Dynamic Pricing Prediction For Cabs Using IBM Watson

Karale Prathmesh Sudhir

Vellore Institute of Technology, Bhopal

Abstract

With the day to day data produced by the customers uber produces a large amount on the daily basis the statistics say that over a fifteen million uber trips are placed and completed in a day. This produces a large data set of rides booked by the consumers and the price for each ride is recorded in the uber data set. With around sixty-five countries (i.e., ten-thousand cities) where uber is a major source of transportation on a daily basis for urban population, uber has great demand and highly used in metropolitan cities. With the rise of carpooling culture coming into picture where the sharing of cabs to travel from a source to destination and uber pool is a great feature which provides sharing of cabs to travel from one place to another with less cost. With this data produced by cab services we use this dataset which includes the type of cab, source, destination, total ride time and cost of ride to predict the price of the ride prior to the start of the ride the consumer will be able to know the price of the ride before taking the ride. There are several aspects involved in predicting the price of the cab ride factors such as- surge multiplier, weather and availability of the cab plays a very important role in creating a price prediction model. We will use techniques of linear regression and Random Forest Regression combined with the machine learning algorithms to predict the price of uber ride. Uber has its own model called ‘Uber dynamic pricing model’ to predict the estimated price for the ride. But in this paper, we will apply the weather data as an additional dataset in order to get more precise prediction that is based on the weather report for the day as well as following week. With this we will be able to get better price prediction model that can be used to predict the price of the consumer’s ride.

1. INTRODUCTION

1.1 Overview

As explained in the abstract the workings of uber dynamic model and about the price prediction model to predict the price of the ride from given source to destination we use the data such as the distance between the source and destination. As weather plays a very important role in deciding the surge in the price of the cab, we take the weather report given for the respective day and by using this weather data we are able to predict the actual price for the given ride at a certain period of time. And with the help of linear and logistic regression we are able to visualize the data into pictures or graphs for better understanding and visualizing the estimation of the pricing with various factors.

The traffic also plays a major role in calculating the surge of price of the ride with increase of the traffic the availability of the cabs becomes limited and when the demand for the cabs start to increase the service provider will not be able to provide the cabs this causes the surge in the price during the peak hours. The peak hours are usually calculated as the time when there is a large number of requests for the cabs and the price is increased during the peak hour. With the help of the driver’s data set as well as the customers dataset we are able to calculate the peak hour for each day.

1.2 Purpose

With the growth of combining the concept of ‘Big Data and Machine-Learning’ and introducing machine learning model for data analysis and the importance of data produced by the cabs on daily basis and how this data can be used by the machine learning to tell the consumer about the exact price of their ride before starting the ride. This provides the consumer to make better choice of cab based on the price predicted by the Machine-Learning model. The proposed system uses the cab dataset and weather dataset to make predictions for each ride booked by the customer.

2. LITERATURE SURVEY

2.1 Existing problem

The author tells us the popularity of uber in the recent years and about the urban citizens who are benefited by the uber. Later the author compares the difference between the competitive taxis and uber and defines new way of calling and also the new way of paying for cabs, the author also tells us about the importance of data produced by the cabs daily and also about the visualization and analysis of data. After that the author tells how the different time and different environments will have an effect on passengers to make different choices.

2.2 Proposed solution

The proposed system helps us for better predictions of cab fare from source the destination using the methods of Linear and Logistic regression. The model uses machine learning technique (Supervised learning) which helps to train the machine with labelled data that is already tagged with some predefined class. Then we test our model with some new unknown set of data and predict the price for them.

3. THEORETICAL ANALYSIS

3.1 Project Flow

1. Install Required Libraries.
2. Data Collection.
   - Collect the dataset or Create the dataset
3. Data Pre- processing.
   - Import the Libraries.
   - Importing the dataset.
   - Understanding Data Type and Summary of features.
   - Take care of missing data & create columns.
   - Data Visualization.
   - Drop the column from dataframe ,merge the dataframes.
   - Observing Target,Numerical and Categorical Columns
   - Label Encoding & Splitting the Dataset into Dependent and Independent variables
   - Splitting Data into Train and Test.
4. Model Building
   - Training and testing the model
   - Evaluation of Model
   - Saving the Model
5. Application Building
   - Create an HTML file
   - Build a Python Code
6. Final UI
   - Dashboard Of the flask app.

