Forecasting Demand for Taxis in New York City

Toomas Roosma
Student ID: 1381691
Github repo with commit

August 24, 2022

1 Introduction

In New York City, the yellow taxicabs are widely recognised symbols of the city. With the arrival of ridesharing services such as Uber and Lyft in 2011 [1] and 2014 [2], the Taxi and Limousine Commission (TLC) has been forced to adapt to the new gig economy. With the introduction of e-hailing apps and green taxis, TLC has already made steps to adapt.

This report assumes the perspective of a taxi or rideshare company with the objective of forecasting demand for pickups in various locations around NYC.

Linear regression will be used to predict demand, based on data provided by the TLC from 2017 to 2019. This project aims to be also beneficial for taxi drivers, giving them insights, into when to work and when not, giving them better control over their work time. Yellow taxi and green taxi data have been used, and they will be collectively referred to as taxi services. Data from For Hire Vehicles (FHV) are also included in this report, forming Mobility Services with taxis. Weather data from the LaGuardia airport was included, as it influences heavily consumer behaviour.

2 Preprocessing, Analysis, and Geospatial Visualisation

2.1 Datasets

This report uses mainly the taxi trip data provided by the New York City Taxi and Limousine Commission (NYCTLC) [3]. Files detailing the outlines of different taxi zones were also provided by the NYCTLC and used for visualisation purposes.

Weather data from LaGuardia Airport recorded by the United States National Center for Environmental Information's (NCEI) Integrated Global Surface dataset [4] was combined with the previously mentioned dataset. This work makes the assumption, that the weather at the moment when the decision to hail a taxi is done, has significant importance, whether or not the consumer will hail a taxi or not. It is also assumed, that the weather at the time of dropoff does not matter because taxis are mainly used to circulate from door to door.

2.2 Preprocessing

2.2.1 Dataset range selection

We are working on the assumption that people are more prepared for bad weather during the winter months, so only data from the summer months (June, July and August) was used. Since 2017, the ride-sharing companies Uber and Lyft have been required to share and publish drop-off date/time and locations [5]. Both these companies have undeniably acquired significant importance in the ride service providing scene in NYC, so data from 2017 and onwards was used.

On the other end, data from 2020 and onwards was not used due to the lockdown and subsequent restrictions put in place by NYC authorities. This means, that the summer of 2019 was the last year included in this study. The world has since slowly gone back to normal, but to avoid model bias data from 2021 onwards was omited. After all the preprocessing, we were left with 167 935 947 unique taxi trip entries.

2.2.2 Feature selection

Although all the yearly datasets provided by the NYCTLC are said to follow one standard, they are not consistent with each other. There are also some outliers who do not follow the given structure.

The weather data recorded at the LaGuardia airport by the NCEI [6] contains many irrelevant fields for this analysis, like monthly averages and quality control codes. There are also data fields that are too local or not applicable for taxi services at the selected period, like wave height and snow accumulations. The type of daily present weather being reported was discarded due to conveying too little differential information about each hour. Cloud coverage was repurposed from an ordinal variable to discrete, as cloud coverage of missing hours was imputed from neighbouring hours. Other weather attributes like windspeed and temperature were imputed similarly.

2.2.3 Eliminating outliers

Numerous techniques were used to eliminate data points that were deemed outliers. Here are a few main criteria:

- **Trips with invalid pickup and drop-off location ID** were discarded from the dataset, as the shapefiles defined only location IDs in the range 1-263.
- **Trips with pickup date outside the defined range** were discarded. Drop-offs outside the selected range were left in because we are interested in predicting pickups.
- **Trips with negative duration** were detected and discarded.
- **Trips with distances over 6 hours** were discarded from the dataset. According to google maps, a trip between Tottenville and Wakefield (extremities of the area of study) would take only 2h during rush hours.
- **Trips with distances over 150 miles** were discarded from the dataset, as the average trip length was 3 miles with a standard deviation of 7 miles. The same trip mentioned previously is 46 miles one way.
- **Trips with no passenger** data or zero passengers were removed.

One of the goals of this report was to include data from FHVs, so only the following data fields were retained from the mobility services dataset:

- Month, day and hour
- Weekday
- Pickup location ID

And the following from the weather dataset:

- Wind speed
- Dew point
- Visibility
- Atmospheric pressure
- Air temperature
- Cloud coverage

