Forecasting the Yellow taxis and Uber ride-share demands in the pandemic

Kasturi Deshpande
Student ID: 1127409
341ac9c1b95f67d6244bcaad4d28b75dbbd9e128

August 21, 2022

1 Introduction

Covid-19 has had an significant impact on the ride-share industry due to the perceived health risks underlying the nature of the virus and the frequent lockdowns. Uber bookings were down by 75% whereas the number of Yellow taxis in New York plummeted from 11,435 to 2,193 during the peak of the pandemic [1]. These numbers have since then shown a slow but positive upwards trend as the world is adjusting to a new normal[2].

The following report aims to predict the demand for pickups for the Yellow taxis and Uber ride-share services in different locations in the New York City (NYC) in any desired month or day given the Covid-19 case numbers and the average positivity rate. Statistical testing using ANOVA was implemented to understand the significance of various features in the prediction of the demand. A Random Forrest Regression model was then trained to predict the pickup demand in various locations for both the taxi types.

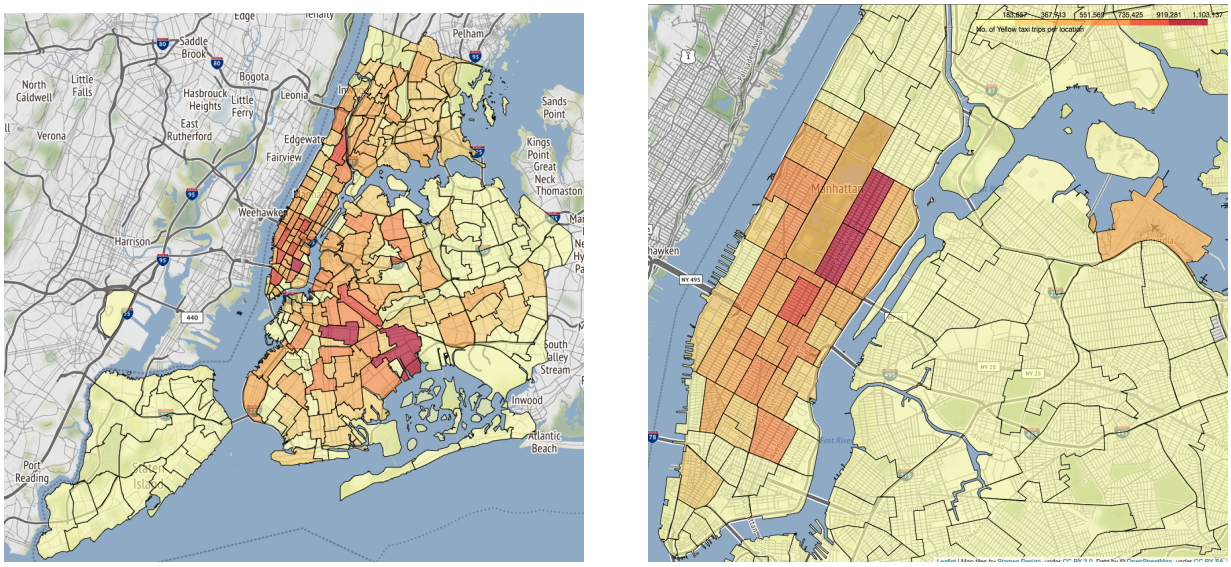
2 Dataset

2.1 Taxi Dataset

The report analyses the data retrieved from the Taxi and Limousine Commission's Official [3] for the Yellow taxis and the HVFHV records which includes the Uber ride-share trips data. Geospatial data was extracted from the same website containing the location data for the different locations in the New York City which is discussed in the further sections.

2.2 External Datasets

An external dataset containing the daily Covid-19 data for each New York Borough was extracted from New York's Health Data website[4]. From this dataset only a few features such as the daily cases numbers and the positivity rate were included in the model training. For the geospatial visualizations of the boroughs, the relevant data was extracted from the City of New York website containing the boundaries for the boroughs[5].



