Assessing When to Use Uber or Lyft Through Machine Learning Techniques

Kevin Qualls – Master's in Data Science at Brown University – Dec. 3rd, 2019

GitHub Link: https://github.com/kevin-qualls/data1030_final_project_December

Introduction

Ubers and Lyfts have made it convenient for many to commute from point A to point B. Clever enough, however, users have figured out that the prices of ubers and lyfts can be lowered if they order it at a certain time or day, and from a certain location. The goal of this assignment is to apply machine learning techniques to help users know when to purchase an uber/lyft at the cheapest rate possible.

Exploratory Data Analysis

The dataset *cab_rides.csv* originally had 14 features and 693,071 rows of data, however after deleting columns not of interest and manipulating the date and time columns, the dataset reduced to 10 features. The dataset had 7.95% of its data missing with `NaN` values, all from the **price** feature (indicating the price of an uber/lyft). Because the **price** feature is our target feature, it is permissible to delete all of these rows.

When analyzing the dataset (before preprocessing), we see in Fig. 1 the box plots of Uber and Lyft are nearly the same, with both having a median price of approximately $17. The minimum price for uber, however, is a few dollars higher than that of Lyft. Furthermore, the range of the outliers for Lyft seems larger (going up to almost $100), where as the outliers of Uber seem clustered between $40 and $70.

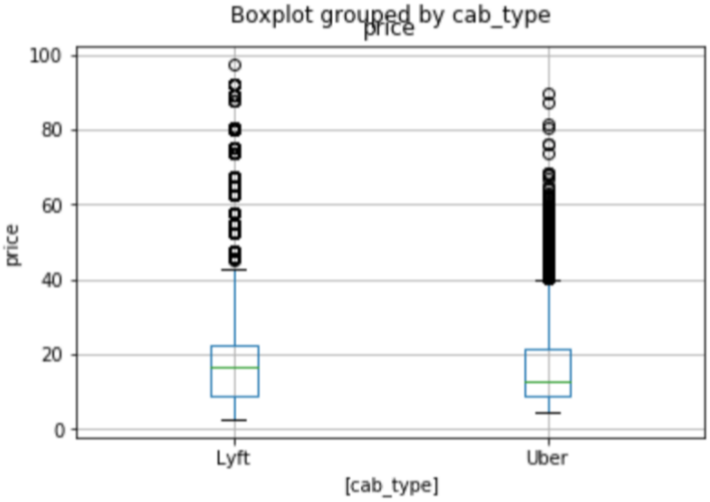

Fig. 1: Price vs. Cab Type

Perhaps Fig. 1 suggests Lyft users travel farther distances than Uber users, which is why there is a higher pay. Or perhaps Lyft is flat-out more expensive. Interesting enough, however, the graph in Fig. 2 shows price and distance have no linear relationship, as one would think.

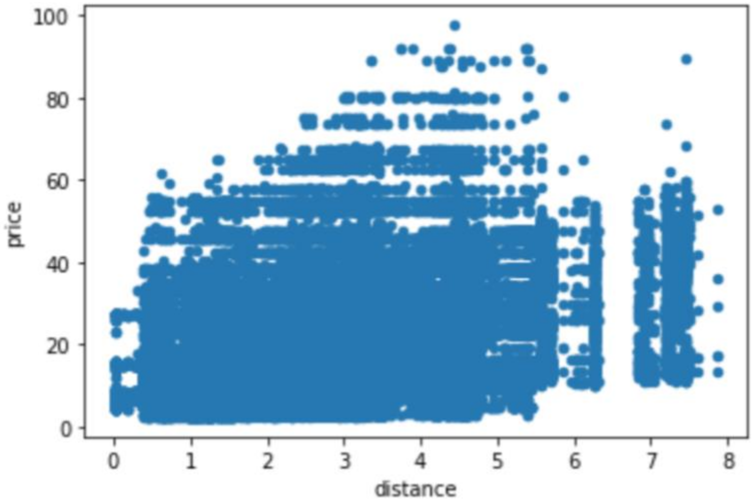

Fig. 2: Price vs. Distance

Perhaps the distance doesn’t influence the cost, and instead, the cost is inflated different days of the week with the surge multiplier. When examining this relationship in Fig. 3., we see how the surge multiplier doesn’t change for different days of the week for a multiplier rate below 3.0. However, when the surge multiplier is 3.0, there is no surge for Friday (blue bar), and a much larger surge for Tuesdays (brown bar). Overall, it seems the surge multiplier doesn’t change on different days, and therefore the price doesn’t change with respect to the day of the week.

