Final Project:

Uber and Lyft Price Prediction

By

Class Section 3 & Group 2

Mihir Koyande

Lynne Chen

Harsh Gandhi

Dhruv Shah

Sahil Sahil

In partial fulfillment of the requirements for

BAN 4550: Analytics Programming

December 5, 2022

Supervised by:

Prof. Theyab Alhwiti

Abstract:

The challenge's significance to us as college students, especially in Boston, motivated us to take it on. To move around the enormous city, we frequently use ridesharing services like Uber and Lyft and knowing how their pricing structures operate and change depending on the situation helps us decide which ridesharing service to employ. The analysis conducted by our team compared 750,000 rideshares between Uber and Lyft in Boston, Massachusetts. Furthermore, the knowledge we have gained from this dataset will be extremely useful in solving real-world problems using data science and machine learning methods. In our analysis, we forecasted and analyzed the cost of ridesharing services from Uber and Lyft using several factors, including distance, time of day, surge multiplier (demand-based pricing), etc. We can see that Lyft has a lower rate for distance than Uber, which suggests that Lyft generally costs less for every extra mile and that long-distance rides will be less expensive on Lyft than they would be on Uber.

1. Introduction

For a sample set of 750,000 rideshares, our team's analysis focused on contrasting Uber and Lyft rides in Boston, Massachusetts. We were motivated to take on this task because it was pertinent to us as Boston-area college students. We frequently use Uber and Lyft rideshares to move around the enormous city and knowing how these pricing models operate and change depending on the situation helps us decide which ridesharing service to employ. We therefore decided to take a road trip to New York in the fall after being inspired by a real-life incident. The first thing we did was make a taxi reservation, and Uber was the only choice that occurred to us. However, when we looked at the overall cost of the vacation, we were astonished because it was much higher than we had anticipated. The next choice was Lyft, which offered some promise because it was slightly less expensive than uber services. Nevertheless, we chose Uber because it is a well-known and internationally renowned corporation. After this occurrence, we made the decision to research this subject and determine how much rides cost on Uber and Lyft. Who doesn't like to save money? The information from this study will assist customers decide whether to use Uber or Lyft in certain situations. Additionally, this will assist the businesses in comprehending and improving both their pricing strategy and that of their competitors. Additionally, the knowledge we have gained from this dataset will be very useful for employing data science and machine learning techniques to solve real-world problems. The purpose of this study is to develop a model for estimating fares for Uber and Lyft in the Greater Boston area. For the two businesses, we develop several linear regression models, and we contrast the variations in their pricing approaches. Based on a range of factors, including distance, time of day, surge multiplier (demand-based pricing), and others, we forecasted and compared the cost of Uber and Lyft rideshares. We obtained our data from Kaggle and developed a price prediction model using this extensive dataset. The main change it will bring about is that it will raise awareness and enable individuals to save time, money, and both. Additionally, on a business level, it will assist the organization in identifying areas where they fall short or where the rival organization excels.

2. Method and Analysis:

2.1 Data Source:

The data used to develop our medium article came from: Kaggle dataset. The contributors of the dataset queried both Lyft and Uber prices in the Boston Area. The queries were done on the apps every 5 minutes for 22 days from late November through mid-December in 2018. The weather data was queried from the Dark Sky API every hour.

2.2 Data Preparation:

The contributors of the dataset queried both Lyft and Uber prices in the Boston Area. The queries were done on the apps every 5 minutes for 22 days from late November through mid-December in 2018. The weather data was queried from the Dark Sky API every hour.

