International Journal of Interdisciplinary Innovative Research &Development (IJIIRD)
ISSN: 2456-236X
Vol. **05** Issue **01** | **2020**

Data Analysis of Uber and Lyft Cab Services

Shashank H.
Department of Computer Application, Jain University, Bangalore, India

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 logistic 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.

***Keywords**: Linear and Logistic regression, Machine-Learning, Pricing Model*

1. INTRODUCTION

As explained in the abstract the workings of uber dynamic model and about the price prediction model to predict the price of the ride form 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.

1.1 Motivation

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 REVIEW

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.

3. PROPOSED SYSTEM

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.

3.1 Requirements

All the requirements of the project are listed below:

- Google Colaboratory
- Anaconda 3
- Sklearn (Machine-Learning model)

Python libraries:

- Pandas
- NumPy
- Matplotlib.pyplot
- Seaborn

3.2 Overview of Architecture

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.

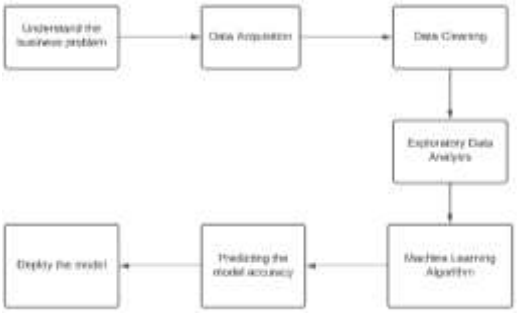


Fig 1.1: Workflow of the model

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.

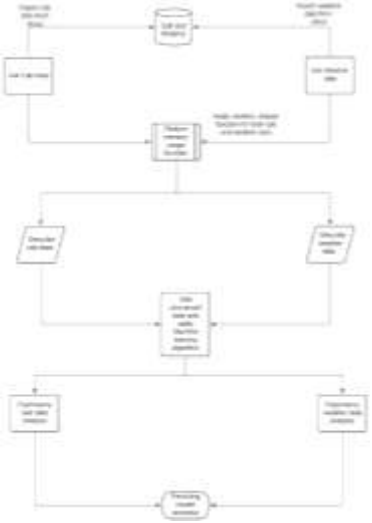


Fig.1.2: Overall workflow of the model

Once the data is cleaned, we have to label the data which will be given to the machine learning model since we are using the technique of supervised learning. The supervised learning process works only with labelled data hence it is important for us to label the data within the dataset accordingly. After the labelling of data, we can use that data and give it to the machine learning model which will use the label data and helps to get derive the graphs for the difference between the actual price and the prediction made by the machine learning model. The graph created by the model is given below.

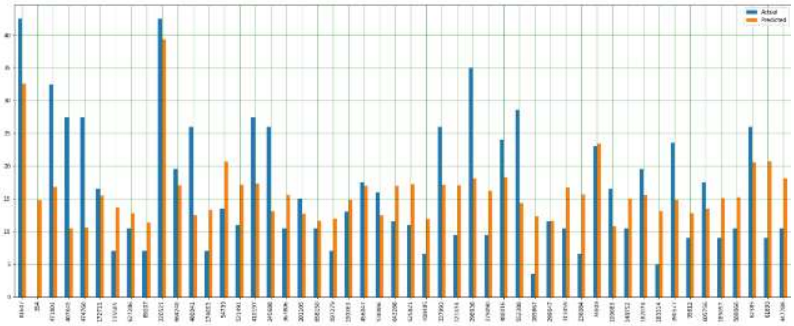


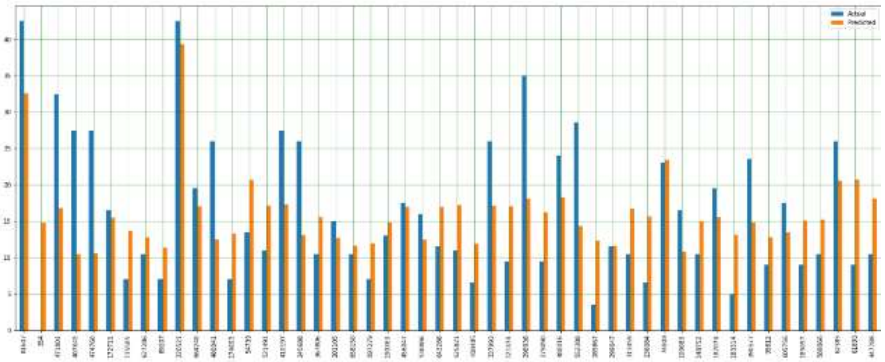
Fig.1.3: Actual vs Predicted price

4. RESULTS

- The data set used for the model:

cab_type	distance	time_stamp	destination	source	price	surge_multiplier	id	product_id	name	datetime	temp	location	clouds	pressure	rain	humidity	wind
0	6276	6276	6276	6276	6276	6276	6276	6276	6276	6276	6276	6276	6276	6276	6276	6276	6276
Lyft	307408	307408	307408	307408	307408	307408	307408	307408	307408	307408	307408	307408	307408	307408	307408	307408	307408
Uber	385663	385663	385663	385663	385663	385663	385663	385663	385663	385663	385663	385663	385663	385663	385663	385663	385663

- The graph of Actual vs Predicted price:



- The chart showing actual vs predicted price:

	Actual	Predicted
107925	34.0	19.210929
584935	14.0	12.745144
452114	9.0	17.504089
9502	11.5	17.707680
9457	20.5	16.587928
465251	0.0	16.231644
692438	16.0	23.459132
429986	8.5	15.951706
343534	0.0	23.459132
461968	13.5	12.414309
167615	33.0	17.071458
436891	10.5	12.287064
544736	3.0	17.275049
61224	0.0	11.141864
305639	8.5	13.126878
484553	21.5	15.162790
407625	9.5	16.078951
335592	7.0	11.141864
155434	7.5	12.796042
70524	5.5	10.302050
538489	7.5	10.709232
579511	13.5	15.519075
5421	9.0	11.828984

5. 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.

6. REFERENCES

[1] J. Guo, "Analysis and comparison of Uber, Taxi and Uber request via Transit,," *IIJRD*, vol. 4, no. 2, pp. 60-62, 2015.

[2] N. G. G. K. Uyanik, "A study on multiple linear regression analysis," *Procedia- Social and Behavioral Sciences*, vol. 106, pp. 234-240, 2013.

[3] Y. J. Y. Zhang, "A data-driven quantitative assessment model for taxi industry: the scope of business ecosystem's health," *Eur. Transp. Res.*, vol. 9, pp. 1-23, 2017.

[4] U. Patel, "NYC Taxi Trip and Fare Data Analytics using BigData," *Department of Computer Science and Engineering University of Bridgeport, USA*, 2018.

[5] J. Chao, "Modeling and Analysis of Uber's Rider Pricing," *Advances in Economics, Business and Management Research*, vol. 109, pp. 639-711, 2019.