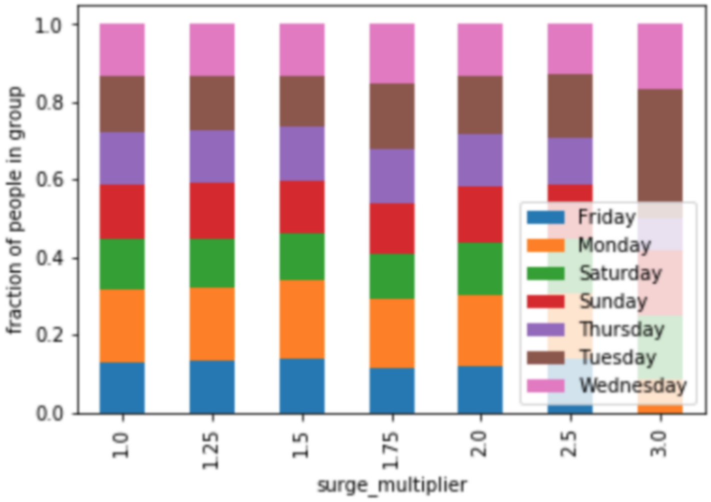

Fig. 3: Surge Multiplier Throughout the Week

Furthermore, the above claim supports Fig. 4, which shows the price is relatively the same for Uber and Lyft throughout the week.

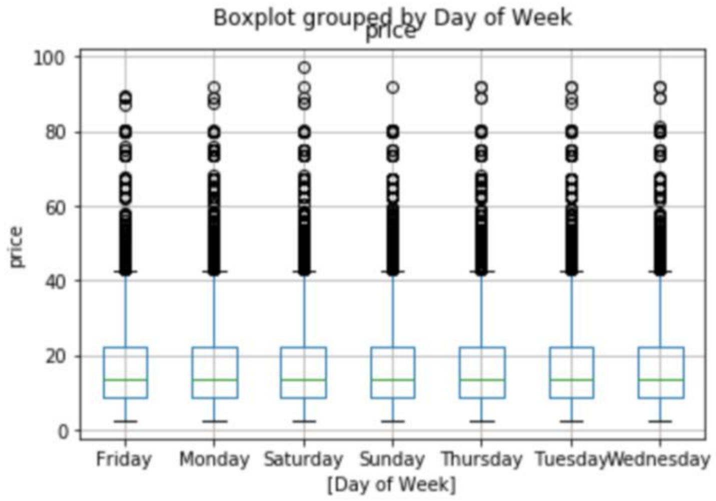

Fig. 4: Price Throughout the Week

Additionally, it seems the prices are the same when it is rush hour or not rush hour, indicated by Fig. 5.