3.2 Hardware / Software designin

All the requirements of the project are listed below:

- Google Colaboratory
- Anaconda 3
- Sklearn (Machine-Learning model)
- Flask
- IBM Watson Studio

Python libraries:

- Pandas
- NumPy
- Matplotlib.pyplot
- Seaborn

4. EXPERIMENTAL INVESTIGATIONS

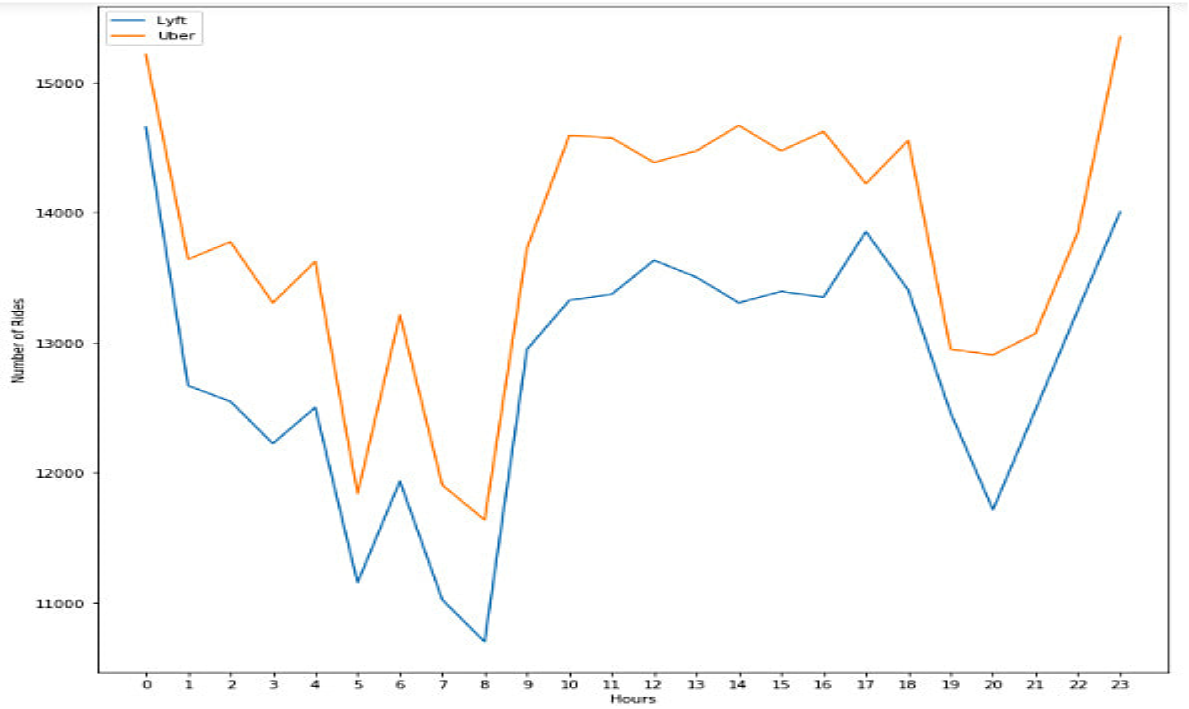
The code for the entire model is done in python the project contains of two parts. First part contains cleaning of data as well as data acquisition and the second part contains exploratory data analysis and implementing machine-learning algorithm. In the first part we will load the data into the google colaboratory using the ‘wget’ command both the cab data and weather data is stored in the dropbox so that the data can be accessed from anywhere and at anytime once the data gets loaded into the program the size of the data is checked. Since the dataset contains large number of data i.e., around six lakh data we use memory reduce function to reduce the size of the data.

