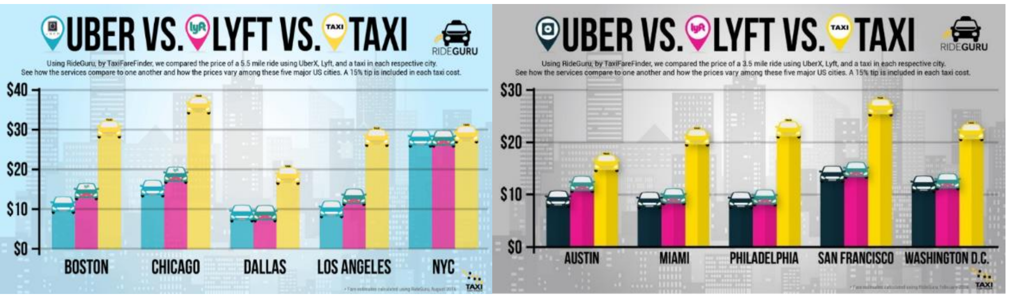
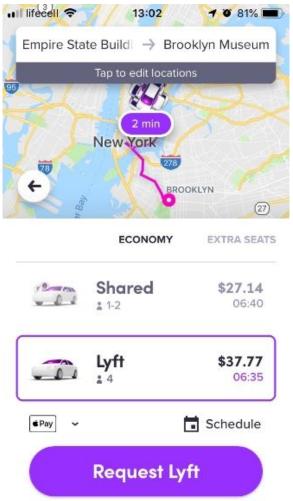
Predicting the Price of an Uber/Lyft

By Devesh Gokalgandhi

Uber and Lyft have dominated the ride-sharing industry in recent times, with combined over a 100 millions active users around the world. You can book rides easily from your phone, selecting the pickup and dropoff location, selecting the type of ride (whether it's a four-seater, six-seater, luxury car, SUV, etc.), and tracking the car to see when the driver arrived. This beats having to call a taxi service, and not knowing when the taxi would arrive from pickup. But a major factor that drove demand for Uber and Lyft over taxi services has been the price of the ride.

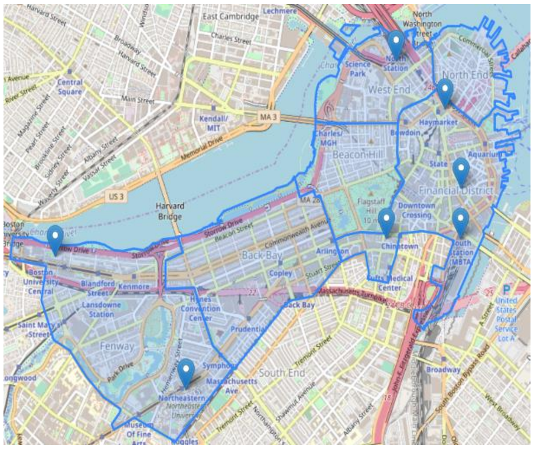


According to RideGuru, Uber and Lyft provide a significantly cheaper alternative to standard yellow taxis/cabs in many major cities in the United States. But what exactly goes into determining the price of these rides by Uber and Lyft? Is it just distance? Does the time of pickup matter? Does the weather affect the demand of rides, leading to an increase in price? I employed machine learning techniques on previous ride data to determine what exactly goes into Uber and Lyft's pricing models to estimate prices for future rides and give some interesting insights for you, the rider, to take into consideration when using these two ride-sharing apps.



