Shared Dockless Mobility: Predicting Usage

Gillian Foster, Saranya Nagarajan, Katherine Wroble, Malik Ouda, Mounika Tarigopula and Nadia Florez

Introduction

Advancements in social networking, location-based services, the Internet, and mobile technologies have contributed to a sharing economy. Shared mobility services offer transportation devices for short-term rental, which reduce transportation costs, lower carbon emissions and improve mobility for all. We use advanced analytics to predict the number of trips in downtown Austin, TX to better meet demand, limit the number of idle scooters, and maximize consumer satisfaction as well as revenues for the companies providing the service. We merge historical shared dockless mobility data with past weather data and create data preprocessing, visualization, and modeling pipelines to perform our analysis.

Data Preprocessing

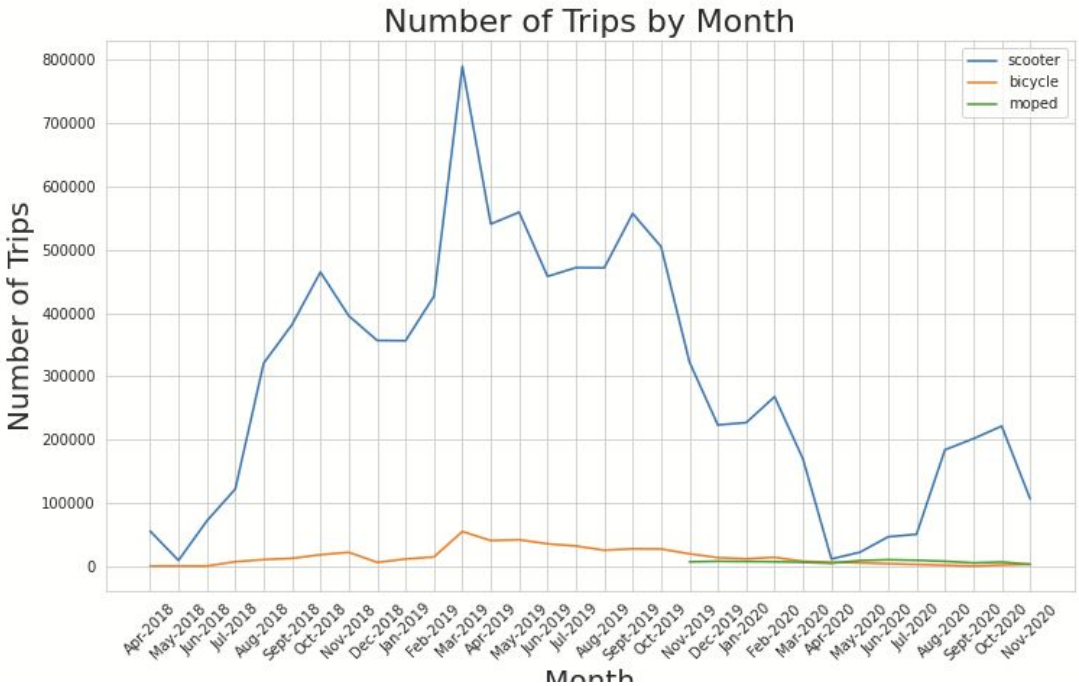
Our dockless shared mobility dataset ranges from April 2018 through November 2020. In this dataset, each row represents a “trip” using a bicycle, scooter or moped. We cleaned and condensed it into a dataset that contained aggregated data by day, and combined it with a daily weather dataset for the same time period.

We first checked for any null values and, given that this number was statistically insignificant compared to the largeness of the mobility dataset (<0.5%) we simply dropped any rows with null values. After dealing with null values, we converted everything into numerical data for ease of modeling by converting any time columns into UNIX timestamps. We also one-hot encoded our categorical data.

