Chicago Taxi Rides

Proposal Paper

Noah Leuthaeuser
University of Colorado - Boulder
nole1337@colorado.edu

Heesuk Jang
University of Colorado - Boulder
heesuk.jang@colorado.edu

Joe Alsko
University of Colorado - Boulder
joal2716@colorado.edu

Rei Isobe
University of Colorado - Boulder
reis9668@colorado.edu

PROBLEM STATEMENT

The database that we will be using was taken from the Chicago Taxi Cabs which recorded the locations of pickup and drop off, time of day, duration, fees, method of payment and tips of each ride in 2016. This database can be used to predict how much tip each location offers, tips vs the duration of the trip, frequently traveled locations, etc. From these knowledge that we can gain from this database, we can combine it with the Census data of Chicago that we found to compare the frequency of routes taken to the income of those locations. This information could be used to help Cab companies to change their fees, such as lowering the initial pickup fee for locations with low income and low frequency to get more customers to take their cabs as their main method of transportation. Our hope is to find connections, such as frequency of the routes and number of pickups and drop-offs at each location, with the Census data to help taxi companies be more profitable by gaining more customers and strategizing their routes.

1 LITERATURE SURVEY

Some studies and surveys have indeed already been conducted on the Chicago taxi service, mostly by city transportation/commerce departments. Our group looked at three conducted by the City of Chicago Business Affairs and Consumer Protection office, the NYC Data Science Academy, and Todd W. Schneider, a Yale software engineer. These studies looked at pickup frequencies, rates, and pick up/drop off locations. Our study also aims to study these attributes; however, unlike these studies, we're looking to find relations between them.

The Chicago BACP study (2014) looked at taxi rates in relation to value to the consumer as well as fairness of income for taxi drivers. The purpose was to develop a model that the city could use to determine the effect of fare changes on taxi drivers' income. This model accounted for different taxi ownerships (part time/full time, lease/ownership). However, the study didn't take into account locations of taxi service and looked at the Chicago taxi service as a whole. Also, the study combined fares and tips into revenue, not looking at each individually [1].

The NYC Data Science Academy study (2016) was more in line with our study (and used our same dataset). It looked at locations in Chicago and the average taxi pickup frequencies at those locations. Another part of this study mapped the ratio of pickups to drop offs in each community as well as average trip ranges. The study found that the airports and central section of the city receive the most frequent pickups and have the highest ratio of pickups to drop-offs. The trip ranges from the airports were consistently high, while the trip ranges from the city center were consistently low. So, in every location besides the airports, trip range seems to increase as pickup frequency decreases [2].

Todd W. Schneider conducted a survey on a similar dataset, but from 2013 to 2016 instead of just one year. His study seemed to be more aimed at trends in the Chicago taxi service as a whole. The study found that the taxi business in Chicago is declining faster than New York City's (55% decrease since 2013). However, his study did look at pickup frequency in certain locations of Chicago as well. Schneider mapped taxis' percent chance of a pickup within 30 minutes to each community in Chicago. Strangely, the study suggests that the airports were areas of very low pickup frequency, completely contradictory to the NYC Data Science Academy study. The rest of the data seems consistent, but this anomaly will be something to focus on in our study. As a side not, his percentage based prediction model could serve as a useful metric in our analysis. Schneider's study was built on his previous analysis of New York City taxis [3].

2 PROPOSED WORK

2.1 Preprocessing

The taxi rides dataset has already undergone some cursory preprocessing before being released to the public. All of the changes made are documented in the press release from the City of Chicago [5]. Major changes include masking time, location, and taxi medallion number for privacy. Pickup and drop-off time is rounded to the nearest 15 minutes, and location is given to the accuracy of the census tract. Implausible values were removed from the data, including negative lengths or costs, or extremely long trips. Some duplicates were also removed.