The Data

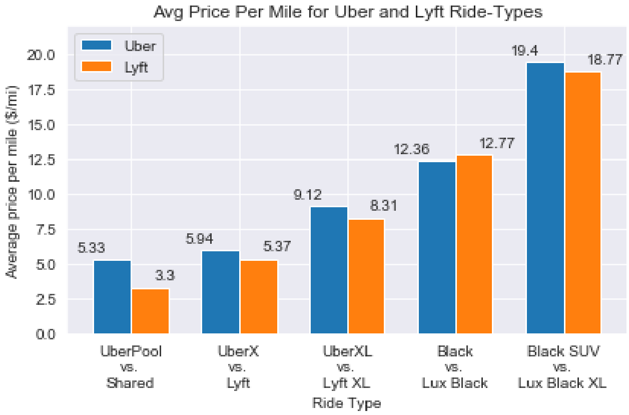
Over 690000 ride records were extracted, taking place in Boston from November 26 to December 18, 2018, using real-time Uber and Lyft API queries along with weather data taken from an open Boston database. Rides were either started or finished in six major Boston neighborhoods (highlighted in blue below): Fenway, Back Bay, Beacon Hill, West End, North End, and Downtown. Within these areas are seven sub-neighborhoods (map markers labeled below) where rides were picked-up and dropped-off: Boston University and Northeastern University (in Fenway), North Station (in North End), Haymarket Square, Financial District, South Station, and Theatre District (in Downtown).



Along with the ride features and weather features, an additional feature representing the occurrence of a Boston Celtics game or Boston Bruins game was added to measure the effect of sports game on the prices of rides going to and coming from the game's stadium (TD Garden, north of North Station).

What does all this data tell us?

Uber's selection of rides is generally more expensive than Lyft's

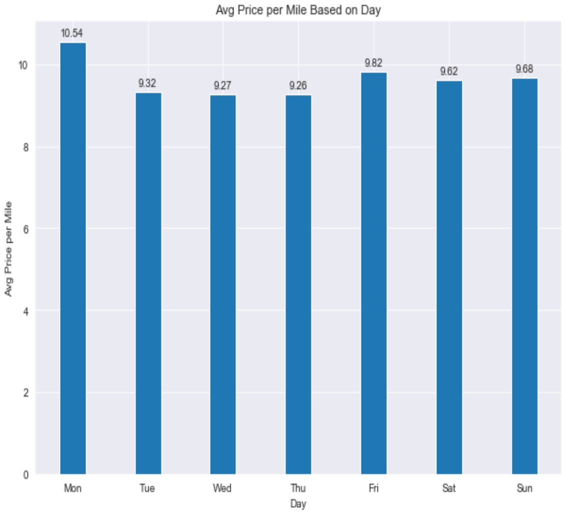


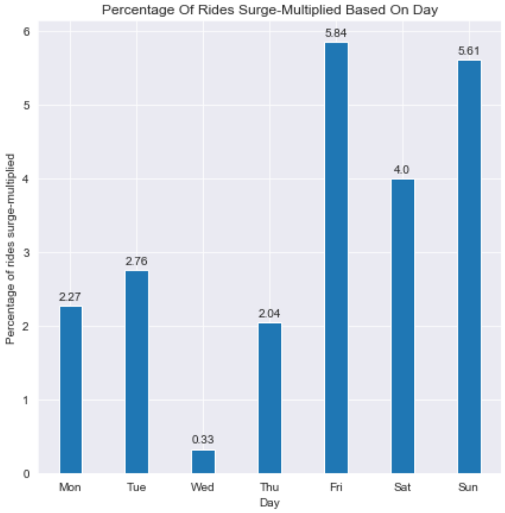
When we compare similar Uber and Lyft ride-types and their average price per mile, we can make the following insights:

- If you would like to share your ride with other riders, for up to two passengers, you should choose to take Lyft Shared over UberPool, as there is a $2.00 difference per mile. This difference can really add up!
- If you would like to request your own ride, for up to four passengers, you should choose a Lyft over UberX, because of a $0.57 difference per mile, or if you would like a more expensive premium black car selection, you should choose Lyft's Lux Black over Uber Black, due a difference of $0.41.
- When it comes to SUV service, with space for up to six passengers, Lyft's selections are much cheaper than Uber's, as Lyft XL and Lyft Black XL provide better value, for a difference of $0.81 and $0.63 respectively, over UberXL and Uber Black SUV. Take this into consideration if you have a big party!

