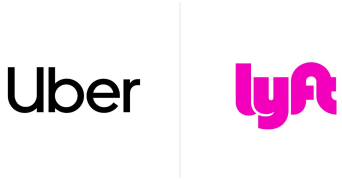
Capstone 1 Final Report

By Samuel Ma

Code:https://github.com/samwrite/Springboard/blob/master/Capstone%201/Capstone%20Project%201%20Final.ipynb

Slides:https://docs.google.com/presentation/d/1eg7ftq8SOl9lUimGDA6ZSRN05DooqWN3NQKAaQlgLdk/edit?usp=sharing



Problem Statement: Is it cheaper to ride with Uber or Lyft?

In this report, we will be exploring whether or not it is cheaper to ride with Uber or Lyft. As both of these businesses are thriving amidst the more traditional forms of public street transportation (busses and taxis), it would be interesting to see which of these services can save you some spare change from the other. People who would be interested in this study would be everyday commuters as well as someone who may want to create a third service to compete with the current two. If there was a way to further optimize pricing and understanding commuter patterns, it may be possible to create more business competition.

Data Acquisition and Cleaning

All data was taken from https://www.kaggle.com/ravi72munde/uber-lyft-cab-prices. There are two csv files, one is weather data and the other is cab data. I took the following steps to clean up the data.

Weather data

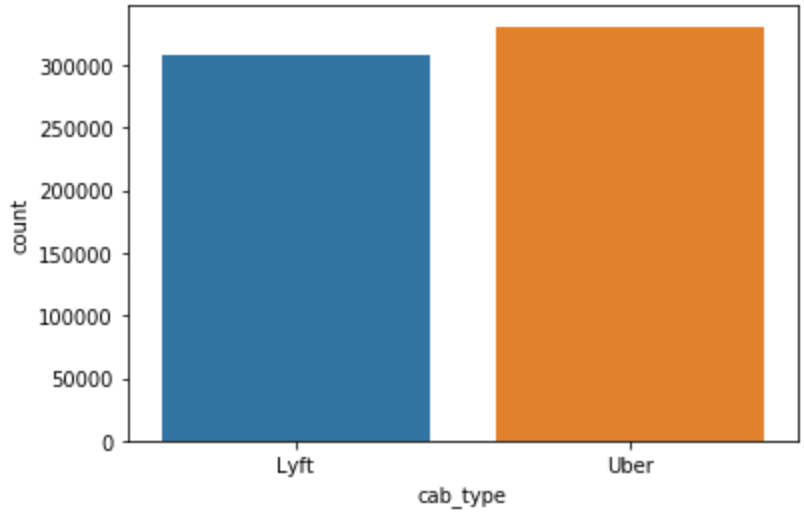
- Examined that there were values for rain which were NaN, changed them to 0 because it meant that there was no rain.
- Changed time_stamp from epoch to human readable form

Cab data

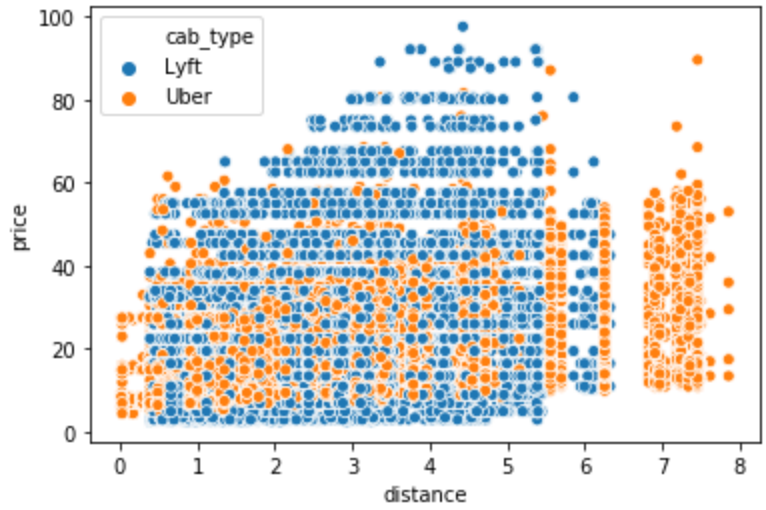
- Extracted all Uber rides with name Taxi, these entries did not provide a price, so I assumed that these were actual taxi rides which will not be a part of this study. They have a different pricing schema which can be explored further as a third comparison in the future.
- Changed time_stamp from epoch to human readable form

These changes removed all non_null values and ensured that all columns now have the same number of values.