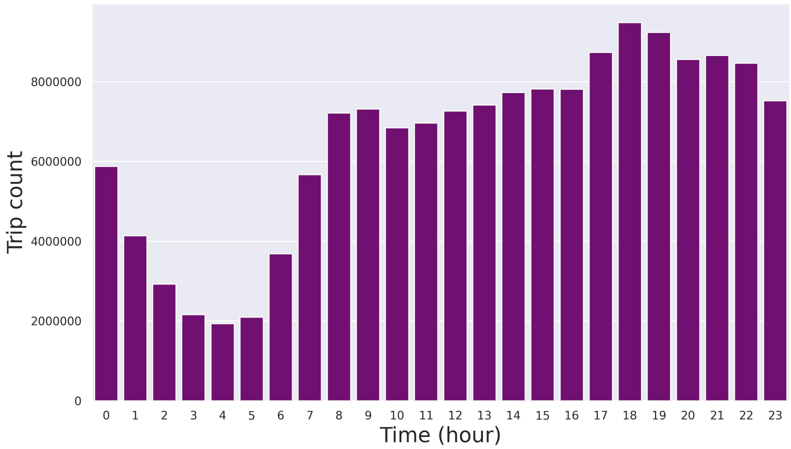

Figure 1: Total pickup count througout the day

The trip amount stayed consistent throughout the with a visible peak at the rush hour around 5-7 PM Figure 1.

3 Modelling

3.1 First iteration

A linear regression model was fitted with the mobility services data from the previously selected temporal range. Location IDs were one-hot-encoded due to them being categorical attributes. Other attributes were normalised. Pickup datetimes were split into components, pickup month, pickup day, pickup day of the week and hour. Minutes and seconds were ignored.

The model achieved a Root Mean Square Deviation of 227 and converged after 17 iterations. A better metric of the model accuracy would be the Mean Absolute Error (MAE), which was 152. This model's performance was below expectations, as the standard deviation of pickups per zone per hour was 189.

3.1.1 Prediction and Error Analysis

As seen in Figure 2, the model tends to over estimate the demand in Manhattan, central Brooklyn and at the John F. Kennedy International Airport. Aprat from the airport, this prediction looks a lot like the population density map [7] from 2010. The map showing the differences between actual demand and predicted demand is clipped at 300 for better viewing purposes. Over estimating the demand in the JFK airport area is expected, as the demand will highly depend on the number of inbound flights, and data, that the model did not have at its disposal.

As seen on Figure 3, the model accuracy is drastically lower during nighttime. For the next model iteration, only times between 5 AM and 10 PM will be kept, in order to help it.

4 Recommendations

Before the model can unlock its true potential, an expert with New York knowledge should take a look at its predictions and give opinions on them. As seen in Figure 4, the trip distance also varies a