Even though in aggregate, the average price per mile for Uber and Lyft is the same ($0.01 difference), four of the five ride-types provided by Lyft are recommended over their Uber counterpart, as the difference in cost is substantial, especially for rides over many miles.

Monday Woes and Long Weekends

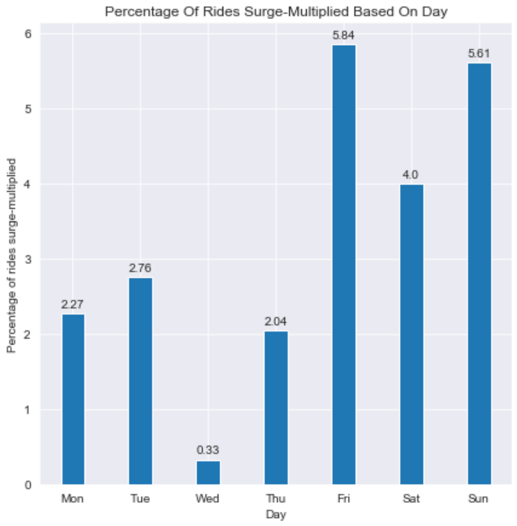


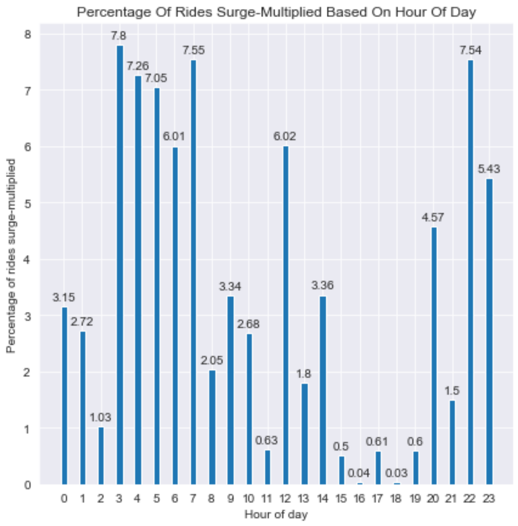


Looking at rides occurring each day of the week, rides taken on Mondays, on average, are significantly more expensive per mile than any other day of the week (10% increase!). Again, this is a substantial cost, as it is average price *per mile.* On such a crucial workday of the week, it recommended for you, who may be considering cheaper alternatives, to take any available public transportation or metro line going to and coming from your office, if it's not an inconvenience.

Rides taken on the weekends are generally more expensive than rides during weekdays, except for Monday. As many people are off from work and want to travel more, the demand of these rides increases, leading to **surged rides**, with applies a multiplier (greater than 1) to your ride cost. The chance that your ride may be surged, and cost higher that normal, is twice as much on the weekends than on the weekdays, as shown on right barplot above. Consider taking advantage of Thursday rides to go out more!

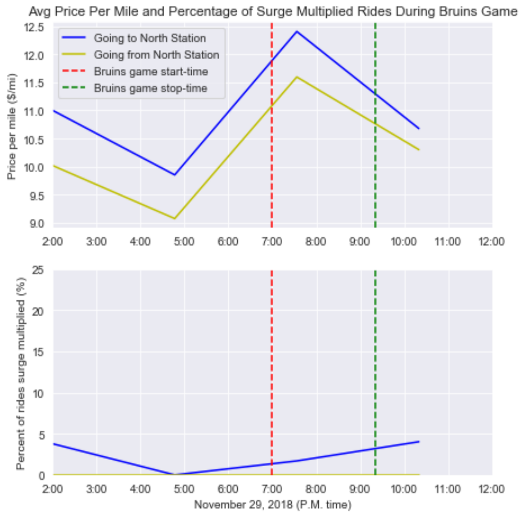
Bar Hoppers Beware!

