Mining For Rideshare Patterns

Saloni Sharma
CSPB Student
CU Boulder
Sash8251@colorado.edu

Ronald Thompson
CSPB Student
CU Boulder
Roth9236@colorado.edu

Jonathan Prindle
CSPB Student
CU Boulder
Jopr2193@colorado.edu

PROBLEM STATEMENT/MOTIVATION

The overall goal of our project is to utilize rideshare data to reveal interesting patterns that would improve the customer experience while simultaneously optimize profits for drivers. In particular, there are three major areas where we will focus our research.

First, we will perform a general analysis of rideshare patterns to better understand how customers use rideshare services. Specifically, we will concentrate on four central questions.

1. During peak traffic times, are there any notable patterns that can potentially ameliorate the customer experience?
2. What are the average durations of rideshare trips throughout the day and how does this affect the customer and driver experience?
3. What times are better and what times are worse to take a rideshare?
4. At what point do drivers hit diminishing returns?

In answering these questions, we will develop a comprehensive picture of the rideshare experience for both the customer and driver in aims of creating strategies to better their experiences.

Next, we will compare city specific rideshare patterns with general rideshare patterns. Principally, we will verify whether patterns with similar attribute values produce similar results. This examination will offer insights into the impact of rideshare on individual communities and countries as a whole.

Additionally, if we have extra time, we would like to explore questions in regard to census data. In particular we would like to examine if rideshare work and tips are related to the area in which a driver works. We would also like to examine which routes are frequented by customers and if they display some type of routinized pattern (e.g. ridesharing to work every day).

In summation, with the analysis of these questions, we hope to improve customer experiences and create a more enjoyable rideshare experience overall.

LITERATURE SURVEY

As rideshare has become increasingly interconnected with our lives, so too has the amount of research in building more effective methods for better performance. The following are a list of sources that investigate data in various directions. Some pose questions similar to our own while others venture to a different path. Each source has been provided with a summary.

1. *Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance* is an article by Todd Schneider that explores how rideshare has affected New York City. This article offers a thorough explanation of rideshare utilization by customers with topics including frequented areas of rideshare, comparisons between different types of rideshares, length of trip, and factors that affect rideshares such as weather.

2. *Uber pickups in New York City*. This is a data set that was explored by the site FiveThirtyEight. Broadly, FiveThirtyEight used the data for a variety of stories to describe how rideshare has been affecting the customer experience. In particular, most stories focused on whether or not

rideshares were helping or hurting the community at large.

3. *New York City taxi fare prediction*. This competition had participants predict the taxi fare between two locations.

4. *An empirical analysis of on-demand ride-sharing and traffic congestion*. This is an in-depth study from Arizona State University that investigates the social welfare impact of rideshare, specifically Uber, on their home city. Essentially, it outlines the benefits and drawback of rideshare.

5. *The Economics of Ride Hailing: Driver Revenue, Expenses and Taxes.* This research article from MIT provides a cost analysis of rideshare. The authors look at the profit margins for drivers of rideshare and whether or not the profit gain outweighs the costs incurred.

6. *Uber Movement*. This website tracks Uber movements and speed. It provides anonymized data such as locations and insights into how fast Uber can get from point a to point b at different times of the day in an effort to improve urban-planning.

PROPOSED WORK

In order to efficiently utilize the raw data we have collected, we will need to preprocess it through steps such as cleaning, integration.

First, we will clean our data sets. All missing values and any unnecessary (outlier) values will be removed as long as they are not a large percentage of the data set. To smooth the data further, we will also bin by fare in 2.50 increments as the Chicago data set only includes such values. We plan to keep any duplicate values as they can demonstrate frequent patterns.

Second, we will integrate our different data sets. Specifically, we will merge the 2014 and 2015 data from data set 2 together to create a single data set for Uber trips in New York City. We will also combine the monthly data files from 2015 from data set 3 to produce one data set. Then, we may integrate data set 2 with data set 3 to create one large set of data for NYC as well as standardize certain attribute values such that they can be better compared with values in dataset 1.

Third, we will transform and reduce the number of attributes in our data with dimension reducing algorithms such as PCA or with backward elimination. We will examine highly correlated values together to potentially remove or simply cluster attributes.