(a) Uber trips

(b) Yellow taxi trips

Figure 1: Observed Uber and Yellow taxi trips

3 Pre-processing

3.1 Data Selection

This report aims to predict the demand for pickups for mainly two type of taxis - Yellow taxis and Uber ride-share services considering the impact of Covid-19. For this, the training data was taken as all the trip records for both the taxi types from January 2021 to December 2021 and the testing was carried out on the records from January 2022 to April 2022. This range of data was selected as lockdowns in New York slowly started lifting up from late 2020. Keeping the concept of a "New normal" as the center i.e. returning to daily life in the midst of a pandemic, the above range suited the best to build a model which can be used in the presence of a health emergency such as the Covid-19 pandemic. Figure (1) shows the demand for both the taxi types in the presence of the Covid-19 cases (fig 2) in 2021.

3.2 Data Cleaning

The initial dataset included 30,904,308 Yellow taxi trips and 174,596,652 HVFHV trips. The data was further filtered and preprocessed based on the data dictionary available on the TLC's website. The HVFHV included the taxi records from four businesses - Juno, Uber, Via and Lyft. Out of this only Uber was selected belonging to the license plate - HV0003. This filtered out 48,467,588 unwanted rows from the HVFHV dataset. The external Covid-19 dataset did not require any data cleaning or imputation procedures.

- **Missing values** - 1,478,695 records from the Yellow taxi dataset were discarded due to empty RatecodeIDs.
- **Payment type** - Yellow taxi trips paid only by card were selected.
- **Ride type** - Only the standard rides from both the taxi types were selected (Uber had stopped its group ride services in 2021 due to the Covid-19 health risks. There were very few trips from and to the airports due to the international travel restrictions. Hence these were not included

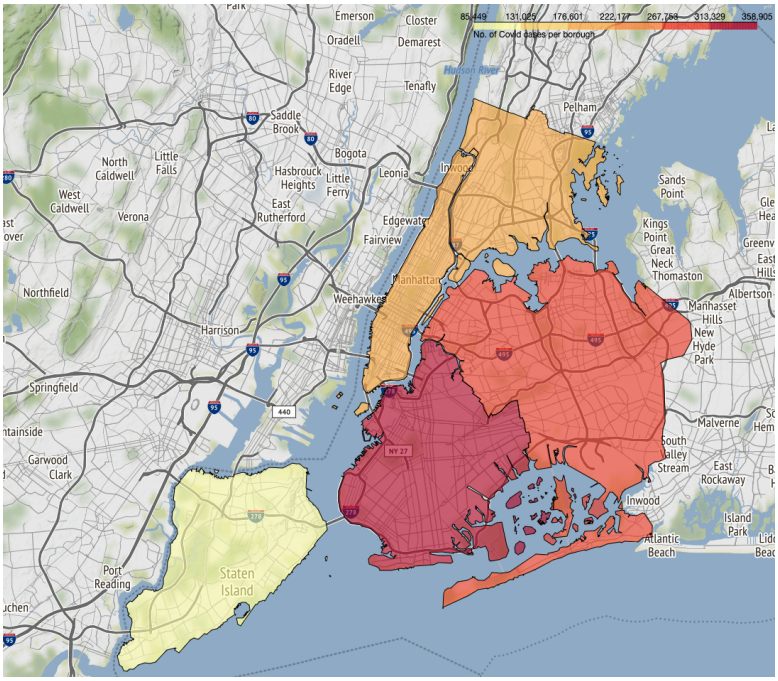


Figure 2: Covid-19 cases by NYC's boroughs

as well.)

3.3 Outlier Analysis

Outliers were removed from both the datasets based on the existing rules in the data dictionary using different techniques.

- **Location ID** - Pickup and drop-off location IDs outside the range of 1 - 263 were discarded.
- **Total amount** - Total fare amount was calculated based on the given base and other fare charges. Values less than or equal to 0 were discarded.
- **Passenger count** - Trips with a passenger count less than 1 or more than 4 were removed as they qualify for invalid trips.
- **Trip duration** - Trip records with a negative trip duration were removed from both the datasets.