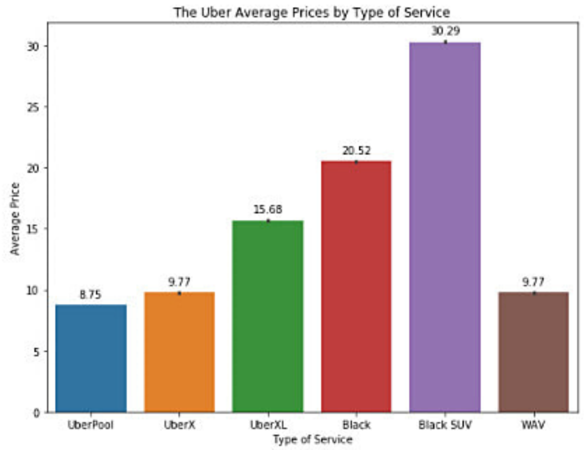
The memory reduce function helps us to reduce the size of the total dataset by compressing the original dataset, since the original dataset is compressed the operations carried on the dataset can happen much faster as the size of the dataset is now small. Next, we have to do the data cleaning for the compressed dataset. The data cleaning helps us to detect and correct the data which is inaccurate or null within the data set we can choose between replacing, modifying or cleaning the inaccurate or corrupt data in our dataset.

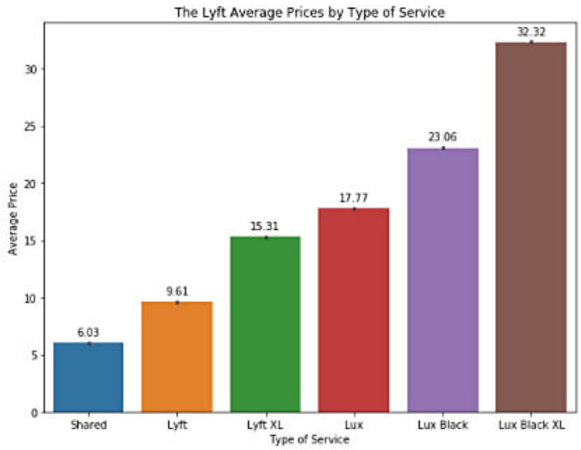
For example, in some of the entries we have distance travelled in the ride is zero this leads to confusion for machine-learning model in order to predict the accurate price for the cab ride so all such type of inaccuracies in the dataset are detected and has to be removed before going to the next process which is exploratory data analysis.