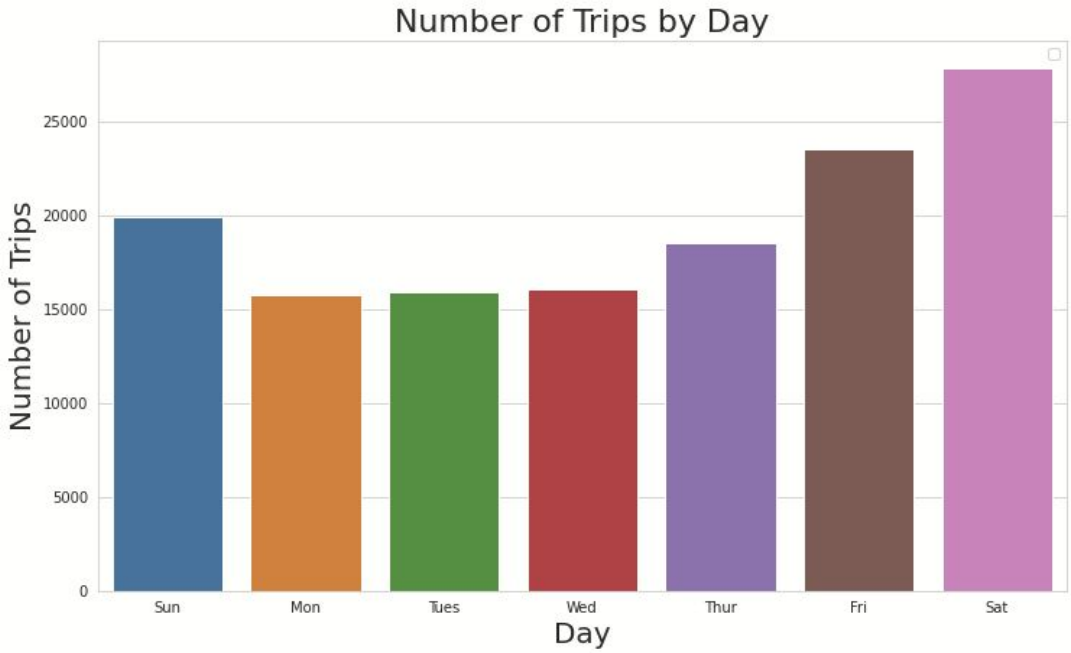
After cleaning the mobility dataset, we then aggregated the data by day, which condensed our ~10,000,000 row dataset down to only about 1,500 rows. This included data on the number of trips per day per vehicle type, the average trip duration and distance, and the date information. We combined this data with a weather dataset, which required no cleaning at all. The weather data included basic weather information for each day, and we kept only some of the relevant attributes like weather and humidity in our final modeling.

Visualization and Summary Statistics:

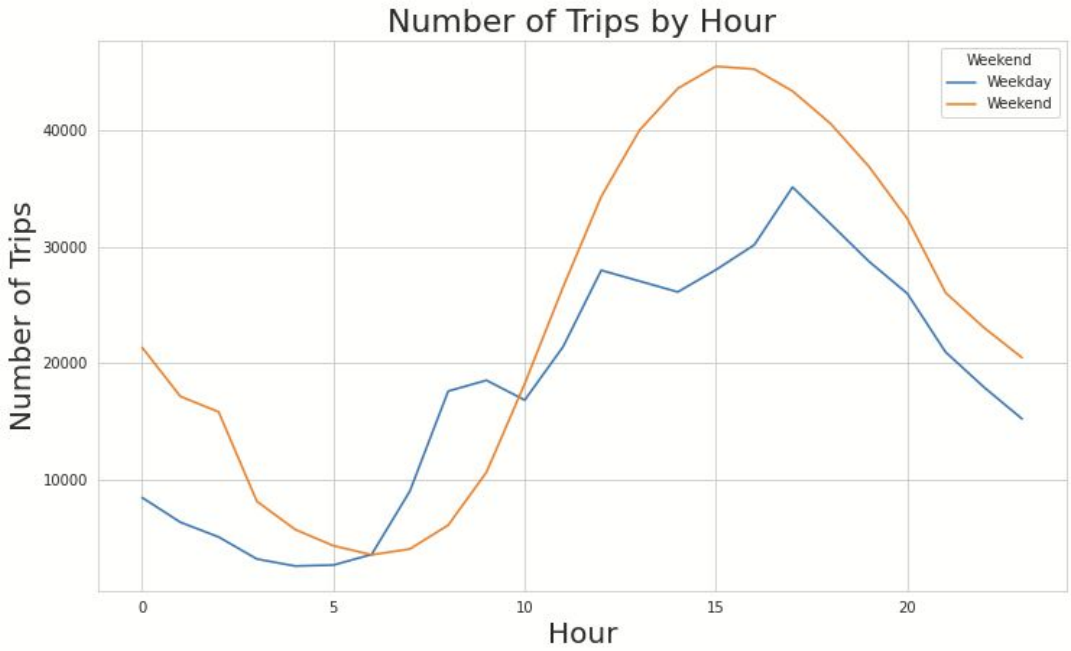
Mobility Data:



We have gathered the daily mobility (scooter, bicycle, moped) data for the time period April 2018 to November 2020. From the above graph we see that different types of mobility have different times at which they have been available to consumers i.e., consumers started using mopeds from late 2019 whereas scooter and bicycles have been around early to mid 2018. Among them, scooters are the most popular choice by consumers. We see a drastic change in usage of scooters by consumers from April 2018 to September 2018. There is a dip in usage of mobility from October to December in both 2018 and 2019 because of seasonality effects. Due to Covid-19, we are seeing sharp decrease in usage of mobility from march and April 2020