Before we explore more of the data, let's first look at what we are given and compare Lyft and Uber prices based on distance travelled.



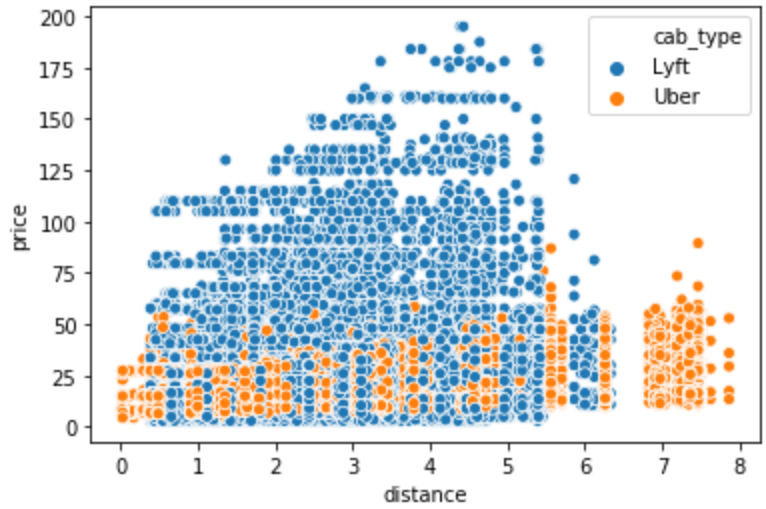
First off we see that the datasets contain nearly the same number of samples each, with Uber having a little more. Due to the large sample sizes, this difference can be made negligible.



Above is the comparison between Lyft and Uber looking at price v distance. From this, we can see that Uber has a rather steady increase in price over distance, and that it is chosen the majority of the time when distance is greater than 6. However, Lyft has prices which seem to be comparatively lower as well as higher than Uber. Another thing we need to factor in will be the surge multiplier. In our data, surge multiplier is a factor which increases the price of a ride due to heavy demand at particular times.

surge_multiplier
1.00 286433
1.25 11085
1.50 5065
1.75 2420
2.00 2239
2.50 154
3.00 12

