A Driving Industry:
Uber versus Lyft

Karin Falconer-Bailey (210025601)

Abstract — This paper aims to assess and explore how price, measured in U.S dollars, alters between cab services, by making use of two popular means of on-demand, ride-hailing services—Uber and Lyft. Through this analysis, multiple linear regression and lasso regression, predictions will be made for the price of a cab fare given explanatory variables.

Keywords —Uber, Lyft, analysis, price, regression, data, k-fold cross-validation.

I. Introduction

In the current global climate, it is understandable for consumers to change their consumption patterns—opting for ride-hailing services as opposed to publicly-funded transportation networks such as buses, trains, and Crossrail. However, statistics show over the past 5 years, Uber has consistently maintained the majority of the United States' ride-hailing market share with 68%, and the remaining 32% owned by Lyft—as of May 2020 (Perri, 2021). With an ever-evolving market insight, under the expectation of further significant growth, this produces the question of what has fueled the demand for one service over another, and exactly how are the prices for each ride determined?

Demand is an important factor regarding price, a fundamental microeconomic principle is the *law of demand* which observes an inverse relationship between price and demand—in other words, demand is expected to increase because of a decline in price and vice versa.

Notably, other various factors can too influence the price of a cab and an even greater reason lies as to why one service is favoured over another. A similar study conducted by Brodeur and Nield (2018, pg. 15), found that "the number of Lyft rides per hour increases by approximately 19% when it is raining". While this paper's dataset also consists of such variables, this study will too explore and provide further explanations of factors affecting price within the transportation industry in Boston, U.S.

(231 words)

II. Analytical data and questions

Collated from Kaggle, this study relies upon one main source of data to draw its conclusions. Despite its single-origin, the data is spread between two sets. The first and most significant CSV file exhibits the study's key variables relating to ride-hailing, i.e., travel distance, price, cab type, price multiplier, ride destination, and the journey's starting point. The second dataset consists of information regarding the weather i.e., temperature, inches of rain, pressure, humidity, and wind speed. These will be combined to form one set of data. Price will serve as the target variable.

With Uber and Lyft being two of the biggest contributors to the cab service industry, this provides a great opportunity to develop insight and analysis of the industry over many segments of observations. The information gathered could prove useful to both companies—to optimize revenue and consumers in lowering and optimizing their reservation price. To deliver widespread coverage of the industry, the following research questions have been derived:

a. What are the factors that influence cab prices?

b. How do cab prices differ across days of the week? — i.e., Monday versus Thursday.

c. How do cab prices differ throughout the day? — i.e., 12pm versus 6pm.

d. Can the price of an Uber or Lyft be predicted based on the information provided?

(218 words)

III. Characterising the data

Prior to conducting any meaningful analysis, it is first important that the data is characterized. Doing so ensures that the data is suitable for analysis and aligns with the study's research aims. As the datasets are yet to be combined, they will be individually characterized. An overall ACRRT quality analysis will be conducted collectively on the data.

Given the nature of the data, the implemented machine learning model will be under supervised learning algorithms. A supervised learning machine model is whereby “a set of [labelled] data (known instances) is provided for the model to make acceptable relations between the features” (Pandey and Rautaray, 2021, pg. 67).

3.1. Cab dataset

The cab dataset contains 637,976 entries and 10 columns. The dataset is primarily qualitative accounting for 60% of the file, however, the data too takes the shape of quantitative: ratio, categorical: nominal, and categorical ordinal data.

3.2. Weather dataset

The weather dataset contains fewer observations with 6,276 entries and 8 features. 87.5% of the information demonstrates quantitative qualities—these values are displayed as quantitative: ratio, and categorical: nominal.

3.3. Quality analysis

The *accuracy* of the data is substantial. There is no evidence of typos or errors. All metric features such as timestamp, price, temperature, and wind speed have a clear assigned metric value consistent with the data’s purpose. Distance has not been assigned any metric value, however, as the data was conducted over the City of Boston, it can be assumed that it is measured in miles.

As aforementioned, the information provided has a combined total of 644,252 observations. From that, a mere 9.38% (60,477) contains missing values. Given that the data takes many observations, this should not be deemed problematic for the data analysis process nor should negate its *completeness*.

Most of the included features are *relevant* in conducting appropriate analysis on the price of Uber and Lyft rides. ‘Product id’ and ‘id’ attributes have been deemed unnecessary and will be removed during the data preparation process.

No inconsistencies have been identified; thus, the data can be regarded as *reliable*.

Despite the notion that the data was assembled 3 years ago, given the nature of the study is to predict prices based on the obtained information, this does not qualify the data as being outdated. Although the base fare of an Uber or Lyft ride may have altered, this study does not aim to predict current prices and therefore remains *timely*.

