- Trip distance is highest during morning hours
- It starts increasing gradually towards the late night hours and decrease during morning hours

8.Distance Per Weekday

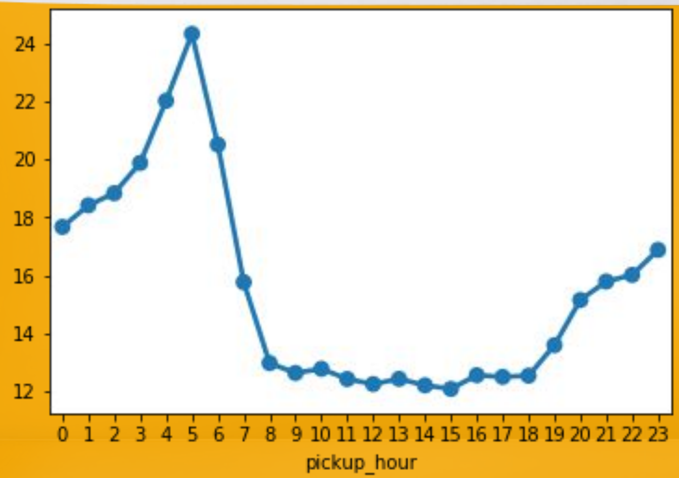


Conclusion

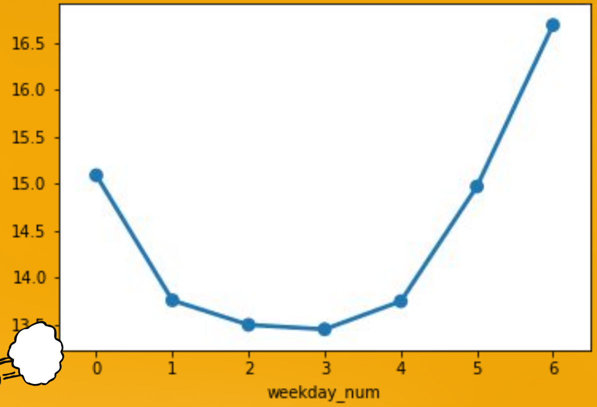
- Sunday is at the top may be due to outstation trips or night trips towards the airport

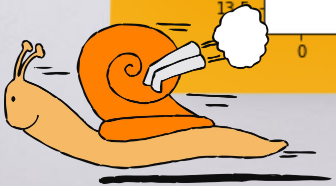
9.Average Speed Per Hour

- Average speed tends to increase after late evening
- Average speed is highest at 5 AM in the morning



10.Average Speed Per Weekday





- Average speed is higher on weekend which is obvious because of more rush
-

Models used

- ★ Linear Regression
- ★ Random Forest Regressor
- ★ XG Boost Regressor
- ★ Lasso Regression
- ★ Ridge Regression

Linear Regression

For Training Data

MSE : 27719319.392683808
RMSE : 5264.913996703442
MAE : 383.8830853770394
R2 : 0.014029143727798214
Adjusted R2 :
0.013951403100655798

For Testing Data

MSE : 24236180.7605149
RMSE : 4923.025569760419
MAE : 385.9898174185425
R2 : 0.018715915606665745
Adjusted R2 :
0.018638544516211275

Reason For Poor Result

None of the feature is linearly correlated with the target variable "46". That is why it is not a good model for the prediction of the trip duration. So let's move ahead and try the random forest regressor. We are not using decision tree regressor because the random forest will anyways consist of almost all its properties. Also, we will not use SVR because it takes too much time to train on this huge dataset even with the default settings. It seems to be not good with high dimensional dataset as well as for the huge instances

Random Forest Regressor

For Training Data

Model Score: 0.9627233035525813
MSE : 946912.4220271672
RMSE : 973.0942513586067
MAE : 4.800507577689049
R2 : 0.9627233035525813
Adjusted R2 : 0.9627217037325101

For Testing Data

MSE : 25468037.98471288
RMSE : 5046.58676579655
MAE : 16.376226808724468
R2 : 0.6344938930322279
Adjusted R2 : 0.6344782064488718

1.Even the variance score is approx 1 which is a good score.