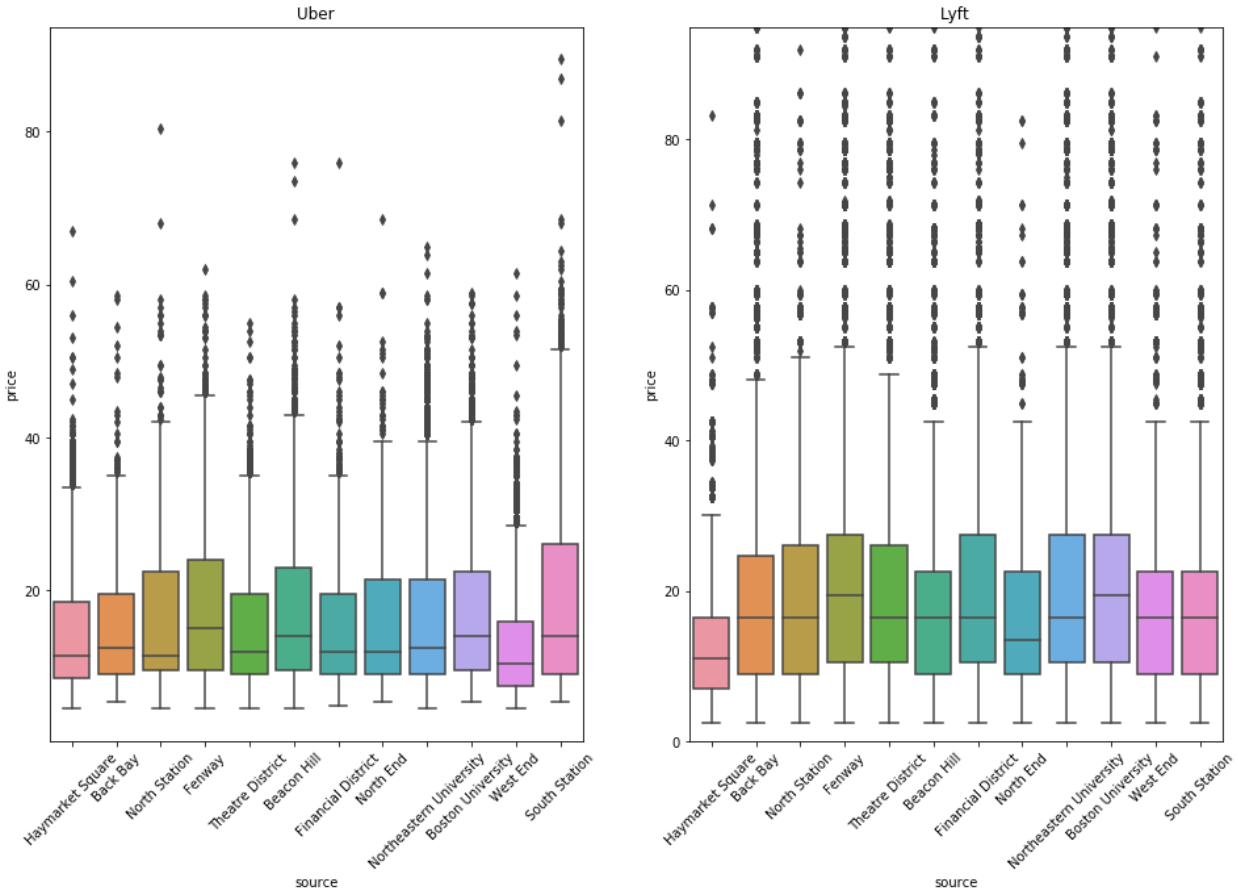
We see that Lyft has surge multipliers ranging from 1 to 3 (although most of the rides stayed at their original price), while Uber only has surge multiplier of 1. Upon further research it seems that Uber applies its surge multiplier to its prices without its customers seeing it (they see the final price at face value), while Lyft shows the multiplier. Since this factor affects Lyft prices, we can apply it to the old data and create the true price dataset.



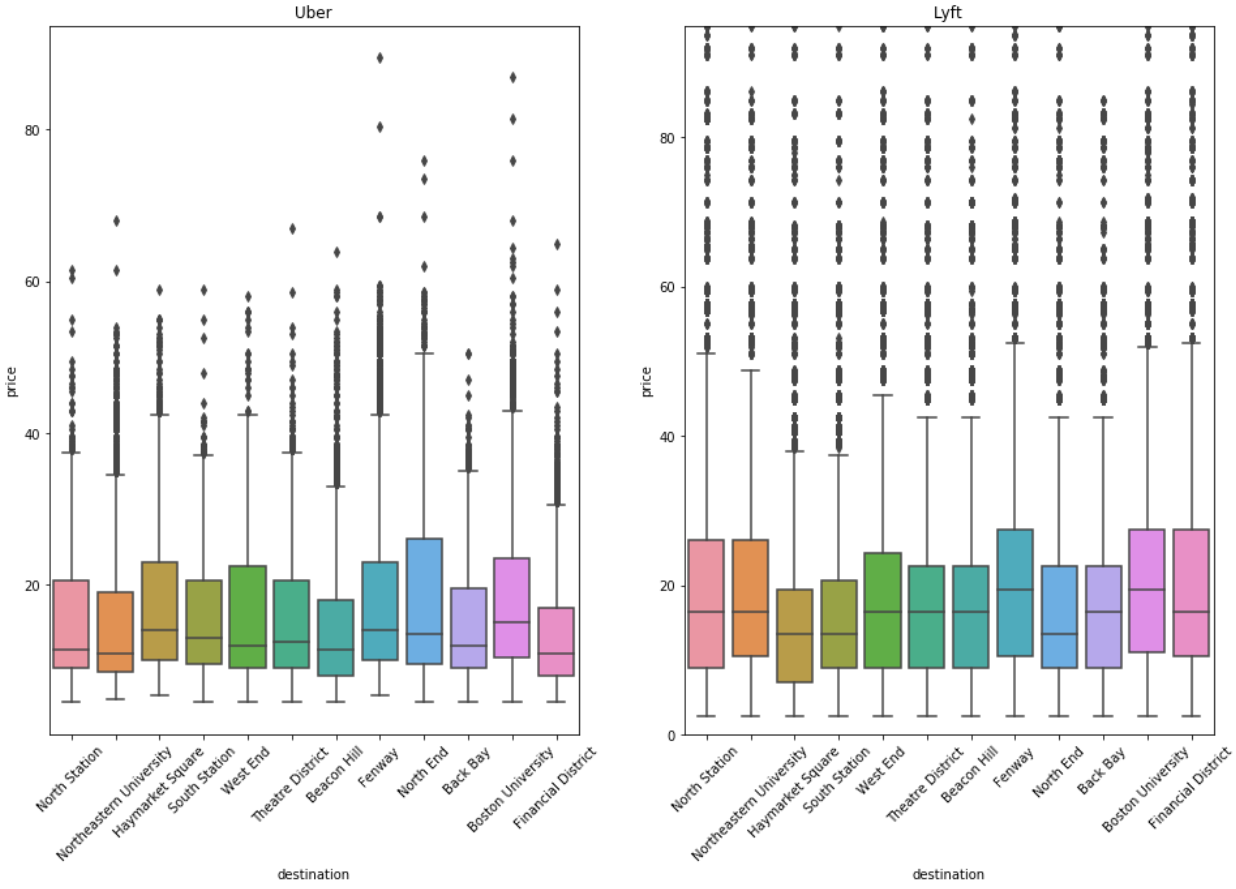
After factoring the surge factor and creating the new true dataset, we see an increase in price for the Lyft rides. At this point it looks like Lyft's pricing is higher by a fair margin.