(405 words)

IV. Analysis

This section will detail a step-by-step analysis of the key stages leading up to the study’s final model.

4.1. Data preparation

Data cleansing and preparation is a fundamental stage of the data analysis and modelling process, so much that according to McKinney (2017, pg. 195), it is often “reported to take up 80% more of the analyst’s time”.

4.1.1 Handling missing values

As aforementioned, a total of 60,477 observations are missing completely at random (MCAR) across both the weather and cab files. Each file’s missing values have been handled alternatively and is discussed below.

The 55,095 missing values are that of cab prices. Opposed to replacing the values with the mean or another metric, these rows were dropped from the data frame. Contrariwise, with the missing 5,382 values of rain data, it is unknown whether they portray a meaningful zero and indicate a lack of rainfall during this period. These absent observations underwent mean substitution, using their replacement with the overall mean of rainfall.

4.1.2 Removal of unnecessary features

Limited features of, primarily the cab dataset, exhibit little to no value and thus, its observations will be removed. The ‘id’ and ‘product id’ elements detail the unique identifier for the customer, alongside the Uber or Lyft’s identifier. As the dataset provides information into the type of cab i.e., Uber XL or Uber Pool, these identifiers, if retained in the dataset, would go unused. The removal of these values creates a clear distinction for the variables which may pose as significant in the predictive modelling process.

4.2 Data derivation

Once the data has been cleansed of any incomplete values, the following stage is to transform the data into the appropriate structure for analysis, also referred to as data transformation.

4.2.1 Changing attribute units

The time feature exhibits its value utilizing the Unix epoch timestamp. Though proving useful, without the knowledge of how to convert such observations into a readable format, its current form is unpractical and if disregarded, may render the model's outcome. Using python's '*datetime'* module, we convert the timestamp into five legible columns: time, date, weekday, hours, and time of day.

4.2.2 DETECTING AND FILTERING OUTLIERS

When detecting and filtering outliers, they must be only altered with proper justification. To identify these outliers, we compute individual boxplots to demonstrate their distribution. *The main variables which exhibit outliers are distance, humidity, pressure, price, price multiplier, and temperature.*

Figure 01. Humidity distribution

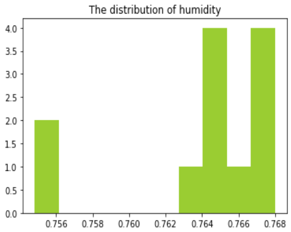


Figure 02. Pressure distribution

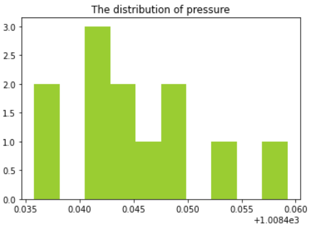
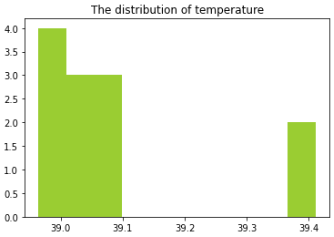


Figure 03. Temperature distribution



Given the nature of the distance, price multiplier and price variables, it can be assumed that these *outliers* will prove significant in the construction of predictive models. These observations are subject to drastic change and therefore, will not be withdrawn from the data frame.

However, the humidity, pressure and temperature outliers exhibited in figures 01-03 are extremely isolated. It is likely that these observations would produce statistical errors and thus will be removed from the construction of any predictive models.

4.3 MODEL CONSTRUCTION

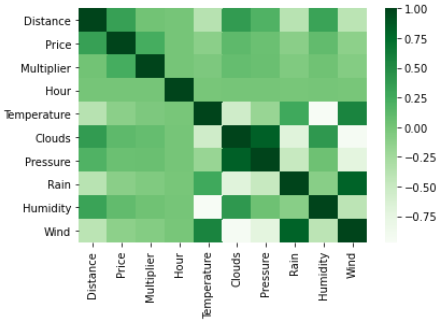
Model construction is a significant process in the edifice of the final predictive model. Due to the dataset's multivariate features, the study will utilize regression models to predict and analyse the variance of which the independent variable's variation is explained by price. This section outlines and analyses the study's feature selection methods and feature scaling techniques.

4.3.1 FEATURE SELECTION

Feature selection is used in instances where the observed dataset contains high dimensionality. As stated by Li et al. (2018, pg.1), the method aims to "[build] simpler and more comprehensible models, improving [the] data-mining performance, and preparing clean, [and] understandable data".