2.RMSE score for the RF regressor of feature extraction group is still very bad along with the variance score.

3.RMSE score for the feature selection group is more or less same as the raw data score. Sometimes the RMSE score for the raw data is better and vice versa. It fluctuates on every iteration and this is quite weird! Let's see if we can improve this further with the most sought after algorigthm i.e. XGBoost!!

XG Boost

For Training Data

Model Score: 0.9848624019219396
MSE : 425574.3598247585
RMSE : 652.3606056658836
MAE : 62.32962075377837
R2 : 0.9848624019219396
Adjusted R2 : 0.9848612083710722

For Testing Data

Model Score: 0.9736287209680698
MSE : 651329.3099000739
RMSE : 807.049756768487
MAE : 64.07043887605909
R2 : 0.9736287209680698
Adjusted R2 : 0.97362664167763

1.As we have restricted ourselves on Feature Selection dataset only. The reason behind this is that if here the dimension increases, the time complexity also increases manyfold. So, better to check for the optimal features. Same thinking was behind the operation on Random Forest Regressor

2.It works exceptionally well for both Training and Test dataset.

Lasso Regression

Feature Selection

For Training Data

MSE : 24916356.007565424
RMSE : 4991.628592710542
MAE : 364.56058751162817
R2 : 0.019128466514946485
Adjusted R2 :
0.01891326872348409

For Testing Data

MSE : 69247125.72219084
RMSE : 8321.485788138489
MAE : 378.18086641395155
R2 : 0.019128466514946485
Adjusted R2 :
0.0059775988651147305

Feature Extraction

For Training Data

MSE : 25318917.498551015
RMSE : 5031.790685089257
MAE : 562.8003408381326
R2 : 0.003281000422185043
Adjusted R2 :
0.0030623257842784524

For Testing Data

MSE : 69604532.89733748
RMSE : 8342.93311162853
MAE : 577.3979610797758
R2 : 0.003281000422185043
Adjusted R2 :
0.0008471225497888035

For Linear, Lasso, and Ridge regressor the model does not works fine.

1.As correlation is a linear model, thus we have shown that the all the other featues (in Linear regression) have a very poor correlation with feature no 46. Here we are using correlation, as because it cathches the linear dependency

2.For the other two regression, we have to catch the non-linear dependency among features. As Mutual Information is an excellent tool to explore non linear dependcy, thus someone can look forward to implement it in order to explore the poor result.

Ridge Regression

For Training Data

MSE : 24916355.148822818
RMSE : 4991.628506692262
MAE : 364.6812006163739
R2 : 0.019128500320699 57
Adjusted R2 :
0.019086403835174126

For Testing Data

MSE : 69247063.40472905
RMSE : 8321.482043766546
MAE : 378.296536367151
R2 : 0.019128500320699 57
Adjusted R2 :
0.006153876908841394

1.For RIDGE regression, the time complexity is very less. So it enables us to explore it for optimal featues, as well as on the extended features too.

2.However, it seems that for both dataset the result is very poor (as like Lasso). So for this large data RIDGE regression prooves to be futile.

Conclusion

We implemented 5 machine learning algorithms: Linear Regression,Lasso, Ridge, Random Forest and XGBoost. We did hyperparameter tuning to improve our model performance. Now in order to avoid unnecessary time complexity we have considered Feature Selected dataset for Random Forest, and XGBoost.

While comparing the results we have found that for XGBoost the results are very well, followed by Random Forest. However for the remaining three ML algorithms the results are very poor. The possible reason behind this is also partially explored by means of correlation. In future one can elaborate the non linear dependency result using Mutual Information or by any other tool.


THANK
YOU