Let’s look at location (specifically the source/destination point of each ride) and see if we see trends there. There is a myth out there claiming that rides could get more expensive due to where passengers are being dropped off as well as where they are picked up. There are a lot of outliers so we can ignore them and use boxplots for better viewing purposes.

Comparing Sources:

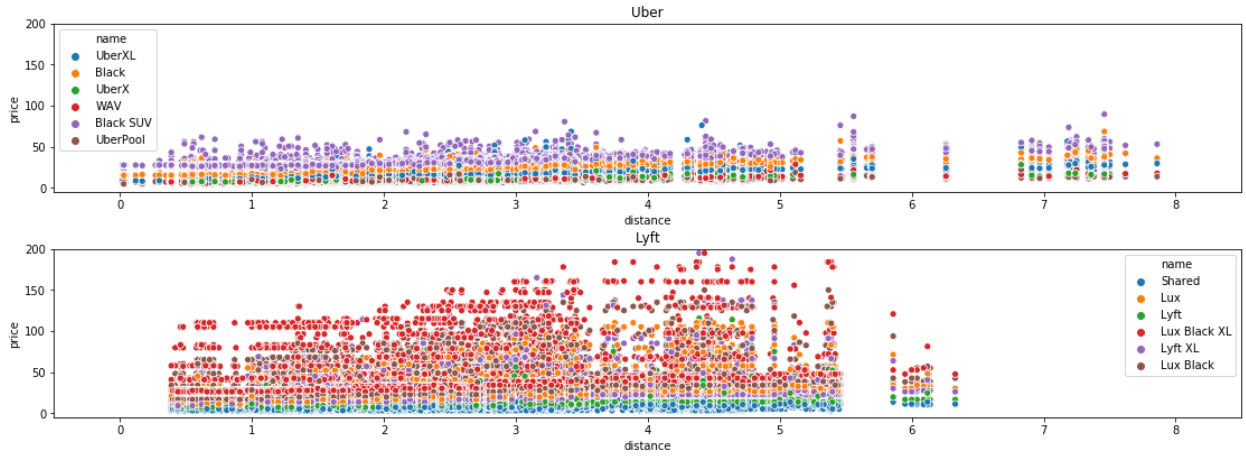


Comparing Destinations:



Overall, Lyft prices are comparatively higher than Uber prices regardless of location. Location also does not seem to be a huge factor in deciding the price of a ride, since there are no convincing locations for both services showing a significant decrease/increase in pricing.

Let's also look at the different types of rides that both services offer.



At first glance, we can see that each of the different rides fall in different pricing ranges. For example, Black SUV Uber rides cost more than WAV Uber rides at a linear rate. Similarly with Lyft, the regular Lyft ride costs less than the Lux Black XL ride.