One method of feature selection has been used, the filter method. The filter method involves utilizing Pearson's correlation to create a matrix, allowing for the identification of the features which hold a strong correlation to price.

Figure 04. Correlation heatmap



While there are in total three main methods of feature selection—including the wrapper and embedded method—the filter method provides numerous advantages in the modelling process. The output of Pearson's correlation saves time and reduces the number of models during training. We can eliminate the responses which have an insignificant correlation with price, while generating models which possess a greater r-squared value.

4.3.2 TRAINING AND TESTING DATA

Considering the data is comprised of thousands of observations, the combined CSV will be split into an 80:20 ratio for training and testing. While there is no optimised way to split data into training and testing, the Pareto principle, coined by economist Vilfredo Pareto which is based on the infamous 80/20 rule "specifies that the 80 per cent of an output or result comes from the 20 per cent of the effects or disproportionate alliance of inputs and the outputs" (Akdeniz, 2019). Given Pareto's theory, it is assumed that in a machine learning context, the 80:20 division of the data will yield an optimal outcome with "80% of effects [deriving] from 20% of causes" (The Data Detective, 2020).

4.4 VALIDATION OF RESULTS

Validating model results is an important process of machine learning and data analysis. Cross-validation acts as a technique to increase the efficiency of a model due to counteracting overfitting in a predictive model.

4.4.1 K-FOLD CROSS-VALIDATION

For all predictive modelling, each model will utilise 10 k-fold as the method of cross-validation. The output of cross-validation allows the analyst to understand the model's accuracy, thus enabling the creation and/or tuning of the final model.

(881 words)

V. FINDINGS

This section details analysis of the findings obtained from both training and testing models. The results of the final predictive model will act to answer the study's research questions.

5.1 MULTIPLE LINEAR REGRESSION

In total, two multiple linear regression models have been computed.

Figure 05. MLR training vs. predicted

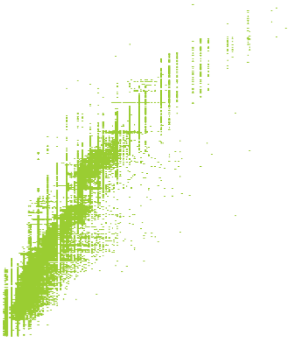


Figure 05 exhibits the training results of multiple linear regression against predicted values. The multiple linear regression model uses a cross-validation method of 10 k-folds and includes all variables of the model.

Overall, it is evident that the model follows a particular shape as its curvature tends to the right. While this is apparent, the model does not possess complete linearity which has resulted in the model being over predicted, this is otherwise referred to as underfitting.

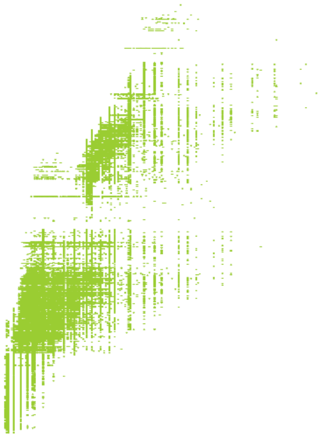
Although this regression model shows underfitting, it has an accuracy and r-squared value of 0.93 and a standard deviation of 0.001. An r-squared value of 0.93 indicates that 93% of the variance in the predictor (price), is explained by the model's response observations.

Figure 06 presents the training results of multiple linear regression against predicted values. This second multiple linear regression model too utilizes 10 k-fold cross-validations, however, given the results of the filter method, this model utilizes fewer features in the aims of increasing its overall performance. These variables include: distance, CabType.Black SUV, CabType.Lux Black XL and CabType.Shared.

Similar to figure 05, figure 06's model contains a particular shape. While the model exhibits greater linearity, as opposed to its more featured versioned, it too displays multiple exceptions. Though this is the case, it can be noticed that there is similarity within the two plots in terms of the plotted actual values.

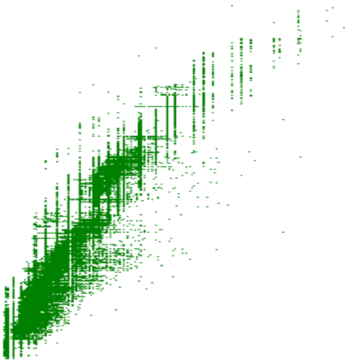
Model 2's 10 fold cross-validation shows an accrued mean accuracy of 0.68 and a standard deviation of 0. Given that this less featured model possessed the ability to generate a lower r-squared value, it demonstrates the opposite. 68% of the variance in price is explained by distance, CabType.Black SUV, CabType.Lux Black XL and CabType.Shared.