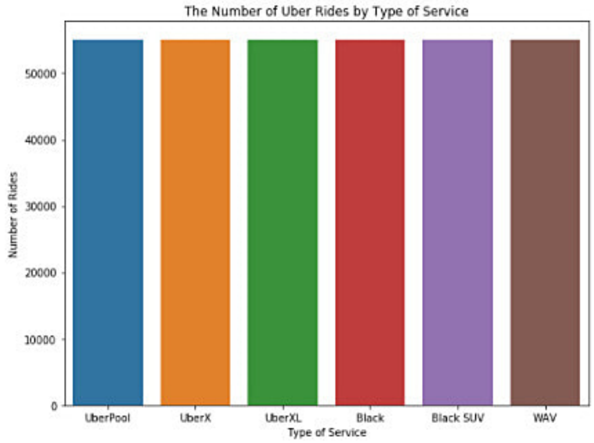
4.1 Graphs

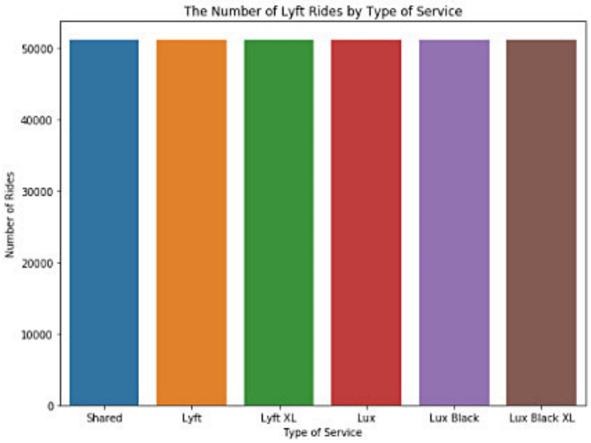
The ride distribution in one day

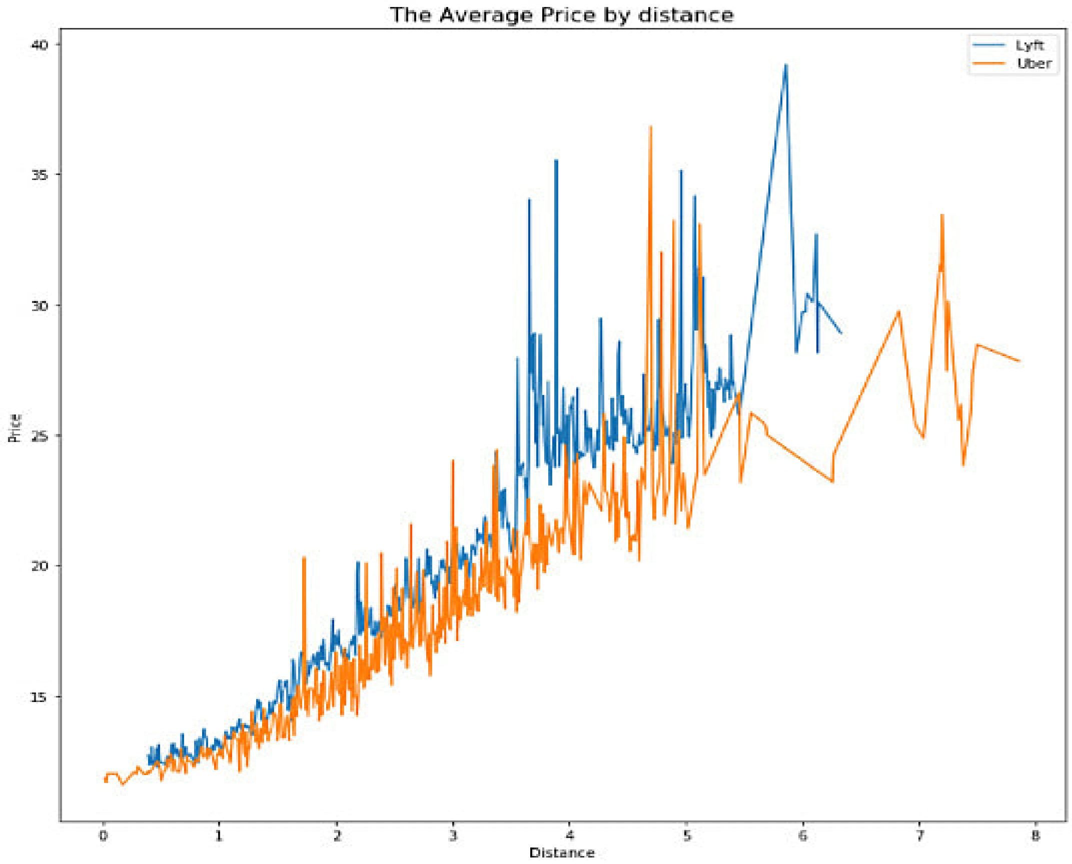



The Uber Average Prices by Type of Service
Average Price
0
5
10
15
20
25
30
8.75
9.77
15.68
20.52
30.29
9.77
UberPool
UberX
UberXL
Black
Black SUV
WAV
Type of Service

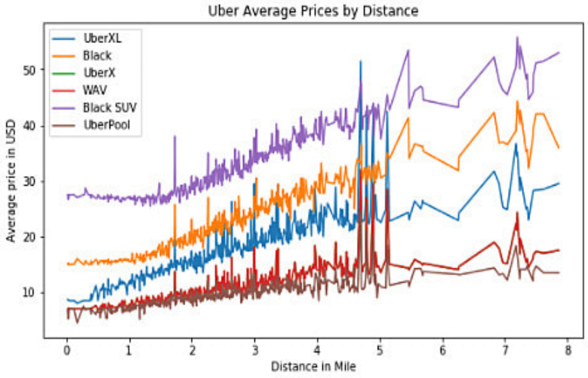

The Lyft Average Prices by Type of Service
Average Price
0
5
10
15
20
25
30
6.03
9.61
15.31
17.77
23.06
32.32
Shared
Lyft
Lyft XL
Lux
Lux Black
Lux Black XL
Type of Service


