Final Project Proposal: Applied Machine Learning

Laura Goyeneche[1] Nathan Deron[2] David Contreras[3]

October 2019

1 Proposed Analysis

The ride hailing industry revolutionized the way in which citizens transported during the first two decades of the 21st century. It is an industry with a global market size of $24.4 billion[4] with particular impact on low skill labor, as many find platforms like Uber or Lyft an alternative to conventional work. On the other side, traditional individual transportation operators (i.e. taxis) engage in political, legal and economic efforts to block these initiatives in cities across the world[5].

In this work, we will use ride and driver data from both taxis and ride hailing companies, as well as neighborhood characteristics data in the city of Chicago. The aim of the work is to find any possible pattern of discrimination between services provided by ride hailing companies and/or taxis and the different neighborhoods in Chicago, particularly based on their level of crime, poverty, presence of minority, etc. In that sense, the expected outcome of this work is a classification model that classify if the ride is made by ride hailing or taxis based on the characteristics of the departure and destination neighborhood. The classification model will allow us to identify variable importance during the prediction. That is, we are going to identify if variable like crime rates, education levels, etc makes a person more likely to ride in apps like Uber or Lyft, rather than take a taxi. Additionally we will predict the difference in cost of a ride using each service and the importance of departure and destination characteristics.

As a consequence, this study seeks to find possible market segmentation between taxis and ride hailing companies, which may be biased against any particular minority. That is the novelty of our work: comparing taxi vs ride hailing companies to identify possible bias in the market segmentation (if it exist). In this last issue, at least two different studies have been made to identify the possible bias against minorities. According to Yanbo Ge et.al. (2016), peer transportation systems seems to have longer waiting times and cancellation rates against those names related to African American communities[6]. On the contrary, a study made by Anne Brown (2018) found that ride hailing companies created access for those previously marginalized by taxis companies and she didn't find any particular difference in terms of quality service for women or minorities[7].

2 Methodology

Four group of data sets will be combined for the analysis. The first group consist of ride-share and taxi trip records uploaded by City of Chicago – Data Portal. Both data sets have information about each trip distance, total fare including tips and extras, payments type, company, pickup and drop-off time and spatial variables. The second one, contain the incidents of crime reported by the Chicago Police Department and characterized by primary type, location, and spatial features. The third group provides information of socioeconomic indicators across community areas, including the percent of housing crowded, household below poverty, individuals aged 16+ unemployed and income per capita. The last group corresponds to the number of police stations, parks and recreation facilities across the city provided by City of Chicago – Data Portal.

Table 1 presents a summary of the primary data sets require for the analysis. As observed, taxi and ride share trips records go up to 187 and 101 million observations dating back to 2013. However, for data manipulation purposes, we limit our observations to two months period (September 01 to October 31, 2019). In addition, in order to align all time variant data sets, we also selected the incidents of crime in the same time period.

Pre-processing will be needed to combine the taxi, ride-sharing, and neighborhood characteristics data into one usable set. Both the taxi and ride-sharing data sets will need to be processed for null values, specifically for missing location data. Census tract data can be imputed based on neighborhood information.

Given the multiple data sets, first we will evaluate whether to exclude and impute variables with missingness in each data set. For instance, 16 out 21 features of interest in ride sharing trips have between $< 1\%$ and 36% of missing values,

Table 1: Data sets characteristics

	Unit	Total Obs.	Project Obs.[8]	Total columns	Columns NA	Percent NA[9]
Taxi trips	Trips	187 M	1.3 M	21	16	< 1 % - 35 %
Ride-share trips	Trips	101 M	8.9 M	21	9	6.5 % - 30 %
Crimes	Crimes	7 M	35,494	22	6	< 1%
Socioeconomic indicators	Community area	77	77	9	-	-
Police Stations	District	25	25	15	-	-
Parks	Parks	581	581	75	-	-

Note: M refers to million observations

of which half present less than 1% of NA values. Variables with high missingness are spatial characteristics such as latitude and longitude, which can be imputed from other present spatial characteristics. Second, we intend to join them guaranteeing that the unit of analysis is ride trips. For that, we propose three stages in our data pre-processing: aggregate taxi and ride share trips into one normalized data, calculate and impute features based on characteristics of pickup and drop-off trip location information, and merge all the data sets properly.