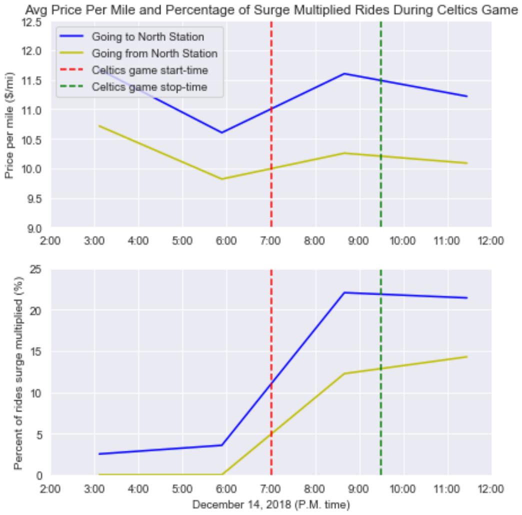




When we observe rides during each hour of day, the average price per mile gradually increases as the hours get later and later, and earlier and earlier. Rides taken at such early hours such as 5:00 A.M. and 8:00 A.M. are substantially more expensive per mile than any other hour. This holds true to even the number of rides that are surged, as hours from 10:00-11:00 P.M. and 3:00-7:00 A.M. see more rides charged higher than normal, than any other hour of the day. As the night progresses, the demand for rides combined with the lack of drivers will cause this surge in pricing. You, who may want to travel to areas full of nightlife/bars/clubs, should be aware of when you may want to get picked-up (before 10:00 P.M.), and avoid having to pay more than necessary. You may also want to leave earlier than when these bars and clubs usually close (before 3:00 A.M), to avoid paying more than usual.