The Number of Uber Rides by Type of Service
Number of Rides
0
10000
20000
30000
40000
50000
UberPool
UberX
UberXL
Black
Black SUV
WAV
Type of Service

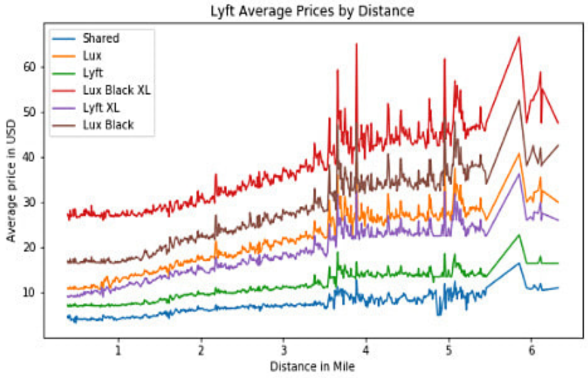

The Number of Lyft Rides by Type of Service
Number of Rides
0
10000
20000
30000
40000
50000
Shared
Lyft
Lyft XL
Lux
Lux Black
Lux Black XL
Type of Service



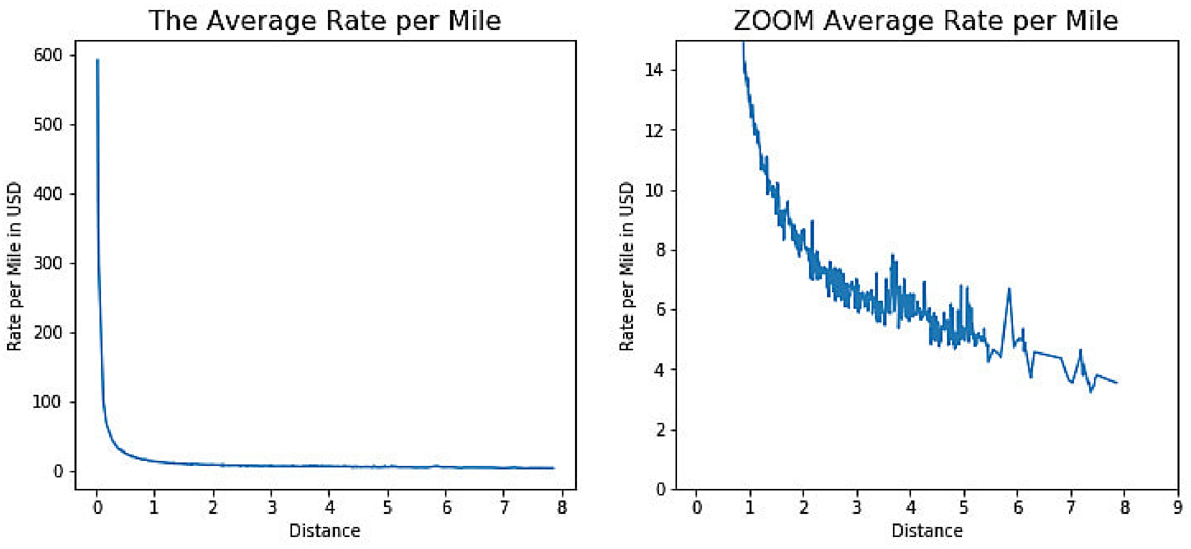