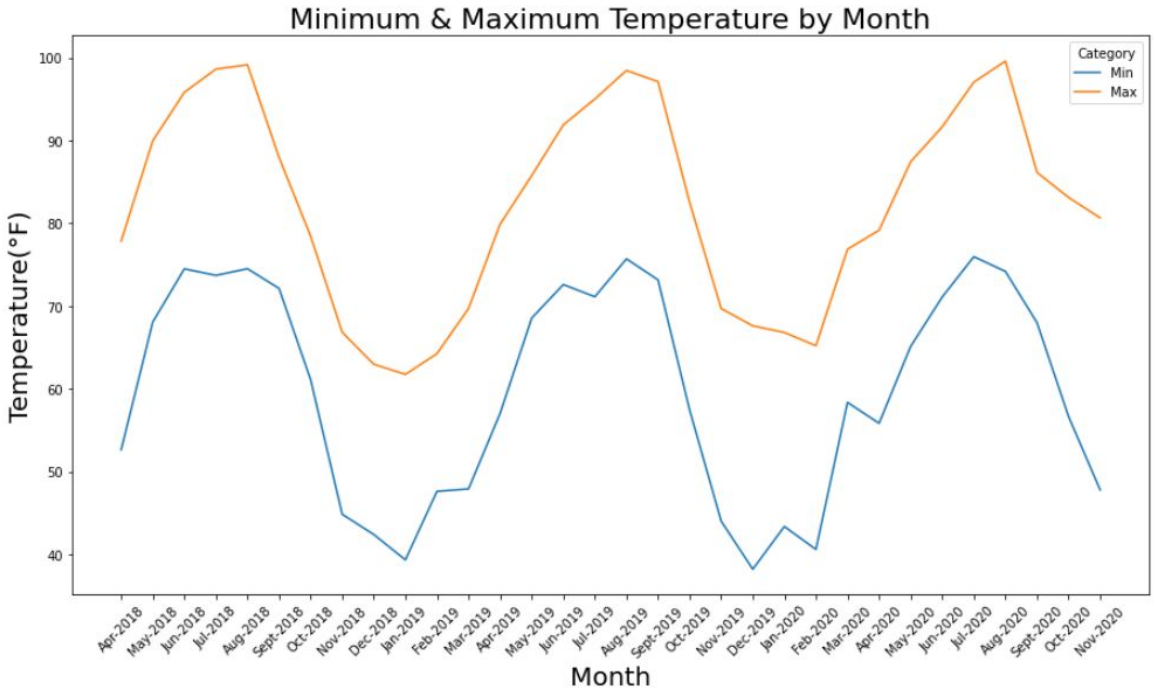
The plot above shows that shared mobility trips occur mostly on Friday, Saturday and Sunday. Particularly, Saturday shows the highest amount of trips.



From the plot above, we see that, on average, weekends have a greater number of trips than weekdays. Weekends see a steady rise in the number of trips throughout morning and early afternoon hours, peak around 3pm and decrease after.

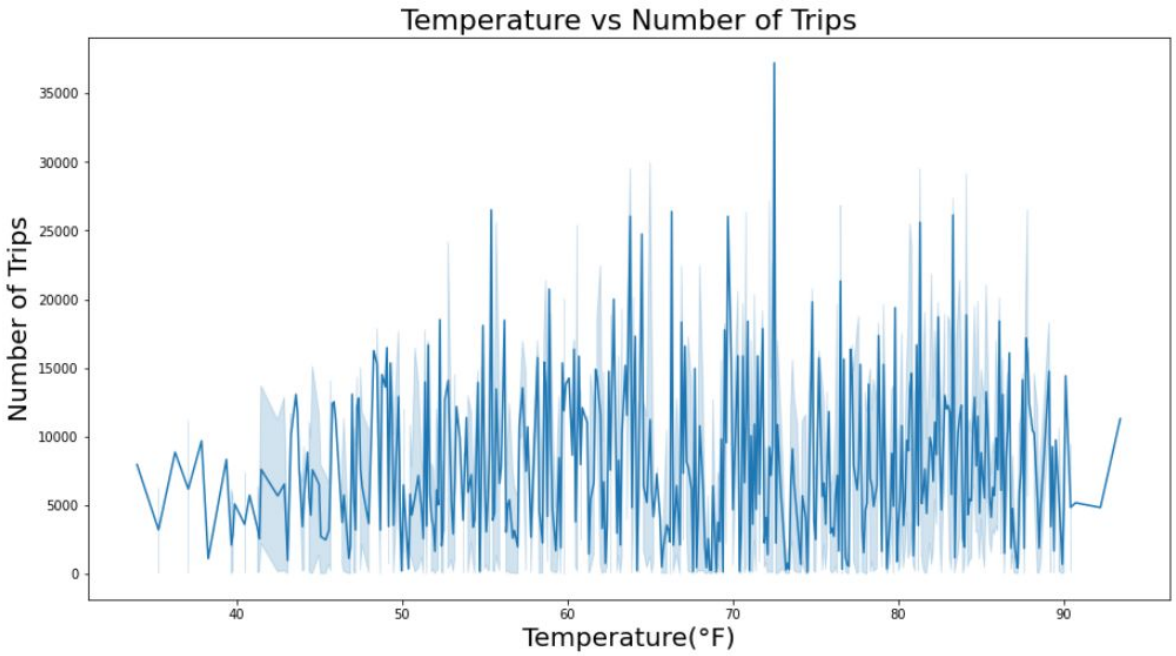
For weekdays, we see three peaks: one in the morning, one midday and one in the early evening. These peaks probably correspond to morning and afternoon rush hours, as well as lunch time.