Figure 06. MLR2 training vs. predicted



The alternative model of the three, uses Lasso regression to conduct predictive modelling, and again 10 fold cross-validation to validate the model.

Figure 07. Lasso training vs. predicted



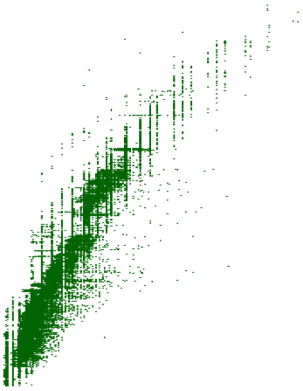
Among the three, Lasso regression acquired a smaller level of accuracy during the validation process, obtaining a value of 18%. Despite its low accuracy, the model achieved an r-squared value of 0.93 and a standard deviation of 0.001 due to containing all features mentioned in the study.

Having computed the final model, we can now try to answer the study's research questions:

What are the factors that influence cab prices?

The final model exhibits a high r-squared value of 0.93, and thus is it evident that many features that explain the price change can be found in the dataset. However, 7% of the price variation can be explained by external explanatory attributes which are not featured in the datasets.

Figure 08. MLR model final



How do cab prices differ throughout the day? It is apparent that price does not alter significantly with the time of day. According to the Ordinary Least Squares summary of the model, for each 0.0002 USD change in price, the time-of-day changes by one unit.

Can the price of an Uber or Lyft be predicted based on the information provided? Through regression coefficients, we can aim to predict the prices based on numerical values inputted into the regression equation. The price multiplier, distance and hour appear to be the only features that appear to have a positive relationship with price, increasing simultaneously. The remaining factors appear to possess an inverse relationship with price, in a similar manner as the aforementioned law of supply and demand.

VI. REFLECTIONS AND FURTHER WORK

Given the results of which the final model exhibits, in relation to the study's research questions, it is not as expected nor as one would have hoped. In future studies, machine learning and data analysis computation, will utilise files with smaller observations. During the stages leading up to model construction, there were multiple instances of NaN value removal which increased the duration spent in the preliminary stages and less time was spent building and analysing the models.

Additionally, I will aim to optimise non-linear regression models during the construction of a machine learning model. Multiple linear regression models and simple linear regression models have less complexity as opposed to neural networks or random forest models. Alternative methods of cross-validation will also be considered and explored to provide models with a variety of performance outcomes and results.

(735 words)

VII. REFERENCES

1. Akdeniz, C. (2019) The Pareto Principle. IntroBooks. Available at: https://www.google.co.uk/books/edition/The_Pareto_Principle/uQiEDwAAQBAJ?hl=en&gbpv=0&kptab=morebyauthor (Accessed 5th December 2021)
2. Brodeur, A. and Nield, K. (2018) An empirical analysis of taxi, Lyft, and Uber rides: Evidence from weather shocks in NYC, *Journal of Economic Behavior and Organization,* 152(), pg. 1-16. Available at: https://www.sciencedirect.com/science/article/pii/S0167268118301598 (Accessed 19th November 2021)
3. Fandango, A. (2017) Python Data Analysis. 2nd ed. Birmingham: Packt Publishing, Limited. Available from: ProQuest Ebook Central (Accessed 23rd November 2021)
4. Li, J. et al. (2018) Feature Selection: A Data Perspective. *ACM Computer Survey.* 50(6), pg. 1-45. Available at: https://doi.org/10.1145/3136625 (Accessed 2nd December 2021)
5. McKinney, W. (2017) Python for Data Analysis. 2nd ed. Sebastopol: O'Reilly
6. *Molin, S. (2019) Hands-On Data Analysis with Pandas: Efficiently Perform Data Collection,*

Wrangling, Analysis, and Visualization Using Python. Birmingham: Packt Publishing.

7. Pandey, M. and Rautaray, S. (2021) Machine Learning: Theoretical Foundations and Practical Applications. Studies in Big data. 87(). Available at: http://dx.doi.org/10.1007/978-981-33-6518-6 (Accessed 2nd December 2021)
8. Perri, J. (2021) The U.S. Rideshare industry: Uber vs. Lyft, *Bloomberg Second Measure.* Available at: https://secondmeasure.com/datapoints/rideshare-industry-overview/ (Accessed 2nd December 2021)
9. The Data Detective (2020) The 80/20 Split Intuition and an Alternative Split method, *Towards Data Science*. Available at: https://towardsdatascience.com/finally-why-we-use-an-80-20-split-for-training-and-test-data-plus-an-alternative-method-oh-yes-edc77e96295d (Accessed 5th December 2021)