So far we can conclude that Lyft rides are more expensive than Uber rides. There are more factors we can explore and will apply some machine learning algorithms to create predictive models.

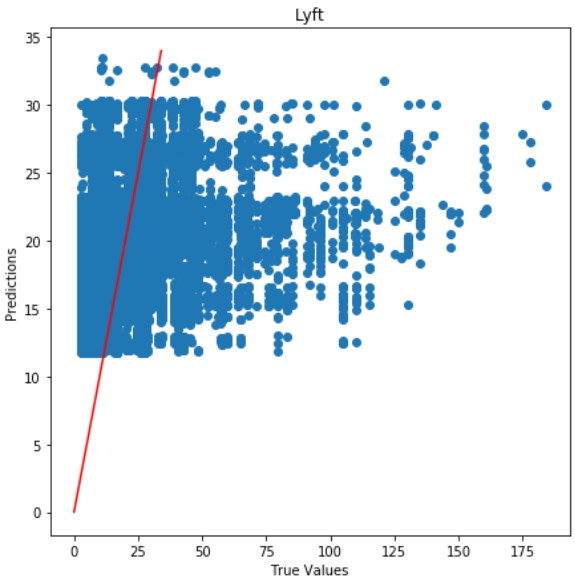
Machine Learning Analysis

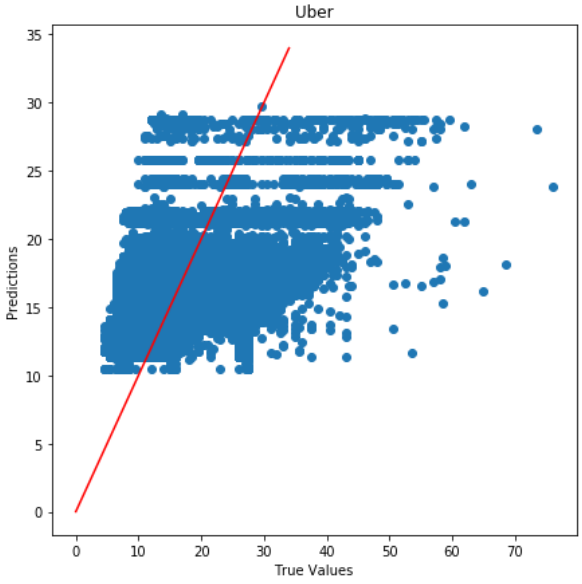
We use Linear regression and Random forests to create models for our Lyft and Uber data to predict prices. This allows for us to determine if our predictions are worth making for future rides or if there are other hidden potential hidden/undisclosed data which affects prices.

Linear Regression:

Right now, the only continuous variables we see in our data are price, distance, and surge multiplier. We will set our y as price, and X as the other two variables to create our train/test values.

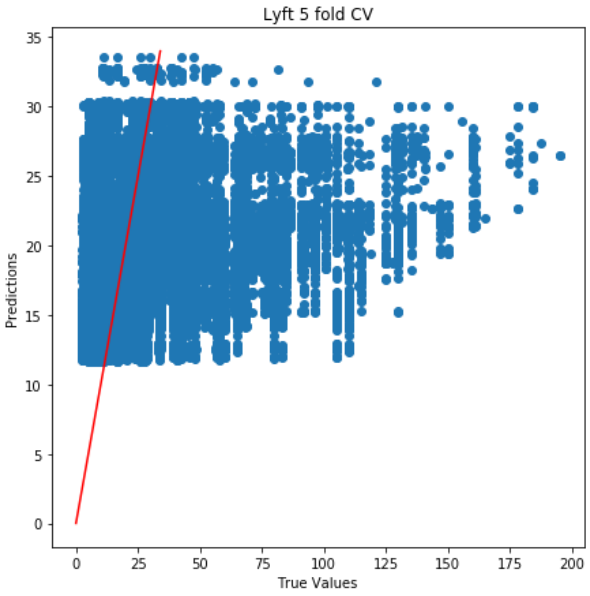
Plotting our y test values to our predictions with our model, we get the following. Note the red line signifies y=x so we can visualize how accurate the model is.