Weather Data:



We gathered daily weather data from [weather underground] for the same period as the mobility data. The above graph shows the monthly maximum and minimum temperature. As we can see, the temperature is the highest at June - July and lowest at December - January.

Plotting daily number of trips against daily temperature, it can be seen that the number of trips are at a lower range when the temperature is less than 40F or greater than 90F. We observe no distinct correlation between temperature and number of trips for the normal temperature range.



Modeling

Our goal was to predict the number of trips taken per day. We used Mean Squared Error (MSE) as the main indicating metric of accuracy. Overall, we tried 5 different modeling techniques, landing on linear regression as the best predictor of the number of trips taken.

Having past success with boosting, we started our modeling process with CatBoost and XGBoost. To our surprise, we got a very high value for our MSE from each of our boosting models. Next, we tried an out of the box Random Forest Classifier. Again, we found that the model returned a high MSE. Wanting to try as many models as possible, we then used a Decision Tree Classifier. Once again, the model returned a very high MSE. Realizing that these complex classifiers were not working with our data, we found that there was an innately linear relationship between our feature space and the number of trips taken per day. Models like CatBoost and Random Forests were unable to translate this linear relationship into an effective model.

Recognizing the need for a linear expression of our variables, we then moved to linear regression. There were two categorical variables that needed to be expressed numerically - weekday and month. We initially tried fitting a regression without this information, but, as expected, trip counts varied in relation to certain months and weekdays. For weekdays, Fridays and Saturdays were associated with higher numbers of trips, while for months, February and March were the peaks. Even without these encoded variables, our linear regression model was already showing a considerable improvement of fit and predictions than any previously discussed model. Our final largest drop in error occured when we recognized the importance of time in our model. By shifting the trip count data back by one day - where we know the previous

day's number of trips, we significantly improved our performance. The final minimum MSEs for each model are listed in the table below for comparison.

Even implementing L1 regularization before and after, we saw an improvement in model performance when all weather data was excluded. Including weather data allowed the model to fit training data better, yet impaired its ability to generalize to test data. When weather data was excluded, we saw an improved balance between training and testing errors. We tested inclusion of individual weather features after realizing this, but none significantly improved the model at all. For example, the average daily temperature showed a fair correlation with the number of trips during data exploration, yet barely at all influenced the model's testing or training performance when included with other variables previously discussed - like weekdays, months, and the previous day's trip counts. We believed that the previous day's trip count may be already capturing the influence of weather trends already, yet fitting with the weather data and excluding the previous day's trip count showed a sharp decline in model accuracy.

Models	**Minimum MSE**
Decision Tree	101,681,458
Random Forest	67,575,173
XGBoost	42,182,216
Catboost	37,000,954
Linear Regression	24,237,005

Conclusion

With this project, we set out to bring together two datasets - the mobility dataset and daily weather data in Austin - in order to find out if we could build a predictive model for the number of trips in a day based on the day's weather data and daily mobility data. To do so, we built out 3 main pipelines: the data preprocessing, visualization, and modeling pipelines. In terms of modeling, we found simple linear regressions to have the best performance, likely due to some underlying linear relationship in the data. We were able to attain relatively accurate models with feature engineering, feature selection, and some hyperparameter tuning.

Based on our discoveries through our exploration and modeling of this data, we concluded that the number of trips taken per day is something that can be modeled with a fair accuracy. Having such a model would allow scooter companies to predict demand for their product and move scooters more procedurally to maximize profit and minimize the amount of idle scooters. In

addition to our basic mobility and weather dataset, we would be able to continue to supplement our data with holiday and event data in addition to weather data which would likely further increase the predictability and consequently the usefulness of our model.

Links to our work

Blog post:
https://nsaranya4.medium.com/shared-dockless-mobility-predicting-usage-c20d203417e5

Presentation: : https://youtu.be/7J0yjKLs65E