Modeling Electric Scooter Demand in the Chicago Land Area

Charlie Marshall
Prof. Chen
Civ 495
12 June 2020

Introduction:

In the summer of 2018, the world was introduced to a new form of ride sharing in the form of electric scooters. These scooters were met with mixed reviews as they were seen as a convenient form of last mile transportation, but also as unsafe and unregulated. Electric scooters are convenient because they allow more flexibility for users than other shared forms of last mile transport such as Divvy Bikes which require users to dock bikes in pre-defined locations. In the summer of 2019, the city of Chicago set out to try to fix many of the safety and regulation problems by carefully designing a pilot program with e-scooter companies, allowing them to place their scooters in certain areas throughout the city from June 15th to October 15th. My goal in this analysis is to use demand data from within the pilot area to predict demand in other Chicago community areas during the same time period. Additionally, my goal is to continue to learn by implementing many of the machine learning and deep learning techniques that I have heard about in class throughout the quarter.

Data:

For this project, I collected data from four distinct data sources. First, the electric scooter trip data I used can be found as publicly available data on the Chicago Data Portal website. This dataset contains information on 711k trips that took place during the pilot period. For each trip, a distance, duration, start and end location were recorded. The start and end locations were given in the form of Chicago Community Areas, Census Tracts, and Longitude and Latitude. After exploring the data further, I decided to use Chicago Community Areas to indicate location because it had significantly more consistent reporting and there seemed to be more available socioeconomic data on the Community Areas than census tracts. The trip data needed to be cleaned so that all tips under 0.25 miles or less than one minute were dropped in order to get rid of cyclical tips and people just trying out the scooters.

Additionally, I used two different data sources for socioeconomic data for the Chicago Community Areas. In general, there are 77 Chicago Community Areas, with 16 of them contained within the pilot area. The Community Area data reported everything from transportation, age, race, and wealth factors within each of the areas. This information was used as factors to predict demand for electric scooters in each area.

Lastly, I used a Chicago weather dataset that was provided to me by Professor Chen. This dataset presented hourly weather data in the Chicago-land area on everything from temperature to wind gusts. This weather data was important because it allowed me to include factors which might explain why there might be more or less riders on one day than another. In general, different transportation modes are very sensitive to changes in weather, so weather data is important to include in time series data.

The final step was combining all these data sources into one dataset which would allow me to measure daily demand in every Chicago Community Area over the pilot period. This step was the hardest and most time consuming for me throughout this project because it required me to

basically make a full df from scratch which included the desired dates (June 15th – October 15th) for every single community area, over 9k datapoints. Additionally, it involved grouping both the trip and weather data by days, not hourly data.

Making sure that the data was in the right format so that it could all be analyzed correctly was a large part of the time spent on this project. In future work with this data, I believe it would be good to spend more time on feature engineering, identifying important features in the data and perhaps trying to reduce the dimensionality of the data.

Exploratory Data Analysis:

The EDA for this project was largely split into three different parts: Community Areas, Aggregate Trip-level, and Time Series. Looking at the Community Areas on a map, I was able to use Folium to visualize the socioeconomic differences between different Chicago Community Areas on a map. Overall, I found the visualization tools to be helpful in spearheading analysis. Looking at the ridership map, three areas on the east side of the pilot area, closest to the downtown area saw the highest overall ridership by a wide margin. From analyzing the demographic data, these three areas also had the youngest median age (early-mid 20s) and were near the top of the highest median income (around $70k). This means that there might be a connection between ridership, age, and income. It would make sense that younger people are more likely to use electric scooters because they have a reputation as not the safest form of micro-transit. Additionally, people with more income would have more disposable income and might be willing to spend it on a leisure scooter ride or try a new form of transit.

However, although these factors seem to show promise, the high ridership in these areas might be due to other factors. For instance, these three areas are close to the city and therefore probably see more tourist traffic than areas further away from the city. Tourists could take advantage of using e-scooters to get around the city with ease because they might not have a car. It would be worth it to further explore adding features about tourist traffic in each area when building the model in the future.

