FINAL REPORT
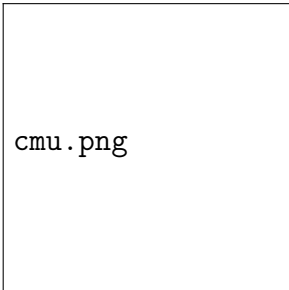
2017 CMU DATATHON

An Analysis of Uber's Disruption of the VFH Market

Authors:

Narain KRISHNAMURTHY
Govind WARRIER
Eric YI

March 25, 2017

cmu.png

1   Introduction

For decades, taxi companies have held monopolistic control over the for-hire vehicle (FHV) market. The high fixed capital costs of scaling an FHV business and local regulatory barriers to entry helped secure the strong monopoly right enjoyed by regional taxi companies. However, Uber launched in 2009, indicating the dawn of a new era in the for-hire vehicle (FHV) market. Technologically-driven competitors like Uber sought to break into the market and rapidly scale, leveraging lower unit costs primarily achieved through automation of the dispatch process. While the threat to incumbents posed by Uber and its peers is clear, it is uncertain how much incumbents have been already affected.

In order to gauge this effect, we focus our study on the effects of Uber ridership on ridership in Yellow Cabs and Green Taxi Cabs. Our question is: Did Uber prove to be a significant impact on Yellow Cab rideship and Green Cab ridership across the NTA regions of New York?

# 2   Non-Technical Summary

2.1   Uber Ridership Visualized
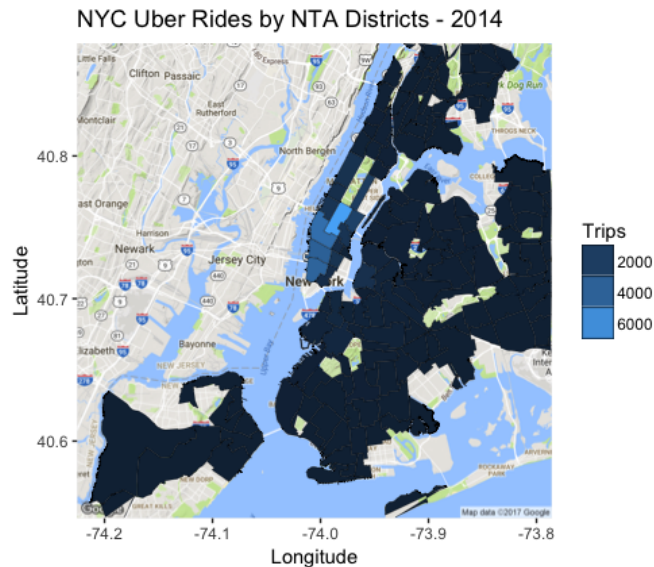
The below graph shows Uber ridership by region for a random sample of

50,000 rides in 2014. The regions with the most rides were in Manhattan, owing no doubt to the region's large population and high population density. The service did however have significant coverage across the city, with all nta areas having Uber ridership.


NYC Uber Rides by NTA Districts - 2015
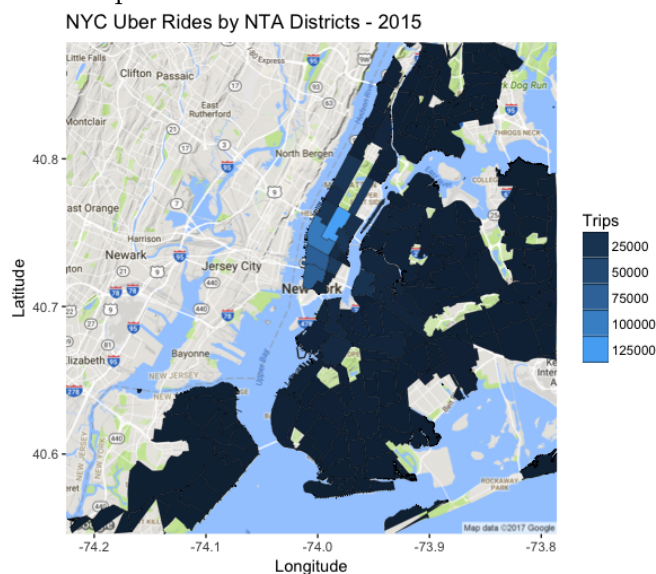
We now examine ridership in 2015 with a larger sample of 1 million rides. The distribution of rides looks almost identical to the 2014 distribution, with Manhattan accounting for the majority of rides and significant coverage over the nta areas of New York.