The Average Price by distance

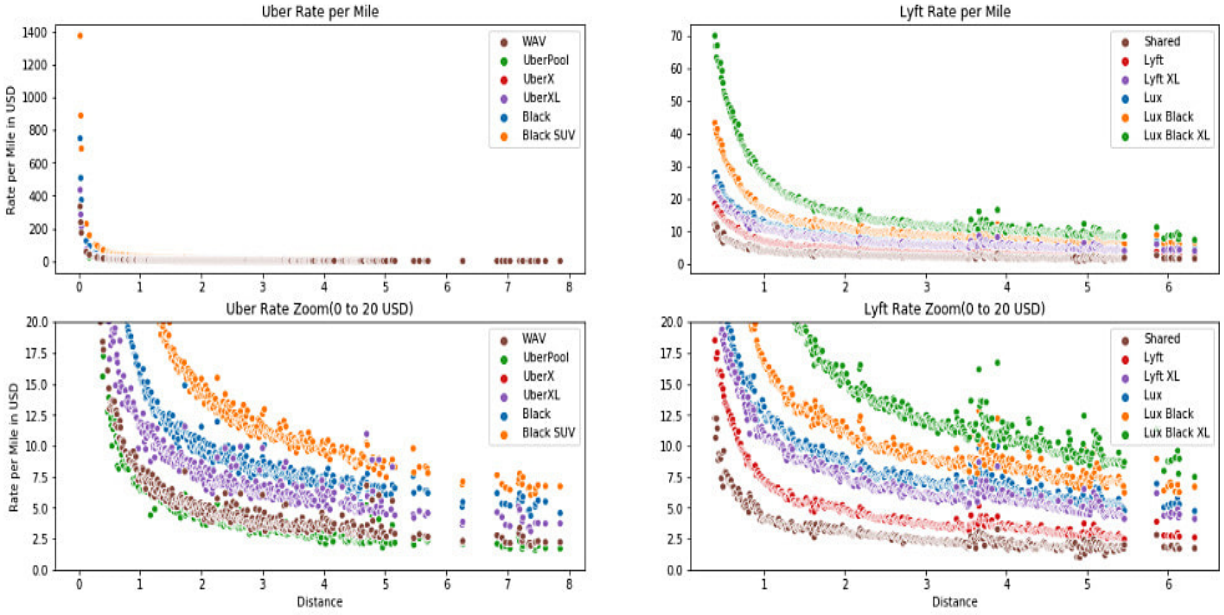


Uber Average Prices by Distance



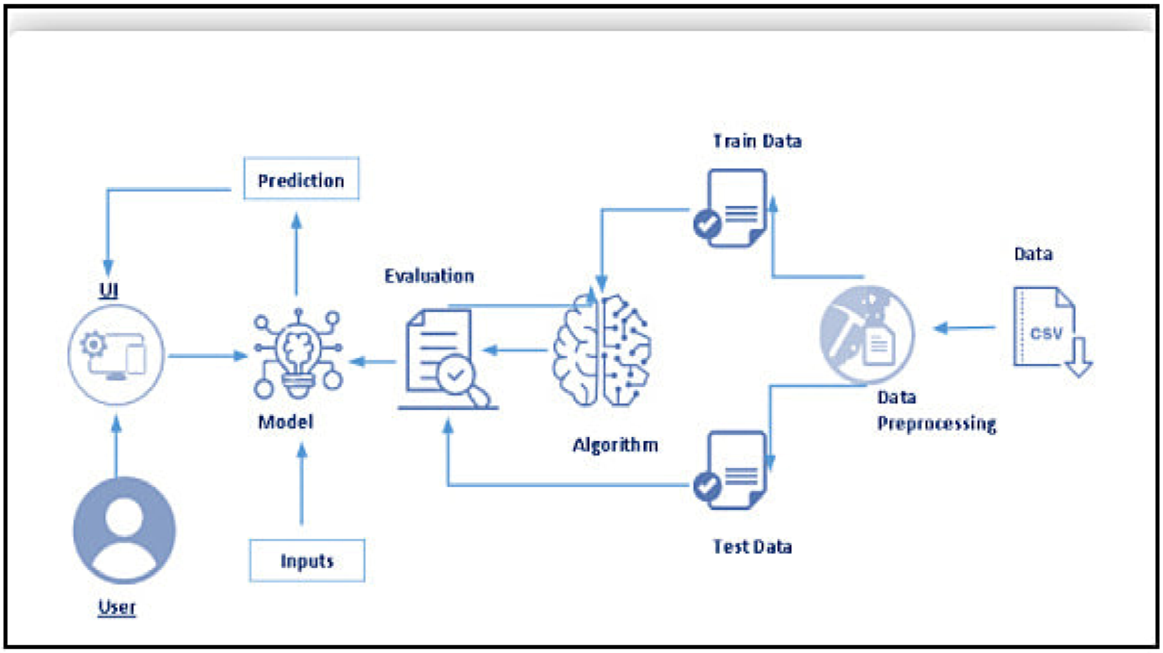
Lyft Average Prices by Distance


The Average Rate per Mile
Rate per Mile in USD
Distance
0
100
200
300
400
500
600
0
1
2
3
4
5
6
7
8
ZOOM Average Rate per Mile
Rate per Mile in USD
Distance
0
2
4
6
8
10
12
14
0
1
2
3
4
5
6
7
8
9