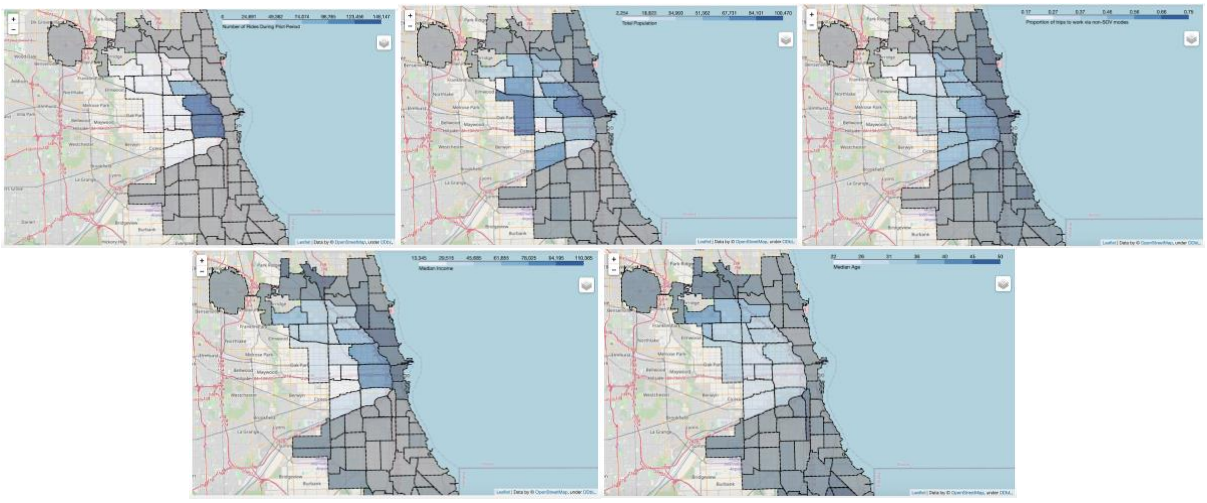

Figure 1: From left to right: daily ridership, total population, proportion of non-SOV trips, median income, and median age for each of the Community Areas within the Pilot Area

Next, I analyzed the aggregate daily trip data. It is important to notice the right skew of all three graphs. Most importantly, the extreme right skew of the daily demand histogram. It is evident from this histogram that in most community areas, the daily demand will be between 0 and 300. In very rare cases, the daily demand is over 500 people. However, although these datapoints are outliers, they are very important to be able to model because they represent a few community areas which consistently have high daily demand, not randomly high days in community areas which usually experience low demand.

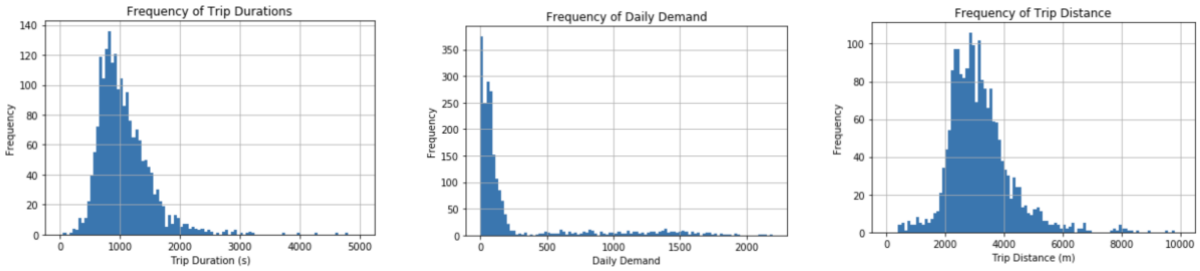

Figure 2: Daily trip durations, demand, and distances

Lastly, I was able to plot the daily demand for each Community Area in a time series. From Figure 3, it is evident that most communities experience daily demand between 0 and 300 daily trips. However, there were two communities which experienced drastically higher daily trip demand. When modeling, it will be difficult, but important to be able to correctly predict daily demand for each type of community.