To get the data into a workable state, we will also need to continue with preprocessing and cleaning. First, the twelve distinct months will be merged into a single dataset to evaluate the data for the entire year. Some features from the dataset are not needed for our purposes and will be dropped. Our analysis of pickup and drop-off location will be on the community area level, so the census tract and geolocation columns for pickup and drop-off location will be removed. The payment extras column is sparse and without specifics on what the extra payment is for, so it will be removed. The taxi ID column will not provide us with any meaningful insights, so it will also be removed.

Some trips recorded a 0 value both for trip length in miles and trip length in seconds. When the value is 0 for both columns, the row will be removed. When only one of the columns is 0, the value will be extrapolated using the prediction technique described in section 4. Tips are only recorded for credit card payments, so any analysis on tip amount will not include cash payments. Because tips are not recorded for cash payments, the fare column will be used for cost analysis rather than the total column.

The selected socioeconomic factors from our secondary dataset will be merged to the taxi rides dataset using a key of community area. Analysis and evaluation methods are described in depth in section 4.

2.2 Difference from prior work

The Chicago BACP study focused on overall Chicago taxi service value and income balance. The NYC Data Science Academy study focused on the relationship between pickup frequency and trip range [2]. Todd Schneider's study looked at the trend of decline in the Chicago taxi business overall as well as pickup frequencies in different locations. Unlike the BACP study - as well as Schneider's - we will focus on trends per community in Chicago [1,3]. Connecting pickup frequency and range per location will be important, but we will also be looking at average tips and payment trends in each location as well. Alongside this, we also aim to study correlations between these communities' taxi trends and their poverty ratings.

3 DATA SET

The primary dataset for this project is the Chicago Taxi Rides dataset provided on Kaggle by the City of Chicago [4]. The dataset includes information about every taxi ride taken within the city for the year of 2016. The features of this dataset include a start and end timestamp, trip length in seconds and miles, pickup and drop-off locations in the form of census tract, community area, and geolocation, payment information including fare, tip, tolls, extras, and trip total, and the taxi company. It is divided into 12 separate files, one for each month of 2016. Each month contains around 1.7 million taxi trips.

We will also use a secondary dataset containing socioeconomic information about the different areas in the city. This dataset is also provided by the City of Chicago [6], and is divided by the same community areas as the taxi rides dataset. The other features of this dataset include the community area name, percent of housing crowded, percent of households below poverty, unemployment information, per capita income, percent aged under 18 or over 64, and a hardship index.

4 EVALUATION METHODS

After we complete the pre-processing of the data in our dataset as described in the Proposed Work, we will apply several different pattern evaluation methods to mine possible patterns we are interested in. It is extremely important to select highly corrective measures to produce accurate and quality results, derived from understanding in depth the characteristics of the data such as the types of attributes and completeness, validity, accuracy, consistency, availability and timeliness of the data in our dataset.

First, we will apply the Prediction Model to fill in some of the unknown or missing numeric values in Trip Seconds or Trip Miles based on the valid Trip Start Timestamp and a pattern we discover between the Trip Seconds and Trip Miles. We can also determine any of the missing numeric codes in Pickup Community Area or Dropoff Community Area based off the given Trip Seconds and Trip Miles for each instance.

Then, using the January dataset as a sample at first trial and acknowledging the existence of null-transactions in the following events, we will use one of the Null Invariant Measures such as AllConf and Cosine to find out if each pair of events below are positively or negatively correlated:

- Trip Total Cost vs. Trip Seconds
- Trip Total Cost vs. Trip Miles

It is likely that the results from the application of predication model and correlation study above will provide us a better significance about how the time block, duration and distance of each trip impact on the trip total cost.

Next, we will apply the Apriori Algorithm using hash-based and pruning techniques to derive the frequent 2-itemsets of pickup and drop-off community area as well as the frequency of pickup in a specific time of a day and a year. We believe that these will be strongly perceptive indicators to justify the possible best or worst places and time blocks for a taxi to make a pickup and/or drop-off, thus again could use the outcomes to generate a more profitable business.