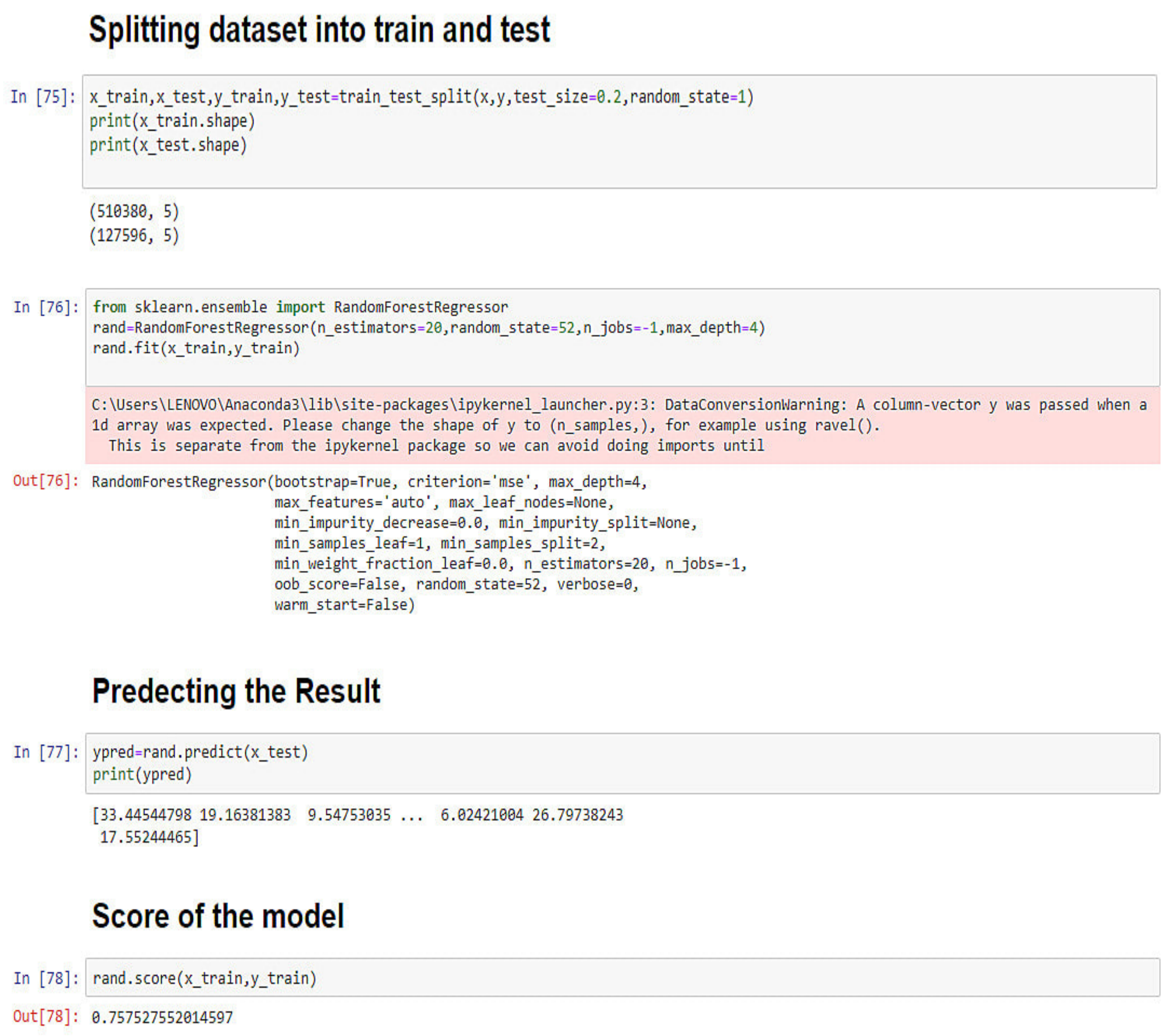

Uber Rate per Mile
Rate per Mile in USD
0
200
400
600
800
1000
1200
1400
0
1
2
3
4
5
6
7
8
WAV
UberPool
UberX
UberXL
Black
Black SUV
Lyft Rate per Mile
0
10
20
30
40
50
60
70
1
2
3
4
5
6
Shared
Lyft
Lyft XL
Lux
Lux Black
Lux Black XL
Uber Rate Zoom(0 to 20 USD)
Rate per Mile in USD
Distance
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
0
1
2
3
4
5
6
7
8
WAV
UberPool
UberX
UberXL
Black
Black SUV
Lyft Rate Zoom(0 to 20 USD)
Distance
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
1
2
3
4
5
6
Shared
Lyft
Lyft XL
Lux
Lux Black
Lux Black XL

5. FLOWCHART

Technical Architecture:



6. RESULT



Splitting dataset into train and test

```
In [75]: x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=1)
         print(x_train.shape)
         print(x_test.shape)

         (510380, 5)
         (127596, 5)
```

```
In [76]: from sklearn.ensemble import RandomForestRegressor
         rand=RandomForestRegressor(n_estimators=20,random_state=52,n_jobs=-1,max_depth=4)
         rand.fit(x_train,y_train)

         C:\Users\LENOVO\Anaconda3\lib\site-packages\ipykernel_launcher.py:3: DataConversionWarning: A column-vector y was passed when a
         1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().
           This is separate from the ipykernel package so we can avoid doing imports until

Out[76]: RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=4,
                               max_features='auto', max_leaf_nodes=None,
                               min_impurity_decrease=0.0, min_impurity_split=None,
                               min_samples_leaf=1, min_samples_split=2,
                               min_weight_fraction_leaf=0.0, n_estimators=20, n_jobs=-1,
                               oob_score=False, random_state=52, verbose=0,
                               warm_start=False)
```

Predecting the Result

```
In [77]: ypred=rand.predict(x_test)
         print(ypred)

         [33.44544798 19.16381383  9.54753035 ...  6.02421004 26.79738243
          17.55244465]
```

Score of the model