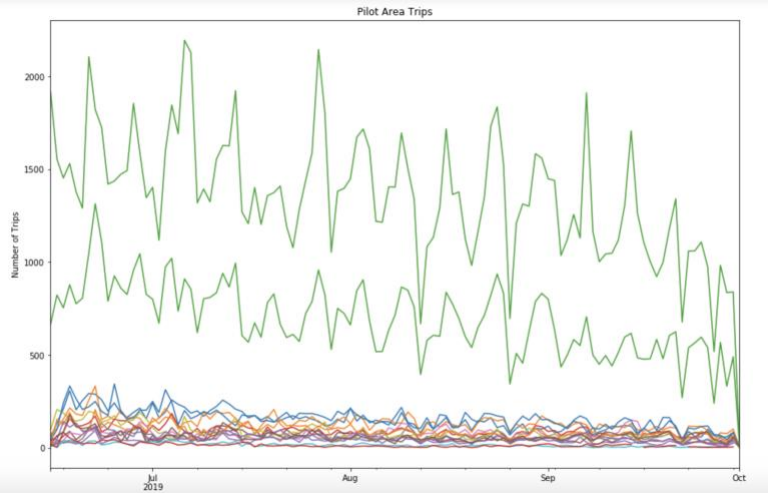

Figure 3: Time Series representation of daily demand of different community areas within the pilot area.

Modeling:

For this project, I tried to develop three different models to predict daily demand in all Chicago Community Areas: LSTM, XGBoost, and Random Forest. I spent a good amount of time trying to train the LSTM recurrent neural network model as this was my focus since our class on neural networks. However, I was very unsuccessful in tuning the model so that it would perform well on test data. In the end, I decided to not use LSTM and use XGBoost and Random Forest to predict daily demand. In general, both of these models were good at predicting on the training data. The XGBoost model had a training MSE of 0.08 and the RF model had a training MSE of 708. However, when used on the testing data, both models performed significantly worse, as seen in Figure 4.

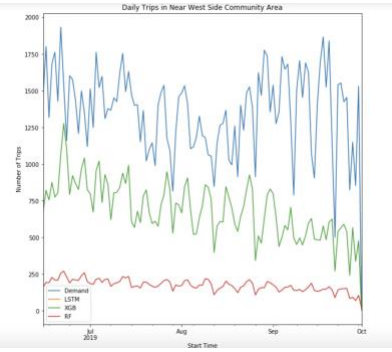

Figure 4: Model Predictions on Test Data vs Actual Daily Demand in Near West Side Community Area

In the case of the Near West Side neighborhood, the XGBoost performed well in modeling the fluctuations in the demand, but poorly predicted the magnitude of the demand. The RF model was even worse at predicting the magnitude. In general, I think this is because a vast majority of the neighborhoods experienced daily trip demands under 300 trips, so the model needed to be very resilient to recognize when demand would be much higher. In the future, I would incorporate more features which indicate past demand, or I will used different models such as an ARIMA model which takes into account past datapoints. Lastly, I would like to continue trying to develop the LSTM model in order to gain more meaningful experience in predicting with neural networks.

Conclusion:

In conclusions, I was able to identify the Lake View, Near West Side, Near North Side, and Lincoln Park community areas as potential hotspots for electric scooter use in the future. This information isn't necessarily surprising given the proximity of all these communities to the loop and the socioeconomic statistics of these neighborhoods. I think the most important takeaway from this is the ability predict how many electric scooters will need to be in certain locations daily and for scooter companies to then be able to place that amount of scooters in an area each day.

In general, I think this project was an important one for me. I was able to get out of my comfort zone and try my had at different modeling techniques which I had not used very much in the past. Took my first try at training and tuning a LSTM recurrent neural network, and while listening to other presentations I took note of other algorithms which might be useful for time

series data in the future including ARIMA. I think there is still a good amount of work that can be done on this project in applying different predictive algorithms and feature engineering so that the model might be better at predicting the magnitude of areas with higher levels of daily demand.