2.3 Data Description:

Dataset including data source and variable descriptions:

No.	Feature	Description	Value	Scale
1.	Cab_Type	Whether the cab is Uber or Lyft	Boolean; 0 = Uber, 1 = Lyft	[0, 1]
2.	Customer_id	Unique identifier for each column	Character, Number	[0, 693071]
3.	Gender	Whether the customer is male or female	Boolean; 0 = Male, 1 = Female	[0, 1]
4.	Hour	Time of the day customer booked a ride	Integer	[0, 23]

5.	Day	Day of the month customer booked a ride	Integer	[1, 30]
6.	Month	Month of the year customer booked a ride	Integer	[11, 12]
7.	DateTime	Exact time when the customer booked a ride	Date Value	[25 Nov – 18 Dec]
8.	Time Zone	Time zone when the customer booked a ride	Unique Value	[1]
9.	Source	Initial source of the ride (Financial district, Theatre district, others)	Boolean; 0 = Financial District, 1 = Theatre District, 2 = Others	[0, 2]
10.	Destination	Destination of the ride (Financial district, Theatre district, others)	Boolean; 0 = Financial District, 1 = Theatre District, 2 = Others	[0, 2]
11.	Distance	Total distance of booked ride	Specific Number	[0.02, 7.86]
12.	Short_summary	The weather when the customer booked ride (Overcast, Mostly cloudy, others)	Boolean; 0 = Overcast, 1 = Mostly Cloudy, 2 = Others	[0, 2]

13.	Long_summary	The weather during a time period when the customer booked ride (Mostly cloudy, partly cloudy, light rain)	Boolean; 0 = Mostly Cloudy, 1 = Partly Cloudy, 2 = Light Rain	[0, 2]
14.	Price	Total charge for the ride	Specific number	[6.99, 59.99]
15.	High Temperature	Whether the high temperature will affect fare prices	Boolean; 0 = Yes, 1 = No	[0, 1]
16.	Low Temperature	Whether the high temperature will affect fare prices	Boolean; 0 = Yes, 1 = No	[0, 1]

2.4 Data Cleaning:

We cleaned up some of the variables of our interests. The time stamp data was in Unix format, so we converted them to fit in the local time zone so that it could be more applicable when assessing how time influences price. We also noticed that there were no prices associated with rows with “cab type” taxi. After removing unusable or irrelevant data, we are left with 637,976 observations.

We Checked if there are any duplicate rows in dataset. We searched for missing values in columns and added all the missing values. There were 55095 missing values in price column out of 693071 total rows. We renamed some columns for our better understanding. We replaced the product values in column like if Taxi then replaced as UberTaxi, shared then Lyft Shared etc. as it will give us good understanding. We created a binary variable for lyft and uber and divided the data of uber and lyft for training and testing purpose.

2.5 Imputation:

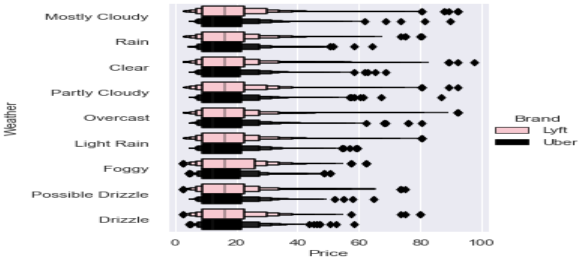
We tried to impute missing values in price column, but it was not appropriate to change the price by putting mean, median, mode or even the logit function as the difference was greater.

The only option we left was to drop the missing values in price columns even after dropping 55096 values we still have large number or rows remain which is sufficient for our analysis.

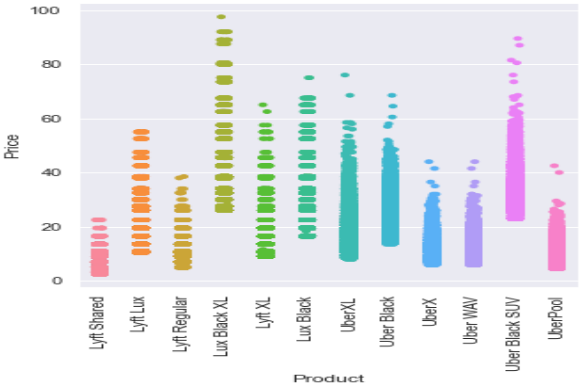
3. Results:

3.1 Charts:

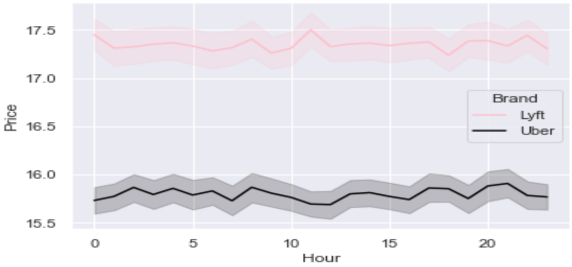
Catplot for price, weather and Brand:



Catplot for price and products:



Line Plot for price, hour and brand:



4. Modeling

4.1 Data splitting

We apportion the data into training and test sets, with an 80-20 split. After training, the model achieves 99% precision on both the training set and the test set. We'd expect a lower precision on the test set, so we take another look at the data and discover that many of the examples in the test set are duplicates of examples in the training set We've inadvertently trained on some of our test data, and as a result, we're no longer accurately measuring how well our model generalizes to new data.

training set—a subset to train a model.

test set—a subset to test the trained model.

4.2 Variable Selection:

Since we believe that most variables have a linear relation with price, we fitted a lasso model for both Uber and Lyft data with cross validation to find the best shrinkage parameter lambda and variables remaining important with that parameter. We found out that only Cab Type and Distance were shown to be important in either model.

4.3 Modeling

a. Linear Regression

Linear regression is one of the most well-known and the simplest way to predict the outputs, which fits a linear model to minimize the residual sum of squares between the predicted values and the true values. Though the main disadvantage of linear regression is that it assumes the linearity between the predicted and the response variables, but data are rarely linearly separable in the real world.

Simple Regression Model:

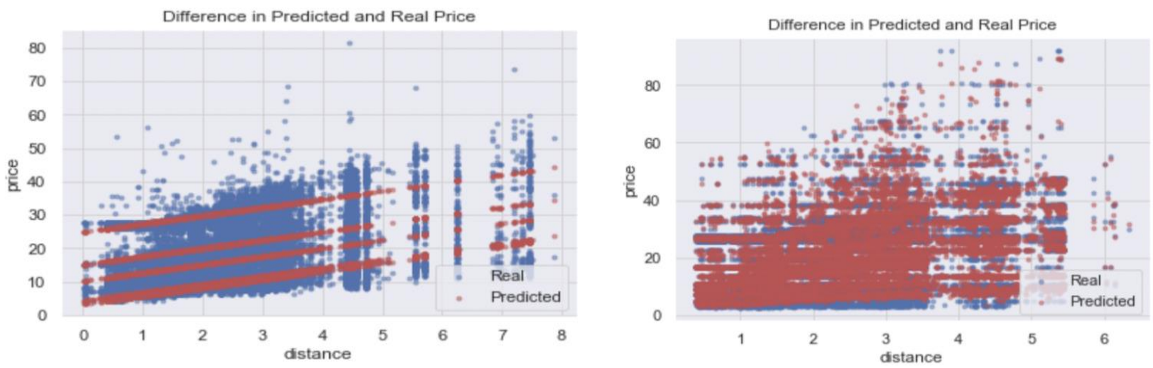
$\hat{y} =\beta0+\beta1x.y\text{^}=\beta0+\beta1x.$

b. Random Forest – It is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. It combines predictions from multiple machine learning algorithms to make more accurate predictions than an individual model. It uses the low bias and high variance to reduce the error.