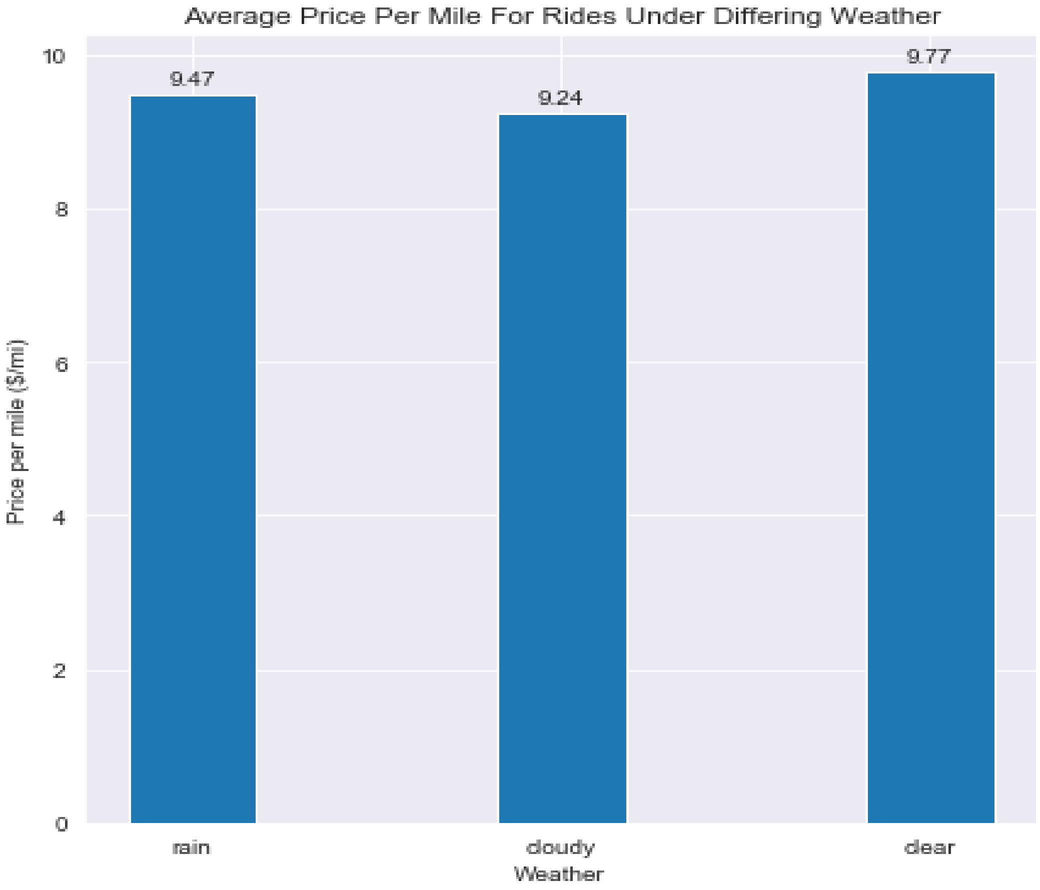
Come Early to the Games





We now look at the effect of a sports game, an event that drives many people to a single location, on the price of an Uber or Lyft. Clearly, rides going to and coming from North Station (location of TD Garden) generally cost less per mile and are not surged to the same extent a few hours before the game begins compared to rides occurring during and after the game. Since the concentration of potential riders, who did not transport themselves to the game, all require either ride-services or public transportation, the demand for Uber/Lyft would increase substantially. This explains why more rides are surged, and cost more per mile, after the game, rather than before. Come at least 2 hours before the game starts and consider taking the metro back home to avoid paying more than usual, unless the price is not surged.

Break out the Umbrella



Finally, we observe the price per mile under differing weather conditions, that are prominent in the Boston area. Weather plays a role in the price of a ride. Surprisingly, rides were 26 cents a mile *cheaper* during rainy and 53 cents a mile *cheaper* during cloudy weather than clear conditions. It makes sense, as more people would be inclined to remain indoors and travel less during cold and rainy conditions. However, take advantage of this! It may be better to travel downtown to your favorite restaurant, or theatre, or bar, and push through the rainy weather, as your ride may be cheaper than usual, and will save you a few bucks!

Predicting price using Regression

Four regression algorithms were set up to determine the best predictive model that would be able to predict the price of a ride, as close to the actual price of the ride. These algorithms were:

1. Multivariate Linear Regression
2. Log-Linear Regression: applying the log function on the price, and then applying the inverse function to predicted value to generate real dollar values.
3. Lasso Regression
4. Ridge Regression

The metrics used to compare each model and measure the accuracy of the model are:

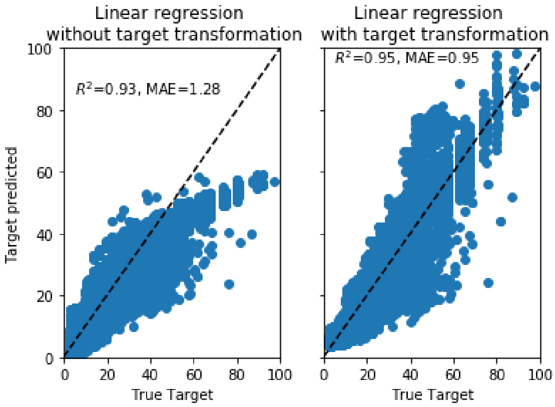
1. R-squared: measure how far the predicted prices are from the actual price
2. Mean Absolute Error: the average error in a set of predictions ($)
3. Mean Absolute Percentage Error: the average percent error in a set of predictions ($)
4. Root Mean Square Error: measure of prediction accuracy

The data is split into training (80% of total data) and test set (20% of total data), where the training data is fit into each regression model, and the test data is evaluated using the fitted models. The results of this evaluation are as follows:

Model	R-Squared	MAP	MAPE	RMSE
Linear Regression	.931	$1.71	12.95%	2.47
Log-Linear Regression	.945	$1.40	9.54%	2.19
*Lasso Regression	.945	$1.40	9.54%	2.19
Ridge Regression	.945	$1.40	9.54%	2.19