3.4 Feature Engineering

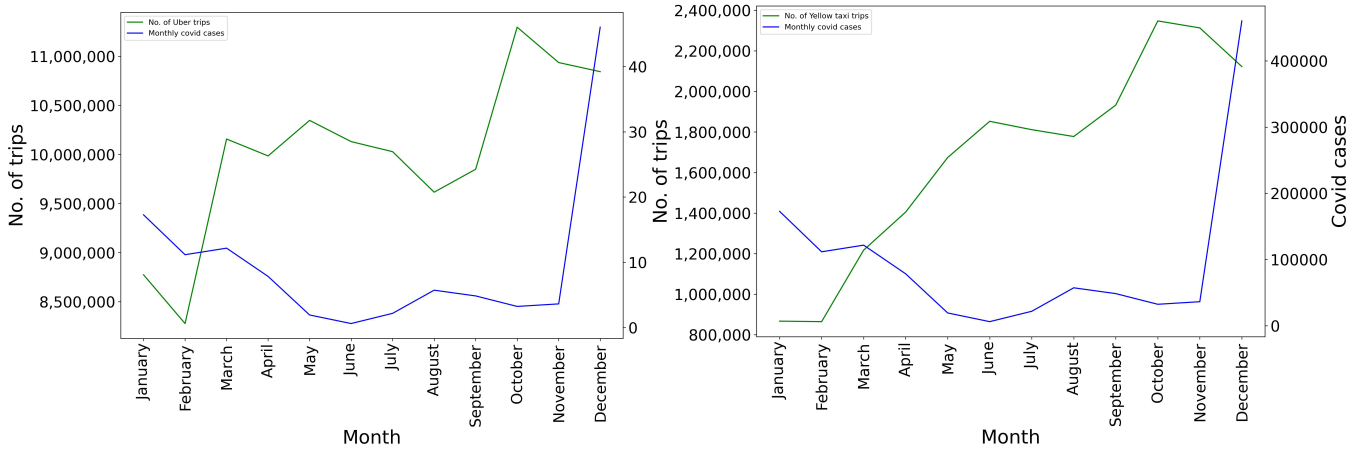
The Yellow taxi trip records and the Uber trip records were merged together after the above preprocessing steps to perform feature engineering. This merged dataset consisted of 146,922,395 records in total. The final training dataset was constructed by merging the taxi datasets (Yellow and Uber together) and the Covid-19 dataset.

From the final taxi dataset the following features were selected -

• Pickup Location ID • Drop-off Location ID • Month • Day • No. of trips • Taxi type (Yellow - Standard or Uber - Standard)

From the Covid-19 dataset the following features were selected -

• Covid cases • Average positivity rate



(a) Uber trips

(b) Yellow taxi trips

Figure 3: Impact of Covid-19 on Uber and Yellow taxi trips

3.5 Aggregations

All the further analysis and prediction was carried out on aggregated dataset. From the final training dataset, data was aggregated by the trip locations, month and then by day. The observed aggregated trips of both taxi types and the Covid-19 cases number in each borough can be seen in figures (1) and (2).

4 Preliminary Analysis

As we know, Covid-19 has affected every industry and caused disruptions in our daily life. It's impact on the transport industry has been significant. Hence the Covid-19 dataset was joined to the existing taxi dataset and this section attempts to investigate its significance in the prediction of the taxi trips using ANOVA.

4.1 Covid cases and average positivity rate

A linear model with interaction was used to test the significance of the external dataset predictors i.e. number of Covid-19 cases and the average positivity rate with the total number of trips per location as the response. As expected, both the factors were significant in predicting the demand for taxis in each location.

	Df	Sum Sq	Mean Sq	F value	Pr(F)	
covid_cases	1	1.5748e+05	157475	18.011	2.197e-05	***
avg_positivity_rate	1	4.0815e+07	40815487	4668.174	2.2e-16	***
covid_cases:avg_positivity_rate	1	7.6740e+06	7673974	877.692	2.2e-16	***
residuals	4366786	3.8180e+10	8743			

4.2 Month and Day

Both month and day of the week were tested by fitting an interaction model to predict the demand of taxis. Similar to the Covid-19 features, these were found out to be significant as expected.

	Df	Sum Sq	Mean Sq	F value	Pr(F)	
month	11	4.0630e+07	3693626	423.054	2.2e-16	***
day	6	2.9095e+07	4849190	555.407	2.2e-16	***
month:day	66	3.4127e+07	517079	59.224	2.2e-16	***
residuals	366706	3.8125e+10	8731			

5 Regression using Random Forrest

The final training dataset consisted of numerical as well as categorical features. Taking this into consideration, a Random Forrest regression model was trained on an aggregated model consisting of 3,057,713 instances and evaluated on 1,498,093 instances. Random Forrest models are robust to missing values, can handle categorical as well as numerical data and produce highly accurate results as they are built as an ensemble of multiple decision trees.