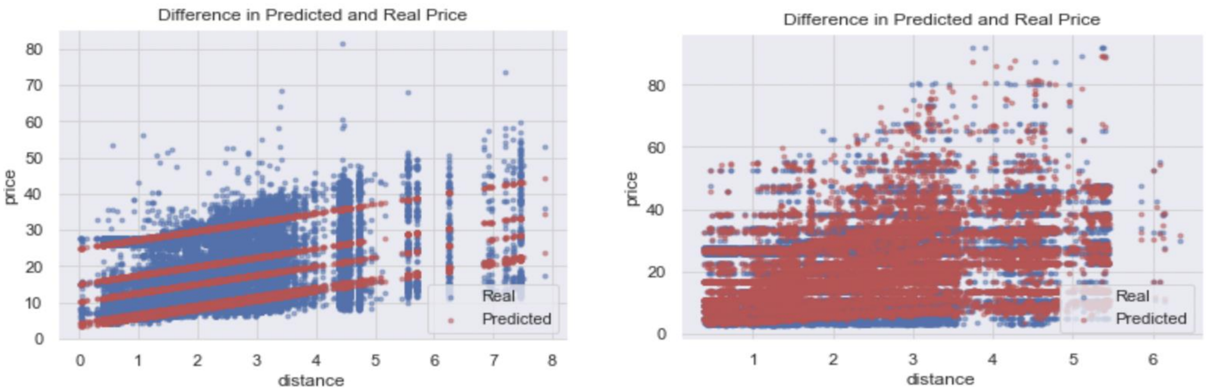
4.4 Model Results and performance:

The above tables show the different outputs obtained by running the two models for our prediction. Random Forest gave us the best predicting.

Linear Regression Plots: fig 1 lyft and fig 2 uber



Random Forest Plots: fig 1 lyft and fig 2 uber

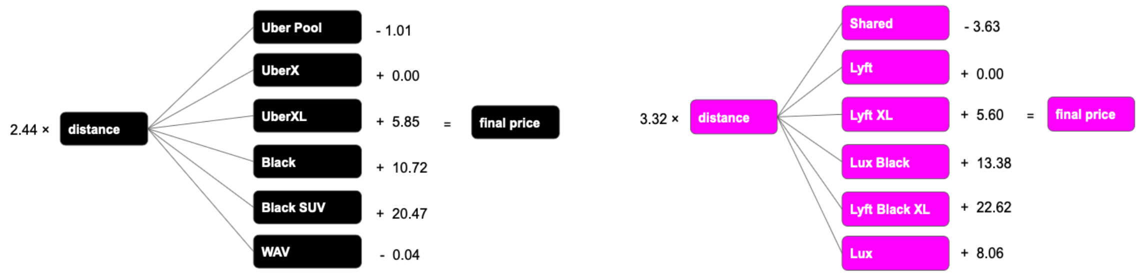


4.5 Model Comparison:

Model	Train R2	Test R2	Test RMSE	Accuracy	
0	Linear Regression	0.936489	0.936337	2.532400	84.991037
1	Random Forest	0.980563	0.975919	1.557507	91.459743

	Model	Train R2	Test R2	Test RMSE	Accuracy
0	Linear Regression	0.920045	0.918768	2.442493	88.128347
1	Random Forest	0.953556	0.947979	1.954609	91.794412

5. Conclusion



A. Price Comparison of Comparable Products between Uber and Lyft

From the graph, we can determine which situations would make Lyft/Uber more affordable than another. Additionally, the outcome influences the variables that our machine learning model for price chooses. Two findings as follows have been contributed to this model:

For shared, regular, and SUV automobiles, Lyft is less expensive than Uber.

Lyft's luxury products are more expensive than Uber's Lux and Lux SUV.

Recommendation:

In terms of the automobile types, customers would prefer using Lyft when they call a shared, regular, or SUV ride.

In terms of the Lux and Lux SUV comparison, Lyft’s price for each ride is higher than Uber’s. If customers have luxury needs, they will prefer to use Uber.