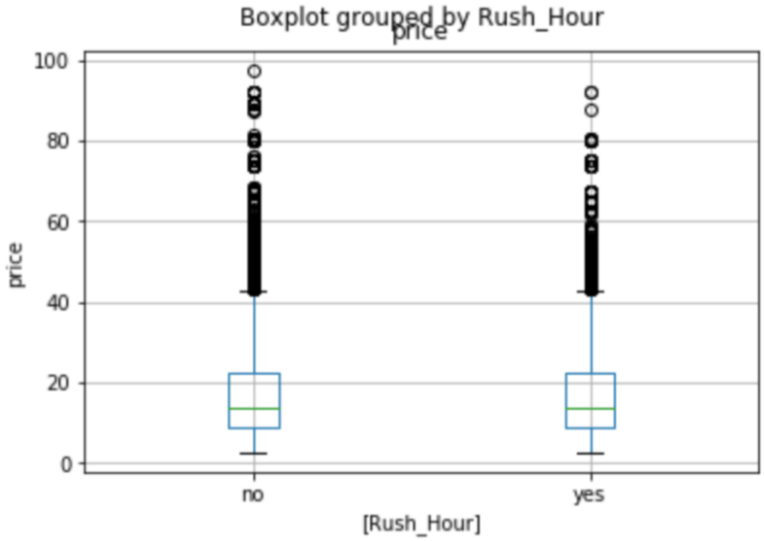

Fig. 5: Price With Respect to Rush Hour

Based on the previous graphs, it seems distance, rush-hour, and day of the week have no influence on the price of an Uber or Lyft, which seems odd. If this were true, then there should be an equal proportion of users who use Uber and Lyft. However, Fig. 6 reveals a significant difference for the proportion of users who use Uber.

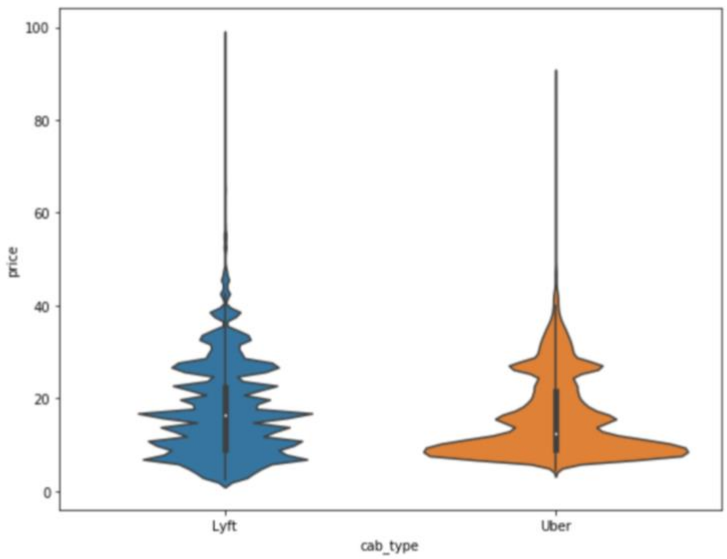

Fig. 6: Price vs. Cab Type

According to the violin plot, more users prefer to take Uber for a $10-$13 trip, compared to Lyft. Perhaps this implies Uber is cheaper for shorter-distance rides, compared to Lyft.

Methods

Splitting and Preprocessing

The original dataset was split into three subsets: a train (60% of the original dataset), test (20%), and CV dataset (20%). The test, train and CV were each preprocessed, with 5 features being one-hot encoded, one being ordinal encoded, one using MinMax scaler, two using standard scalar, and the target feature using the label encoder. An X subset and Y subset was then made for each preprocessed set (X_train, Y_train, X_test, Y_test, X_cv, Y_cv). The X sets had a final count of 39 features.

Correlation

When trying to find the top features that are most correlated with price, the feature Month_Day is the highest for test set (0.005), x2_South Station is the highest for the train set (0.006) and distance and surge_multiplier for the CV set (0.341 and 0.223, respectively). The results for the later result make sense, since distance and surge multiplier should influence the price ; plus the results from the other data sets were insignificantly small.

Model Performance

I made 5 models (Linear Regression, Feature Engineering, Random Forest, XGBoost and Simple Pipeline) shown in Table 1. I also compared the R2 scores of each model to the Baseline model, to see how each model faired. For “Feature Engineering” I made a column by multiplying the price of an Uber/Lyft with the distance it travels. For the Random Forest Model, I ran it 4 different times by tuning the n_estimators and max_depth parameters for each one. Lastly, I made a Simple Pipeline for my train, test, and cv data, because a K-Fold CV model took too long for my computer to process ; this was also the same for an SVM Model.