For the demand prediction, 5 categorical features and 2 numerical features were used to train the model to get a continuous output for the taxi demand. For the model evaluation and Mean Absolute Error (MAE) was used as the evaluation metrics. The trained model performed well giving the results -

- Mean Absolute Error (MAE) - 43.96

This indicates that the model prediction of the demand of taxis at a particular location was accurate to within 44 trips.

5.1 Prediction and Error Analysis

The training dataset is unbalanced due to the volume of Uber trips that are made per day which is comparatively quite higher than the Yellow taxi trips made per day. For this, the Random Forrest Regression model was observed to be best as it works well for unbalanced datasets as well.

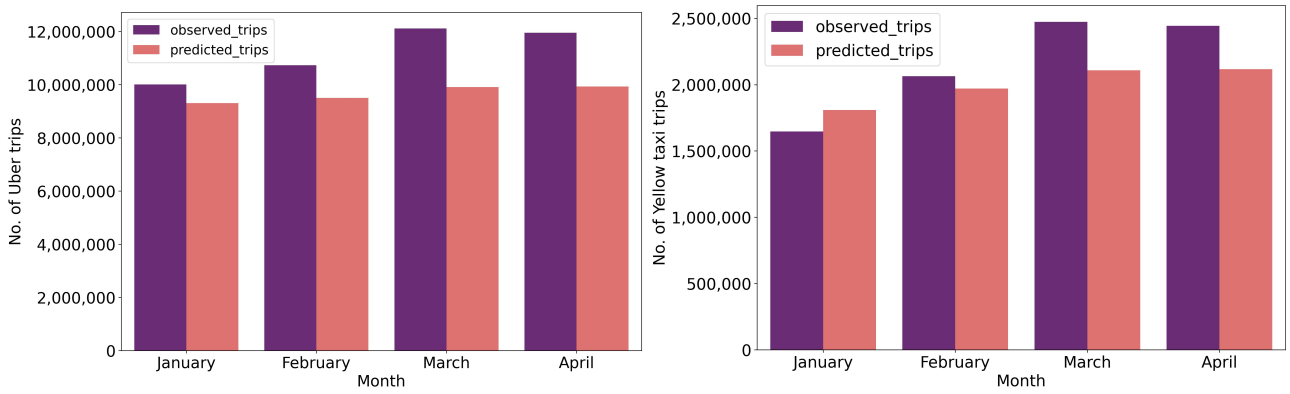
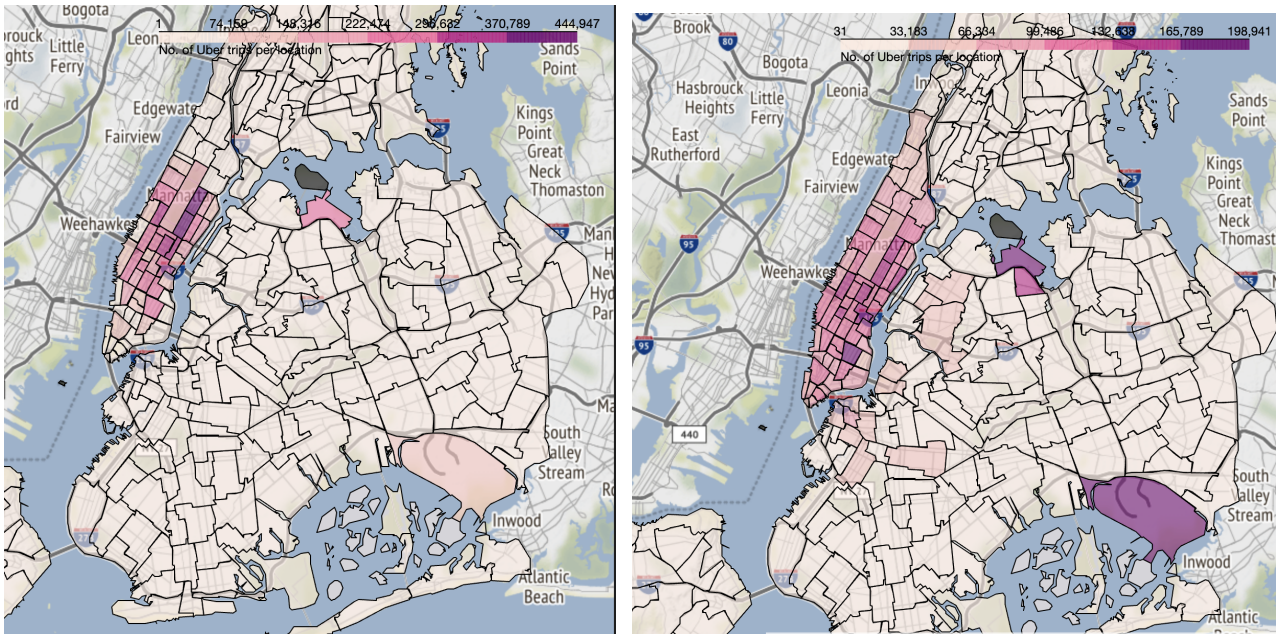