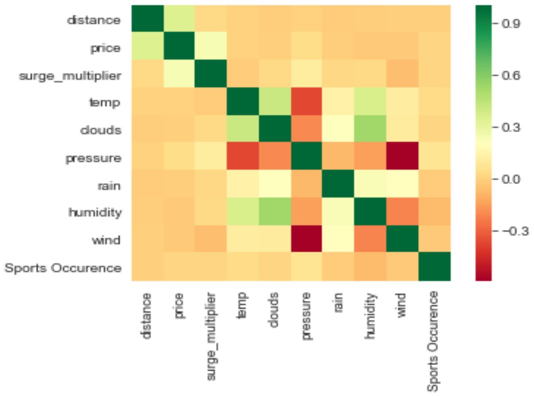
*For Lasso, the model improved for lower and lower alpha values. The best alpha value for evaluating the test data using lasso, was 0, which is the same as normal linear regression

Three of the four models produced the exact same metrics when predicting the price of an Uber/Lyft ride from the test data, with an error of $1.40, or 9.5%, within the actual price. Because of this, we choose the simplest model of the three, log-linear regression, as the model of choice for predicting price on unseen future data. Log-linear regression captures the non-linearity between the independent features and the price, as shown when fitting the training data:



Normal linear regression underpredicts significantly, capping predicted prices under $60, for rides that cost near $100. When we apply a logarithmic transformation on the price, the model can predict much more accurately, and capture these higher priced rides.

Feature Importance



Heatmap of Correlations. Greener shade for positive correlation, redder shade for negative correlation

The top five independent features most important, or most correlated, to the price (according to the regression coefficients) are:

1. Ride-Type (UberPool, Lux Black, etc.)
2. Surge-Multiplier
3. Distance

4. Cab-Type (Uber or Lyft)
5. Rain

Two features in the dataset, humidity and Sport's Occurrence, were found to not be statistically significant, towards the prediction of ride price. Removal of these two features caused the model to remain unchanged, it it's predictive power of test data.

Biggest Takeaways

- Using more than 600,000 actual Uber and Lyft rides from Boston in 2018, I built a model capable of predicting ride price to within $1.40 (9.5%) of the actual price on average. This can help riders or competitors better understand Uber and Lyft's pricing models
- Uber's selection of rides is on average more expensive than taking a similar type of ride from Lyft. This was true for all ride-types (UberPool, UberX, UberXL, Uber Black SUV) except Uber Black, which was cheaper than Lyft Black.
- Rides on average were more expensive and more likely to have a surge multiplier during late night hours and on weekends (Fri, Sat, Sun). Bar-hoppers, beware!
- Rides starting or ending near professional sports stadiums (Celtics or Bruins games), on average, were more expensive during and after the game, than before the start of the game. Sports fans should consider taking Uber/Lyft to the game and the Metro home. Weather plays a role in the price of a ride. Surprisingly, rides were 26 cents a mile *cheaper* during rainy and 53 cents a mile cheaper during cloudy weather than clear conditions.
- To build the model, tried a variety of regression techniques, including Linear, Lasso, Ridge, and Log-Linear (log-transformation of dependent variable). Ultimately, Log-Linear regression produced the most accurate model because of the non-linear relationship between ride price and independent features.

Hopefully, more complex models can be implemented, such as regression trees, gradient boosting, or random forest, to reduce the error as much as possible. Data from all major cities should also be included, as well as, other features such as ride duration, traffic data, etc.

To get more involved, check out my repository for this project, to get access to the data and IPython notebooks of how the wrangling, analysis and modeling took place. Hopefully, you, as the rider, can use these insights to approach using these ride-sharing applications more frugally, and avoid having to pay ridiculous prices to get to your destination.

Devesh Gokalgandhi is a student at Springboard's Data Science Career Track. His Linkedin and Github can be found here: https://www.linkedin.com/in/devesh-gokalgandhi-a20b3b123/, https://github.com/dgokalga/Springboard-Data-Science