```
In [78]: rand.score(x_train,y_train)

Out[78]: 0.757527552014597
```

7. ADVANTAGES & DISADVANTAGES

This project can potentially be profitable in the real market.
If a taxi driver could know in advance (and with precision) which boroughs or areas are going to have the biggest demand, he could optimise his workday by driving only around those areas. He could choose whether to earn more money in the same time or save that time for his family/personal life. Either way, it will improve his life.

8. APPLICATIONS

Solving an existing problem.
There has been a lot of debate in the past regarding how Uber is literally eating the traditional street hail taxi market. Taxi drivers are afraid that they cannot compete with the kind of on-demand fare-adjusted service Uber provides, based in cutting-edge technology. I think that traditional taxi drivers could also make use of advance technology like this machine learning app in order to improve their service and profitability.

9. CONCLUSION

This project gives us basic understanding of how we can use machine learning in order to predict the cab fare from given source to destination before starting the cab ride. The model created is able to give us the predictions which are not exactly equal to the actual the price fluctuation is around the difference of ten to twenty rupees compared to the actual price. Since the model is good but not the best, we can improve the predictions of the model by using the Fine-tuning technique. If fine tuning is applied to the existing model, we are able to get higher accuracy than the proposed model.

10. BIBLIOGRAPHY

- **https://www.analyticsvidhya.com/blog/2021/06/uber-and-lyft-cab-prices-data-analysis-and-visualization/**
- **https://github.com/NikhilKumarMutyala/Price-Prediction-in-Ride-Hailing-Services-based-on-Weather-Conditions**
- **https://www.kaggle.com/ravi72munde/uber-lyft-cab-prices**

APPENDIX

(Source Code to the Project has been given in terms of Google drive link.)

➤ **https://drive.google.com/drive/folders/114mC4cJK816mlXveivUHENRfBJ8NmYpf?usp=sharing**