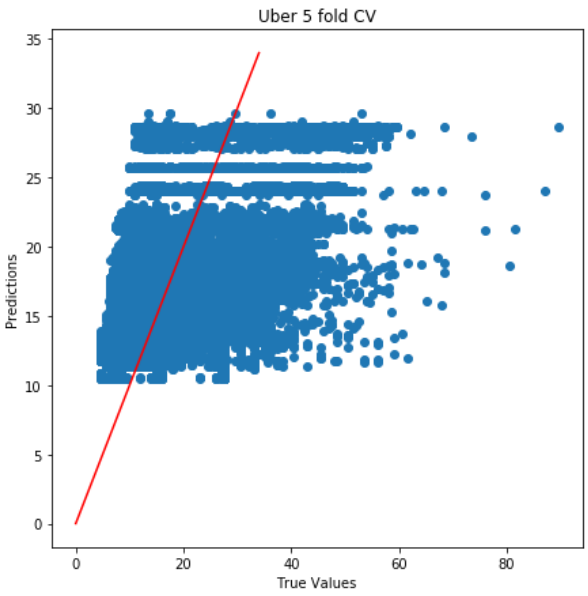




We see here that our model is not perfect but gives a good idea of how well it fits with our data.
Score: 0.10168683165668924
Score: 0.10985819294871513

We use 5 fold cross-validation to get the result below.

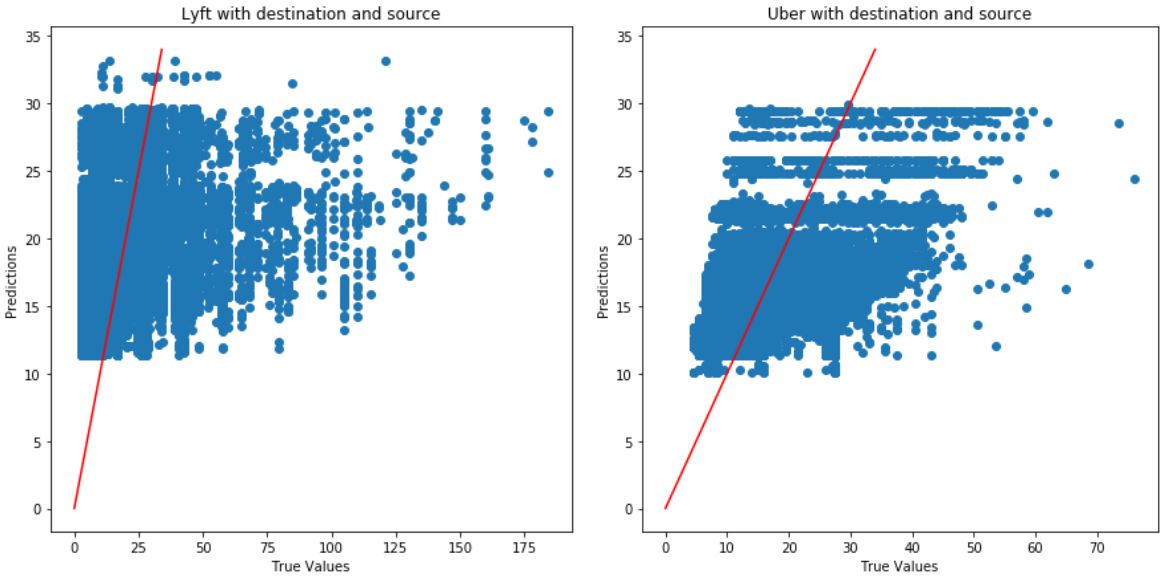




The charts are not the exact same as before but they give an idea that the validation shows consistency.
Lyft CV scores: [0.09242405 0.09496105 0.09443755 0.09730579 0.09674657]

Lyft Cross-Predicted Accuracy: 0.09518297341864324
Uber CV scores: [0.11508637 0.11607605 0.11216367 0.11169404 0.11003401]
Uber Cross-Predicted Accuracy: 0.11302726305689426

We can also take a look at if we use dummy variables for our categorical features and add them into our model.



Lyft CV scores: [0.09778943 0.10019079 0.10019576 0.10310951 0.10102531]
Lyft Cross-Predicted Accuracy: 0.10168683165668924
Uber CV scores: [0.11660139 0.11751953 0.11358684 0.11298346 0.11148821]
Uber Cross-Predicted Accuracy: 0.11164952921809902

With the dummy variables there is very minimal change in accuracy. With these findings we see that our Uber model is slightly more accurate than Lyft with this method.