This analysis will look to predict the geography-specific likelihood of a public taxi or a ride-sharing service picking up a given passenger, and the geography-specific cost for each of those services. We hope to compare cost differences between the two types of services in relation to crime statistics in the pickup and drop-off areas.

The empirical analysis is based on predicting the likelihood of a public taxi or a ride-sharing trip given the geographic-specific crime risk. Formally, our outcome (Y_i) is one for ride-sharing trips and zero for taxi trips. We will determine the importance of the number or proportion of crimes around the pickup and/or drop-off location (C_i), and a group of covariates such as location (police beats, community area or district) and city characteristics (X_i) on this outcome, as well as on another set of outcomes: $Price_t$ and $Price_r$. Note that the sample of interest is limited to trips in Chicago in a specific period (W).

To achieve this we will utilize logistic regression, regularized regression, gradient boosting, and neural networks. After comparing outcomes we will choose the best method for each of our two tasks. We will use ROC plots and AUC to determine the usefulness of our taxi/ride-sharing pick-up predictions, regularized regression for variable selection and analysis of variable important to determine if crime is important in predicting whether an individual uses a taxi or a ride share service.

Between combining our data sets, imputing missing values, fitting multiple models for two different machine learning tasks, then choosing and tuning our model, there is significant work to complete. In addition to the interpretation, this provides sufficient work for the final project.

3 Limitations and Use

Limitations arise from the use of regularized regression to determine variable importance, as these methods can prove suspect. Since we do not have access to ride-sharing algorithms we cannot be sure what causes fluctuations in prices. Additionally there is difficulty determining causality here, where there may only exist a correlation between crime, prices, and type of transportation access. However these limitations do not sufficiently overshadow the potential benefits of this work.

This pipeline could be used by regulators looking to increase competition between taxi and ride-sharing services and protect equal access to transportation across geographic areas. For instance, a regulator in the City of Chicago may input their new taxi and ride-sharing data into the pipeline to find that in a particular neighborhood, ride-sharing drivers are disproportionately more likely to pick up riders and they charge a higher price due to crime in the area. This could inform expanding other transportation options in the neighborhood.

Notes

[1] lgoyenec
[2] nderon

[3] dlcontre

[4] Marketwatch. *At 19.8% CAGR, Ride Sharing Market Size is Expected to Exhibit 103600 million US$ by 2025.* Available in `https://www.marketwatch.com/press-release/at-198-cagr-ride-sharing-market-size-is-expected-to-exhibit-103600-million-us-by-2025-2019-03-15`. Retrieved on 10/27/2019.

[5] Levi, Mark. *Lyfts, Ubers may gain ground in war on taxis, getting access to cab stands to pick up riders.* Available on `http://www.cambridgeday.com/2019/02/12/lyfts-ubers-may-gain-ground-in-war-on-taxis-getting-access-to-cab-stands-to-pick-up-riders/` Retrieved on 10/27/2019.

[6] Yanbo Ge, Christopher R. Knittel, Don MacKenzie, and Stephen Zoepf. *Racial and Gender Discrimination in Transportation Network Companies.* NBER Working Paper No. 22776. October 2016. Available on `https://economics.stanford.edu/sites/g/files/sbiybj9386/f/zoepf.pdf` Retrieved on 10/28/2019.

[7] Anne E. Brown. *Ridehail Revolution: Ridehail Travel and Equity in Los Angeles.* University of California, Los Angeles. Available on `https://escholarship.org/uc/item/4r22m57k` Retrieved on 10/28/2019.

[8] Due to limitations manipulating the data, the present projects will use taxi and ride share trips, and crime incidents records from September 01 to October 31, 2019.

[9] Variables with high percentage of NA in both taxi and ride share trips correspond location variables like latitude and longitude. Other variables like distance (miles or seconds), fare, etc. have less than 1% of NA values