B. Uber VS Lyft: Price Comparison on Distance

Distances undoubtedly influence price, but we're interested in how differently Uber and Lyft weigh this issue in their pricing models. In our dataset, the queries of distance in Lyft have a smaller range compared to Uber. This is a human-made difference that we could not draw any conclusions. The price range for "Lux" and "Lux SUV" in Lyft, however, is clearly much more than in Uber, and the cost rises more in Lyft than in Uber as the distance grows. According to this occurrence, we concluded the following two findings:

Lyft's high-level products are more sensitive to distance variation than Uber's.

The price estimation in Lyft’s app has a larger gap compared with real price than that in Uber.

Recommendation：

In terms of Lyft’s distance sensitivity, we suggest to customers when they call the ride which distance is over 6km, they should use Uber or regular Lyft ride and do not call "Lux" and "Lux SUV" in Lyft. Regarding the Lyft price estimation gap, customers should call an Uber ride priority when the destination is outside of the base mileage; if they call a Lyft ride, they must keep proof of the estimated price; when the actual price is significantly higher than the estimated price, they can contact Lyft to find a solution. According to the boxplot of all items, we concluded the following three findings:

For shared and normal products, the weather does not change the price for either Uber or Lyft.

For luxury products, Uber’s price of “Lux SUV” in “drizzle” is slightly higher than in “clear”, but the price in “rain” is almost the same as in “clear”. Moreover, a similar pattern does not appear in its “Lux” product. Therefore, there is no apparent and consistent relationship between weather and price for Uber.

For Lyft, the price of the "Lux SUV" and "Lux' products is somewhat more in "Mostly Cloudy" and "Rain" than in "clear." This suggests that severe weather has an impact on Lyft's luxury car costs, and it may help to explain the large price variation in Lyft’s luxury products. However, we could not explain the price in “Possible Drizzle” is the lowest among all weathers in Lyft’s” Lux”. Therefore, the relationship between weather and price in Lyft still needs more investigation to support.

Recommendation:

In terms of the weather diagram, we suggest customers prioritize calling a regular Lyft or Uber ride when meeting severe weather and then may consider calling a Laux SUV.

Even though we do not have a specific investigation for an impact on Lyft's luxury car costs, but the safest way is when encountering bad weather, try not to call for a Lyft’s luxury rides.

C. Uber VS Lyft: Analysis for Product Types and Hours

In Boston, rush hour occurs from 7 am to 9 am and 3 pm to 6 pm, and it may also have an impact on prices. We excluded weekend data from the heat map because there is no rush hour on the weekends. According to this occurrence, we concluded the following two findings:

For Lyft, the price of "Share" products does not alter with the hours. The cost of "Normal" products is slightly more expensive at 1 PM than it is at other times. For “SUV”, 5 am is higher than other times and for "Lux." 10 am are higher than other times. For Lyft's "Lux SUV," the price at 10 am is higher than at other times. However, for all types of products, there is no consistent pattern in Lyft at certain times having higher prices. As a result, for Lyft, there may not be a general rule governing the relationship between hours and cost.

In order to explain why we couldn’t find a simple and universal relationship between hours and price; we did some research on surge prices. Surge prices are based on the logic of fairness, economic equilibrium, and location-specific to affect prices at different times. By adjusting prices, Uber and Lyft will increase their normal price with a “surge multiplier” to match driver supply for rider demand at any given time.

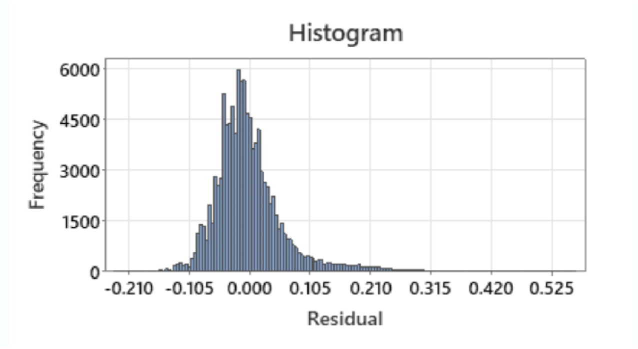
Recommendation:

In terms of the Surge prices adjustment, we suggest customers do not call a Lyft rider at busy times, such as a "Normal" rider at 1 PM, an “SUV” rider at 5 am and a "Lux." rider at 10 am because the prices are higher than other times.