Random Forests:

The second machine learning algorithm we can use are random forests. With this method we can see the importance of how each feature affects our model and predictions.

We can reuse the training and testing datasets from earlier, and using 10 decision trees, we find the following:

Without dummies
Lyft Mean absolute error: 8.45 degrees.
Uber Mean absolute error: 6.71 degrees.
Lyft Accuracy: 28.75 %.
Uber Accuracy: 50.42 %.

With dummies
Lyft Mean absolute error: 8.46 degrees.
Uber Mean absolute error: 6.72 degrees.
Lyft Accuracy: 28.85 %.
Uber Accuracy: 50.4 %.

This time again, it seems like Uber accuracy is higher with this method

Feature importance without dummy variables
Lyft:
Variable: surge_multiplier Importance: 1.0

Uber:
Variable: distance Importance: 1.0

These make sense due to them being the only factors being used, we can look at results including dummy variables below.

With dummy variables
Lyft:
Variable: distance Importance: 0.87862
Variable: destination_Financial District Importance: 0.01224
Variable: source_Boston University Importance: 0.01208
Variable: source_Northeastern University Importance: 0.01039
Variable: destination_Haymarket Square Importance: 0.00875
Variable: source_North Station Importance: 0.00737
Variable: source_Theatre District Importance: 0.00734
Variable: source_West End Importance: 0.00689
Variable: source_Fenway Importance: 0.00687
Variable: destination_North Station Importance: 0.00559
Variable: destination_Boston University Importance: 0.00519
Variable: destination_South Station Importance: 0.00509
Variable: destination_Fenway Importance: 0.00474
Variable: destination_North End Importance: 0.00474
Variable: source_South Station Importance: 0.00387
Variable: destination_West End Importance: 0.00342
Variable: destination_Theatre District Importance: 0.00336
Variable: destination_Northeastern University Importance: 0.00329
Variable: source_Beacon Hill Importance: 0.00266
Variable: source_Financial District Importance: 0.00234
Variable: source_North End Importance: 0.00198
Variable: destination_Beacon Hill Importance: 0.0016
Variable: source_Haymarket Square Importance: 0.0016

Uber:

Variable: distance Importance: 0.95012
Variable: source_Beacon Hill Importance: 0.00617
Variable: destination_Beacon Hill Importance: 0.0043
Variable: source_North End Importance: 0.00354
Variable: destination_Financial District Importance: 0.0035
Variable: source_Northeastern University Importance: 0.00314
Variable: source_Financial District Importance: 0.003
Variable: destination_South Station Importance: 0.00296
Variable: source_Boston University Importance: 0.00283
Variable: source_Fenway Importance: 0.00254
Variable: destination_North Station Importance: 0.00228
Variable: destination_North End Importance: 0.00223
Variable: destination_Theatre District Importance: 0.00184
Variable: destination_West End Importance: 0.00156
Variable: source_South Station Importance: 0.00147
Variable: destination_Haymarket Square Importance: 0.00146
Variable: source_West End Importance: 0.00124
Variable: source_Theatre District Importance: 0.00114
Variable: source_Haymarket Square Importance: 0.00112
Variable: destination_Northeastern University Importance: 0.00106
Variable: destination_Fenway Importance: 0.00095
Variable: destination_Boston University Importance: 0.00085
Variable: source_North Station Importance: 0.00069

As we see here the individual locations do not have much importance, and the biggest factor is the distance.

Final Conclusions:

In our initial glance at the data, we predicted that Uber prices would be cheaper than those of Lyft. Variables we used to check this data were distance, and location. Regardless of these factors, Lyft prices were comparably more expensive than Uber rides. We also took a look at the different services that each of these companies offered and compared them relative to each other. Generally, the cheaper rides have similar pricing, but when looking at the more luxurious and expensive options for each company, Lyft shows a drastic increase in price over Uber.

We attempted to use machine learning algorithms, namely linear regression and random forests, to see if we could predict Uber and Lyft prices given distance and location. Our tests showed that distance is the biggest factor in determining ride prices, with location being a very small and negligible factor.

There are still other factors and variables which can be considered which we had not covered in this project. Some interesting things to consider moving forward would be to include

the weather data that was initially provided but was not used (how much does rain affect pricing?), as well as comparing side by side the different types of rides that each service provides.