2.2 Yellow Cab Ridership Visualized

NYC Yellow Medallion Taxi Rides by NTA Districts - 2014
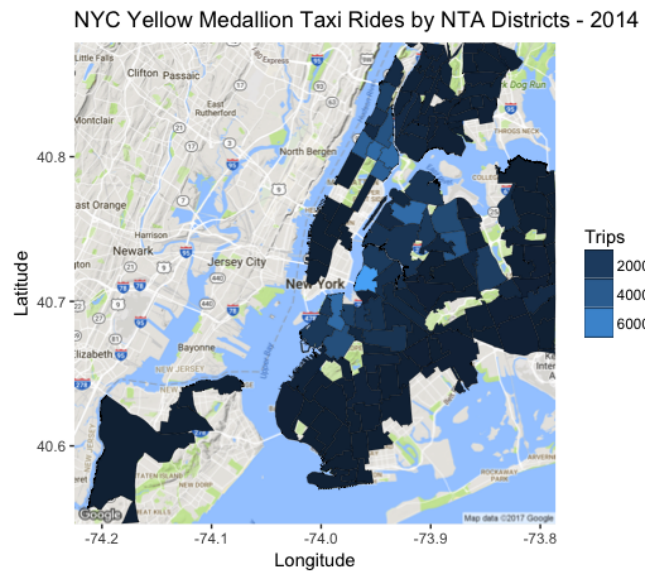
The following is the distribution of ridership for 50,000 Yellow Cab rides taken in 2014. Yellow Cab historically is most prominent in Manahttan while Green Cab is historically most prominent in the outer boroughs. Despite this, a significant portion of the high ridership regions for Yellow Cab are outside Manhattan, most notably in Queens and Brooklyn. In addition, Harlem accounts for a significant portion of ridership while the downtown regions have low ridership. This is the opposite of the distribution of Uber rides, which showed the highest ridership in downtown and low ridership in Harlem.

NYC Yellow Medallion Taxi Rides by NTA Districts - 2015
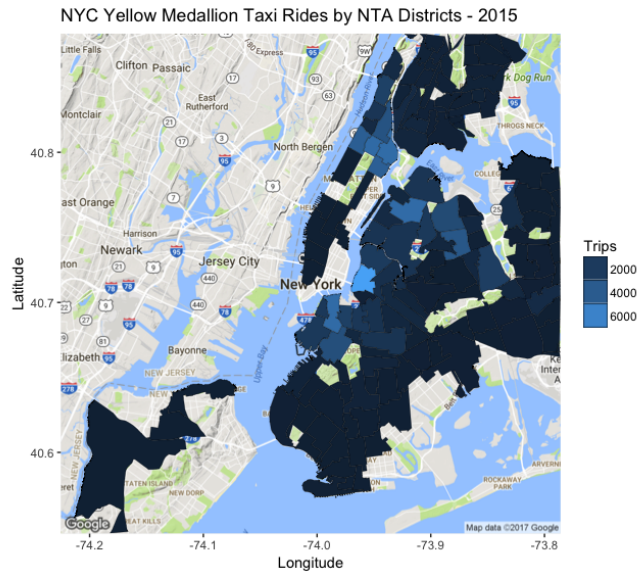
This is the distribution of a sample of 50,000 Yellow Cab rides in 2015. The same exact trends play out as in the 2014 sample: high ridership in Queens, Brooklyn, and Harlem, low ridership in Downtown.