We will also look at how Per Capita Income by Each Community Area is correlated with the frequency of pickup and tips as a part of trip total cost. It is very important to perceive cash tips are not recorded mainly because they do not go through the payment systems. As a result, we will closely monitor if this does not result in a possibly skewed outcome.

Subsequently, we will do a Cluster Analysis to see how similar in the frequency of pickup and/or drop-off from one area to another and one time block to another. In order to measure the magnitude of similarities we will also apply one of the distance functions such as Euclidian or Manhattan Distance.

To make sure the selective methods we applied are consistently validated, we will lastly try to replicate some of statistics that are studied in the previous work for comparison.

5 TOOLS

5.1 Python3 and iPython

The rich features of the high performance in preprocessing, statistics, and graphics using iPython (interactive form of Python3) provided by Jupyter Notebook will allow us to more readily focus and solve domain problems, rapid-prototype code and more quickly and easily experiment with ideas. In this interactive notebook environment, Python 3, Pandas, Numpy, and Scipy will be used primarily for preprocessing, mining, and analysis of the data in our dataset. As data visualization tools, we will use Matplotlib and Seaborn for simple graphical representations and Bokeh and possibly Plotly for rich and interactive visualizations.

5.2 Git and Github

We will use Git and Github to manage source code. It will allow us to easily and seamlessly collaborate and keep track of the various changes through version control.

5.3 Slack

We will use Slack for better team communication, especially as a venue to share any resources such as ideas, documents, and tools that enable our team to produce quality results.

6 MILESTONES

We have a few project deadlines that we will use to make sure we are on the right path. These deadlines are: Part 3 is due April 10th, signups for the presentation spots are on April 19th, and the project will be due on May 1st. We will be meeting on every Monday's after class, as well as some Fridays. At each meeting on Mondays, we will have milestones to accomplish. For next Monday, March 12th, we will have the final decision on our data cleaning, taking out unnecessary attributes that we will not need for this project. By the following Monday, March 19th, we will apply the data cleaning to each file, which has the data for each month in 2016. After Spring Break, we will have a meeting on April 2nd, by then we will integrate all the months together and analyze the dataset. By April 9th, we will integrate the Census data of Chicago to compare the cab ride database and finish Part 3. By April 16, we will finish the code and start working on Part 5. By April 23rd, we will have project presentation slides finished and begin working on the final report. On April 30th, we will finish the Final report and the Peer Evaluation.

6 SUMMARY OF PEER REVIEW SESSION

During the peer review sessions, we received feedback about our goals and our work as related to other groups. We were advised to focus our goals to be more specific. They were fairly broad and we have shortened them into a few small goal statements. We were also suggested to work with another group that is looking at taxis in NYC, and we'll try and communicate with them going forward

REFERENCES

[1] Nelson/Nygaard Consulting. "Taxi Fare Rate Study." City of Chicago, www.cityofchicago.org/content/dam/city/depts/bacp/publicvehicleinfo/publicchauffer/chicagotaxifaresstudyaug2014.pdf.

[2] Wu, Yiming. " 2016 Chicago Cabs Analysis." NYC Data Science Academy, 18 Sept. 2017, nycdatascience.com/blog/r/2016-chicago-cabs-analysis/.

[3] Schneider, Todd W. "Chicago's Public Taxi Data." Todd W. Schneider, toddwschneider.com/posts/chicago-taxi-data

[4] City of Chicago, "Chicago Taxi Rides 2016." City of Chicago, https://www.kaggle.com/chicago/chicago-taxi-rides-2016

[5] City of Chicago, "Chicago Taxi Dataset Released." City of Chicago, 16 Nov. 2016, https://digital.cityofchicago.org/index.php/chicago-taxi-data-released/

[6] City of Chicago, "Census Data – Selected socioeconomic indicators in Chicago." City of Chicago, https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2/data