In addition to the busy hours, we suggest customers take public transportation at non-office hours, weekends, and holidays to avoid the higher price by taking Uber or Lyft.

5.1 Limitation:

One of the major limitations for our study was the data collection method. All data was collected during the month of November, in a particular part of Boston. Hence, the model might not accurately predict fares during other times of the year throughout the state. Furthermore, for the Uber dataset, conditions for inference were not met as we can see a strong skew to the right in the histogram of residuals. This indicates that residuals are not normally distributed.



5. 2 Future Work

This project can be approached m ways and from different angles. The models can be improved to increase prediction precision. For instance, we could think about the interactions between the variables, such as the predictors for distance and cab types. With the ideal settings for the parameters, we can also investigate various Random Forests prediction error rates. We intend to consider extraneous data to include traffic conditions and timeframes. We plan to get Uber and Lyft fare data from more areas in Boston and fit the data into out model. We also plan to study why we saw large standard deviation of fares for a given source and destination.

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. Uranic, "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.

[6] Napitupulu, J. H. (2015, April 22). Conditions and inference of linear regression. Data Science, Python, Games. Retrieved April 20, 2022, from

http://napitupulu-jon.appspot.com/posts/conditions-inference-linear-regression-coursera-statistics.html

[7] Normal probability plot of residuals. Penn State: Statistics Online Courses. (n.d.). Retrieved April20,2022,from https://online.stat.psu.edu/stat501/lesson/4/4.6#:~:text=Skewed%20residuals,terms)%20are%20not%20normally%20distributed.

[8] Shashank H. - Department of computer application, Jain University, Bangalore, India 2018 https://scholar.google.com/scholar?hl=en&as_sdt=0%2C22&q=Uber+and+Lyft+price+prediction&oq=Uber+and+Lyft+price+predic#d=gs_qabs&t=1667604686132&u=%23p%3DsuJCF_MW6OUJ

[9] Matthew Battifarano, Zhen Qian - Transportation Research Part C: Emerging Technologies October2019 https://scholar.google.com/scholar?hl=en&as_sdt=0%2C22&q=Uber+and+Lyft+price+pr

ediction&oq=Uber+and+Lyft+price+predic#d=gs_qabs&t=1667604728495&u=%23p%3D1tOONHiJMfMJ

[10] "Lyft." Wikipedia, Wikimedia Foundation, 05 December. 2022. https://en.wikipedia.org/wiki/Lyft

[11] "Uber." Wikipedia, Wikimedia Foundation, 05 December. 2022. https://en.wikipedia.org/wiki/Uber

[12] "Kaggle Datasets." Retrieved 05 December. 2022.

https://www.kaggle.com/datasets/brllrb/uber-and-lyft-dataset-boston-ma

[13] RideGuru Team. (2022, November 02). The New and Improved Lyft Pink Membership. Rider.gure.com.

https://ride.guru/content/newsroom/the-new-and-improved-lyft-pink-membership

[14] Shokoohyar, Sina, Ahmad Sobhani, and Anae Sobhani. "Impacts of trip characteristics and weather condition on ride-sourcing network: Evidence from Uber and Lyft." Research in transportation economics 80 (2020): 100820

[15] Parajulee, Simran. "Predicting Uber and Lyft Fares Using Linear Regression." (2022).

[16] Chao, Junzhi. "Modeling and Analysis of Uber's Rider Pricing." 2019 International Conference on Economic Management and Cultural Industry (ICEMCI 2019). Atlantis Press, 2019.

[17] Zhou, Bei, et al. "Analysis of Factors Affecting Real-Time Ridesharing Vehicle Crash Severity." Sustainability 11.12 (2019): 3334.