2.3   Green Cab Ridership Visualized

NYC Green Boro Taxi Rides by NTA Districts - 2014-15
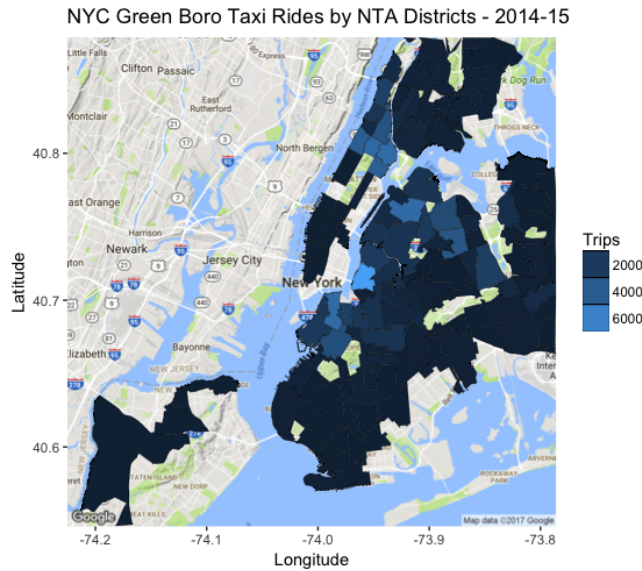
We now show the distribution of a sample of 100,000 Green Cab rides that

took place during 2014-2015. Being historically associated with travel in the Outer Boroughs, it's no surprise Brooklyn and Queens contain multiple high ridership regions. What's as interesting is that Harlem also has many high ridership regions for Green Cab. The distribution of rides is fairly similar to that for Yellow Cab.

2.4   Conclusion

It appears that the amount of Uber fares in a given NAT region on a particular day is an important factor in predicting the corresponding number of fares for both yellow medallion taxis and green boro taxis. This is particularly interesting since both companies are largely concentrated in regions where Uber is not highly prevalent. This suggests that although the taxi companies are avoiding Uber-heavy areas, they are still under significant competitive pressure from Uber. This effect would probably be much more exaggerated if the taxi companies attempted to compete with Uber where Uber is most concentrated (Manhattan).

3   Technical Summary

For each region, we model ridership for each day in 2014-2015 as a function of Uber ridership on that day as well as well as the region's demographics. We fit both a rule-ensemble model and logistic lasso model to predict Green Cab ridership and Yellow Cab ridership.

3.1 Green Cab
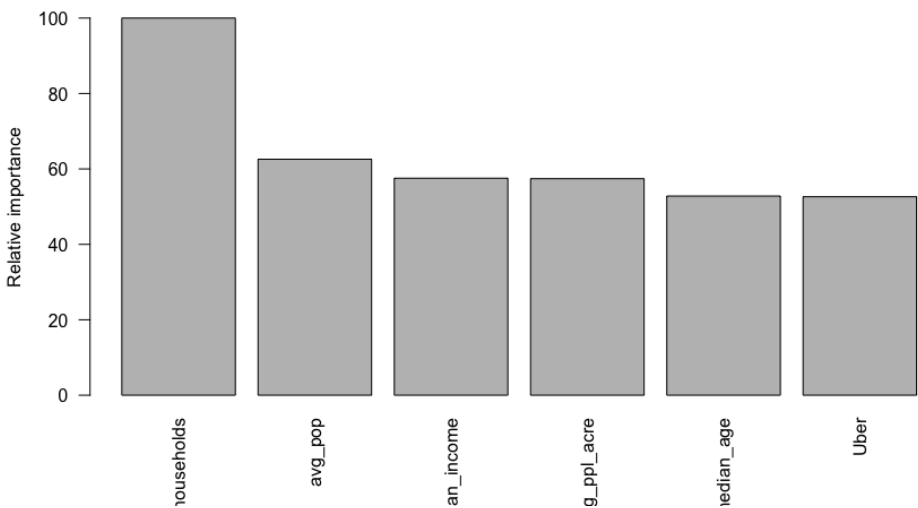
The cross-validated test error for the Rule Ensemble was 50% better than for the lasso regression. Thus, we decided to use the Rule Ensemble method to determine the predictive power of each variable. Thus, according to the model, Uber proved to be a significant factor on ridership of Green cabs. We include
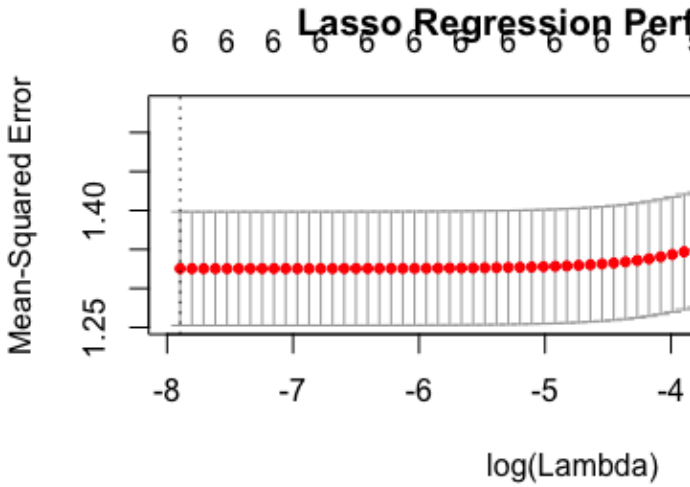
the MSE of the Lasso Method for completeness.