Table 1: Mean Squared Error and R2 Score of Different Models

Model / Method	MSE	R2 score	Baseline
Linear Regression	~ 304	~ 0.161	~ -2.19 E -05
Feature Engineering	~ 303	~ 0.162	~ -2.19 E -05
Random Forest (Tune Parameter 1)	~ 301	~ 0.168	~ -5.43 E -05
Random Forest (Tune Parameter 2)	~ 329	~ 0.093	~ -2.31 E -05
Random Forest (Tune Parameter 3)	~ 302	~ 0.165	~ -3.15 E -05
Random Forest (Tune Parameter 4)	~ 312	~ 0.139	~ -1.95 E -05
XGBoost	~ 398	~ -0.10	~ -10.2
Simple Pipeline	CV: ~ 312 ; test: ~303	NA	NA

Overall, all models performed really poorly, as there were high MSE scores and low R2 scores. Even my Baseline R2 scores seemed inaccurate as they were all negative. Perhaps it's safe to conclude this dataset cannot be modeled.

Results

Comparison to Baseline Model

For my Linear Regression model, feature engineering model, and random forest model, the baseline is nearly 0 (i.e. the mean of my target variable price). I suppose the R2 scores for these models are less than 0.2 away from the baseline, yet this doesn't really tell anything, as the baseline shouldn't be 0.

Feature Importance

Out of all 39 features, distance and surge_multiplier are the most important when determining the price of an uber/lyft. The distance feature is the tall bar in Fig. 7, with a score of 0.686, and the surge_multiplier is the other bar with a score of 0.294. The other 37 features add up to 0.02, which is why they seem like they don't appear in Fig. 7.

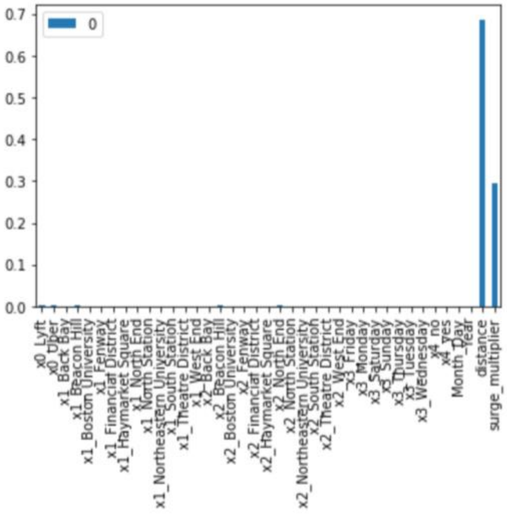

Fig. 7: Feature Importance

Distance traveled and surge multiplier should contribute the most to the price of an uber/lyft, since, for example, a 100 mile trip cannot cost the same as a 2 mile trip, not can the price during a period of high-demand can cost the same as the price in a period of no demand. The results from feature importance support basic mathematical and economic concepts.

Outlook

Future Improvements

The weak spot of this data set is that there's 693,000 rows of data from less than 3 weeks in the Boston area. Because the time-frame is so small and the location is so specific, the results cannot be generalized for any user who will purchase an uber or lyft. The easiest way to account for this is to find a dataset that spans a couple of years from multiple regions throughout the country.

In my EDA section, there seems to be no linear relationship between price and distance, although the distance traveled was classified as an important feature of price in my Results section. The discrepancy between the two graphs implies the dataset is too specific, and needs to have a wider range of time and location.

Additional Techniques

Perhaps I could have subsetted a sample of my data to a manageable number of rows to perform SVM and K-Fold cross validation. As powerful as these techniques are, however, they cannot make up for how specific and homogenous the collected data is.

Conclusion

Ultimately, machine learning techniques can provide insight to help users know when and where to strategically purchase an uber/lyft for a cheaper rate. Better conclusions can be drawn with a dataset that considers multiple years and geographical locations.

References

Source of Data: https://www.kaggle.com/ravi72munde/uber-lyft-cab-prices