Overall, we will begin to work with smaller sample training sets and reserve test sets with data matching certain attributes such as month of year.

Our process will differ from the previous examples cited in the literature survey section in a variety of ways. We will be using more comprehensive data sets, which span over various regions in contrast with one specific area. The preprocessing and integration of our data will be more targeted to address our specific goals of customer experience improvement and profit optimization as well.

DATA SETS

1. **<u>Dataset 1</u>**: City of Chicago. This data set provides 129 million different data points with 21 categories. We have a variety of numeric attributes such as trip miles, trip seconds, binary attributes such as shared trip authorized, and interval attributes such as fares, which have been rounded to the nearest $2.50, and tips rounded to the nearest $1.00. Overall, this dataset provides comprehensive data and the relevant attributes in regard to the topics we are focusing on. City of Chicago Rideshare Data.

2. **<u>Dataset 2</u>:** Uber Pickups in New York City. This data set offers the nominal attribute of location (for pickups) in NYC and ratio interval attributes in the form of time. This dataset will allow us to track when and where Uber pickups occurred. This will allow us to measure the frequency of trips at different times. Uber Pickups

3. **Dataset 3:** NYC Taxi and Limousine Commission. This dataset offers various numeric attributes in regard to fare, tip, trip distance, etc. This dataset will allow us to examine rideshare rates, tips, and distances of the trip. In conjunction with dataset 2, we will be able to measure peak travel times and measure when travel times are less strenuous for passengers and more beneficial for drivers. NYC TLC Trip Data

EVALUATION METHODS

We will focus on a variety of ways to evaluate our data in order to ensure the best results.

First, we will compare our results with other existing analyses, specifically from those we cited in our literature survey, for aspects such as frequented locations and times of trips. We will use the following questions as a guide to evaluate our results. Does our analysis show a positive or negative correlation of high fares with certain locations and at specific times of day, and does this match the prior work done on rideshare data?

We will create hold outs of our data when running models to see initial sample performance. We will test to determine if we can use them to predict times or locations with high fares or tips. Additionally, we will test and see if the analysis of Chicago data can predict similar attributes in New York and vice versa.

Finally, we will try to match anecdotal experiences with our findings for both riders or passengers and drivers.

TOOLS

We plan to use Python as the main programming language for the project (via Jupyter Notebooks and VS Code). We will also utilize several resources and libraries provided by Python to implement our mining process:

- SciPy, Pandas, Matplotlib, NumPy, GeoPy, GeoPandas, Tableau, Dropbox, SQLite Database.

We plan to use the previously stated datasets in their CSV format, which have been downloaded already. Our code will be stored on Github. Given our dataset sizes, we're exploring options for storage which include Dropbox or Google Cloud. Additionally each member will retain their own personal copies.

MILESTONES

We are planning to have the following tasks completed by the marked dates:

March 27th – Preprocessing: have data cleaned, integrated and transformed by this point.

April 10th – Select and test models: have tools functional and working on small sets of our data. Refine any questions or other aspects of the project that need to be refined.

April 17th – Progress report: have completed progress report submitted to Moodle.

April 24th – Finalize models and evaluate: have all analysis on data completed at this point. Write an initial README file for github and begin a rough draft for the final report.

May 1st – Submit project: have the project presentation, final report, README and other files in GitHub finished and submitted.

REFERENCES

[1] Patricia S. Abril and Robert Plant, 2007. The patent holder's dilemma: Buy, sell, or troll? *Commun. ACM* 50, 1 (Jan, 2007), 36-44. DOI: https://doi.org/10.1145/1188913.1188915.

[1] Todd Schneider, 2015. Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance. toddwschneider.com

[2] 538, 2017. Uber pickups in New York City.. Kaggle.com/UberPickups

[3] Kaggle Competition, 2019. New York City Taxi Fare Prediction. Kaggle.com/TaxiPrediction

[4] Ziri Li and Yili Hong and Zhongju Zhang. 2018. An empirical analysis of on-demand ride-sharing and traffic congestion. Scholarspace.edu

[5] Stephen Zoepf, Stella Chen, Paa Adu, and Gonzalo Pozo, 2018. T*he Economics of Ride Hailing: Driver Revenue, Expenses and Taxes.* *Princeton.edu*

[6] David Schnurr. 2019. Uber Movement https://medium.com