3.2 Yellow Cab

The lasso method proved to be a very poor fit for modeling Yellow Cab ridership, so we decided not to continue using or fitting it. The rule ensemble method proved to have similarly good performance on the Yellow Cab data as on the Green Cab data. We show significance of factors below from the model:
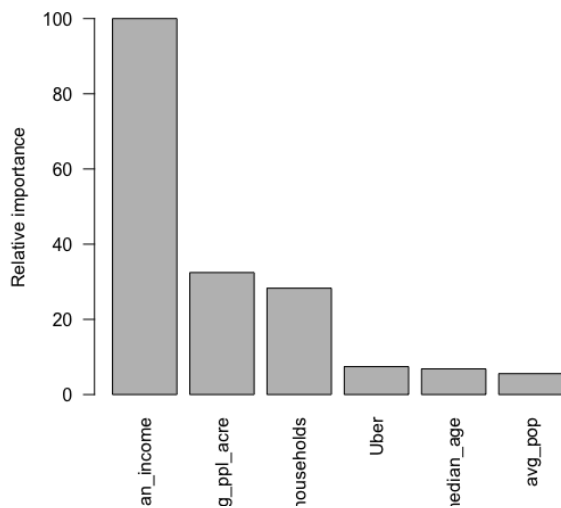
For Yellow Cabs Uber proved to be more significant than two of our demographic factors indicating that Uber did indeed have a significant impact on the ridership of Yellow Cabs.

3.3 Conclusion

In conclusion, we found that a rule-based ensemble outperforms regression with LASSO penalization for both the yellow medallion cabs and green boro cabs. Further, we find that the frequency of Uber trips appears to be very important in predicting the frequency of fares for the two taxi companies. This effect is particularly interesting due to our visualizations of the geographic concentration of Uber and the taxi companies' businesses.

4 Technical Issues

4.0.1 Determining NTA Region for Rides

One major problem we encountered with the ride data sets (for all three services) is that the individual rides didn't have a tag indicating the NTA region where the ride began. The rides did have lat/long coordinate pairs, so we used the point in polygon method to determine for each ride which NTA region it belonged to. The method is relatively slow, so this impacted the number of data points we could use when fitting our model. The geographic.csv data set provided vertices for the boundaries of each NTA region. These boundaries contained thousands of points for each region. Adding more boundary points increased the runtime of the point in a polygon method, so we had to decide on a limit for the number of vertices in each boundary. After testing at log-scaled intervals (e.g. 10, 100, 1000) we found that having a number of vertices in the thousands did not increase the run time much more than hand a number of vertices in the tens. In fact, the difference on checking 5,000 points for 1000 vertices vs 10 vertices was only about 20% (i.e. the 1000 vertices version took only 20% longer to fit). Thus, we used 1000 vertices in our boundaries.

4.0.2 Disporportionate Data Sizes

The datasets for the different types of rides had vastly different sizes. We used stratified sampling to account for this.

4.1 Tables and Figures

Use the table and tabular commands for basic tables — see Table **??**, for example. You can upload a figure (JPEG, PNG or PDF) using the files menu. To include it in your document, use the includegraphics command as in the code for Figure **??** below.

4.2 Mathematics

LaTeX is great at typesetting mathematics. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with $E[X_i] = \mu$

and $\mathrm{Var}[X_i] = \sigma^2 < \infty$, and let

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_i^n X_i$$

denote their mean. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$.

4.3   Lists

You can make lists with automatic numbering ...

1. Like this,

2. and like this.

... or bullet points ...

- Like this,

- and like this.

We hope you find writeLATEX useful, and please let us know if you have any feedback using the help menu above.