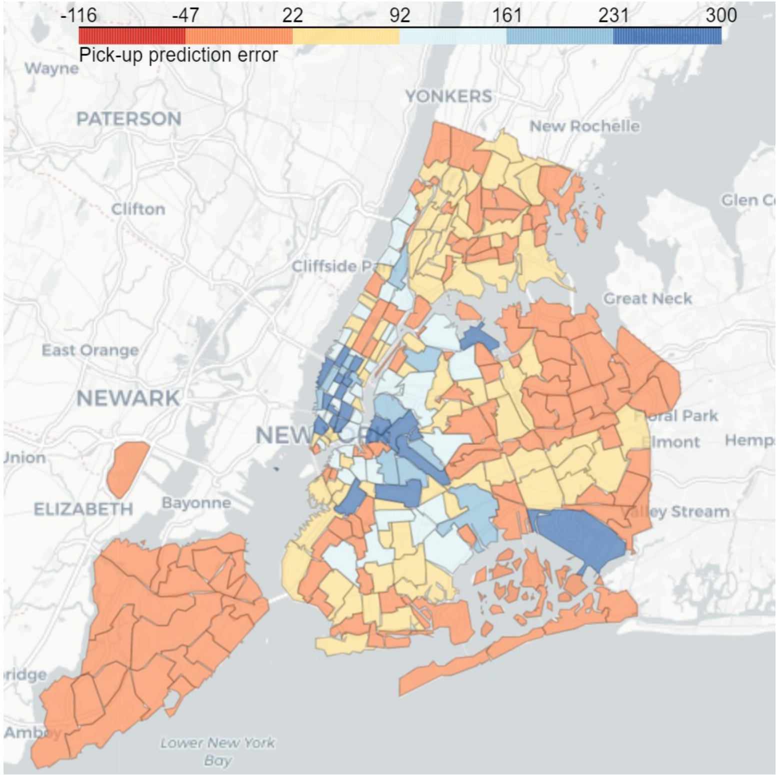

Figure 2: Pick-up prediction error on August 17th, at 17:00

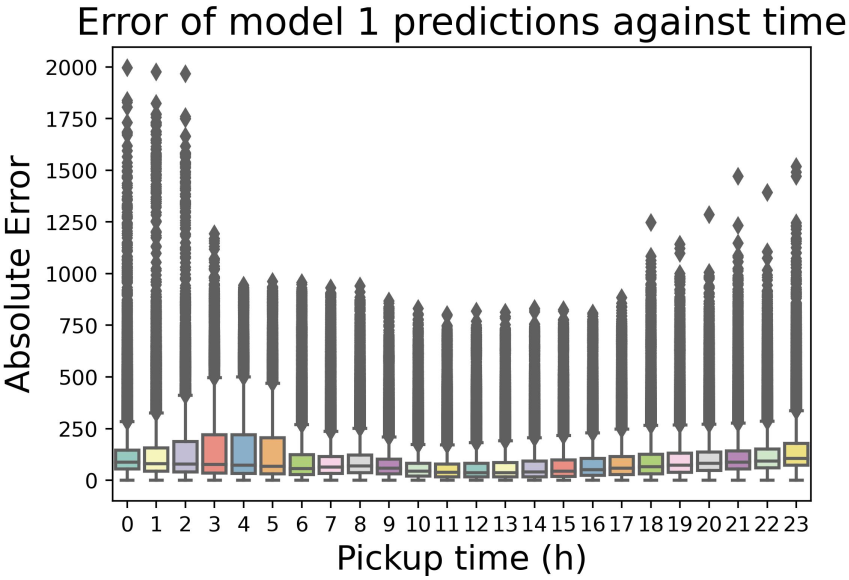

Figure 3: Absolute error of predictions with respect to time

lot depending on the time of the day.

New York City's legislators could also push further laws to force ride-sharing companies to share more of their data. It is clear, that if we could use the number of users concurrently opening the application and actively looking for a ride, instead of using a proxy of pickups, accuracy could be improved.

4.1 Discussion

The linear regression model does not perform exepcionally well, but accounting for the fact that the pickup number per hour per zone varies greatly (mean of 764 and standard deviation of 1024 [1]). The most appropriate metric to grade our model is Mean Absolute Error (MAE), because of how easy it is to interpret, especially for our target group. Our model achieved an MAE of 123, with an R^2 of 0.58.

The NYCTLC has not yet published 2022 summer data on their website, so there it is not possible to evaluate the model's performance with current data. However, with a more thorough cleaning and data preprocessing, greater accuracy can be achieved. Normalisation did not have a big impact on model performance.

Before the model fullfills its second goal of giving the taxi drivers decision power over their working times, a new metric should also be put into place, because demand is not the only metric for profitability. As seen on Figure 4, trip lengths vary greatly depending on the time of the day. This variance can not be explained by weather data only and needs an expert insight into NYC work culture.

Looking into the feasability of a neural network is also recommended. This project looked into it but abandoned the lead due to the high computational requirements of training such a model. It is highly probable that a (deep) neural network predicts the data better, as human reactions to weather can not be explained by mathematical formulas.

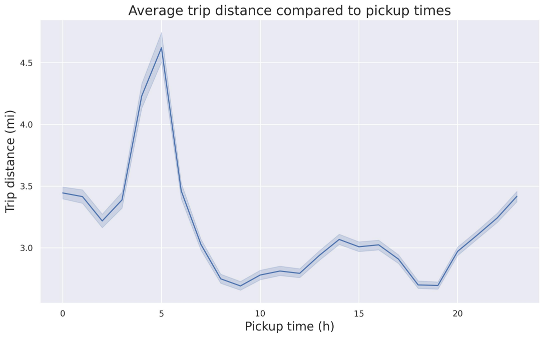

Figure 4: Trip distance over pickup time

5 Conclusion

This model proved itself interesting enough to pursue further research. However, due to the lack of sharing information from other ride-sharing companies like Uber, predicting the demand is limited to proxy data and weather data. With data from them, more precise amounts of passengers could be estimated, which can also be used as a proxy for demand.

[1] Excluding zones and hours with zero trips

References

[1] Lisa Eadicicco. *Uber just went public — here's what the app looked like when it first launched in New York City in 2011*. https://www.businessinsider.com/uber-old-app-new-york-city-ipo-2019-5. Accessed: 2022-08-21.

[2] Christine Lagorio-Chafkin. *Lyft in New York City: Let's Try This One More Time*. https://www.inc.com/christine-lagorio/lyft-another-nyc-launch-attempt.html. Accessed: 2022-08-21.

[3] New York City Taxi and LimousineCommission. *TLC Trip Record Data*. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: 2022-08-22.

[4] National Centers for Environmental Information. *Integrated Surface Dataset (Global)*. https://www.ncei.noaa.gov/access/search/data-search/global-hourly. Accessed: 2022-08-20.

[5] New York City Taxi and Limousine Commission. *TLC Trip Records User Guide*. https://www1.nyc.gov/assets/tlc/downloads/pdf/trip_record_user_guide.pdf. Accessed: 2022-08-20.

[6] National Centers for Environmental Information, US Air Force - 14th Weather Squadron. *FEDERAL CLIMATE COMPLEX DATA DOCUMENTATION FOR INTEGRATED SURFACE DATA (ISD)*. https://www.ncei.noaa.gov/data/global-hourly/doc/isd-format-document.pdf. Accessed: 2022-08-20.

[7] U.S. Census Bureau. *Population Density by Census Tract New York City*. https://www1.nyc.gov/assets/planning/download/pdf/data-maps/nyc-population/historical-population/pop_density_1950_2010.pdf. 2010.