Figure 4: Observed and Predicted taxi trips for Uber (left) and Yellow taxi (right)

Figure(4) shows that the model underestimates the demand for Uber trips whereas it overestimates the demand slightly for the Yellow taxi trips in the month of January. But overall the model results in a good prediction of the demand. To investigate this further, the observed and predicted demand in Manhattan was compared to check the model performance.

From fig (5) it can be observed that model the overestimates the demand for Yellow taxis in a few locations. This can be due to the imbalanced training dataset. As the training dataset contains high volume of Uber trip data (due to the high volume of daily record of Uber trips), the model overfits the



(a) Observed trips (b) Predicted trips

Figure 5: Observed and predicted Yellow taxi trips in Manhattan

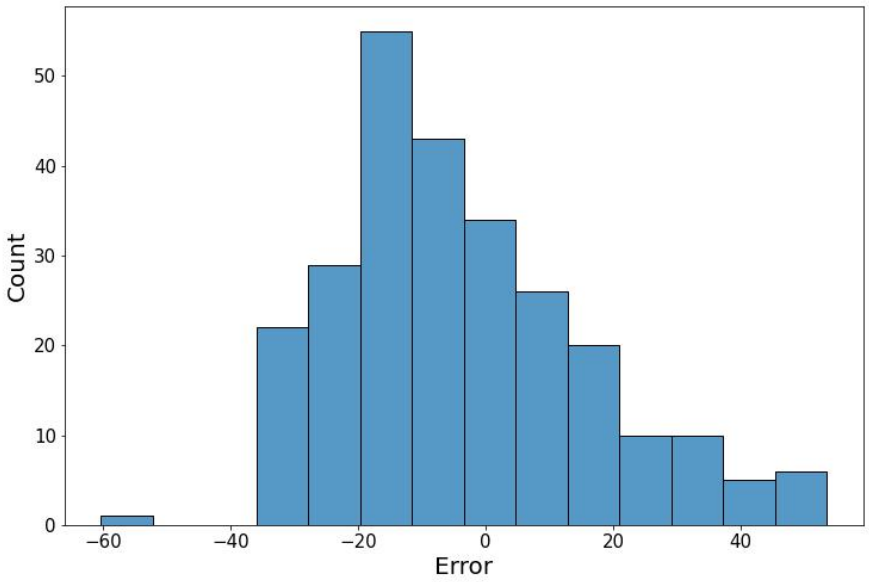


Figure 6: Prediction error for the Pickup locations

trip instances for the Yellow taxis. This suggests that the model is not sufficiently generalizable for predicting the demand for Yellow taxis. This can be improved by including more training instances of the Yellow taxi trips. Figure (6) shows the distribution of the prediction error which indicates some level of over-fitting although a lot of predicted values are centered around the true value.

6 Conclusion and Recommendations

This report aimed to predict the demand of Yellow taxi trips and Uber trips in the locations around the NYC given the impact of external factors such as Covid-19. The above constructed model was fairly simple and yet produced considerably good results given that the mean absolute error was observed to be 43.97. The model accuracy can be increased by adding more yellow taxi training instances to make the dataset more balanced. Techniques such as hyper-parameter tuning which was not implemented for the above trained model due to insufficient computational power can be used to fine tune the current model. Along with this, other machine learning models such as a neural network can be implemented by taking more features such as hospitalization rate, vaccine related data, etc. to understand the relationships between the different features to produce the best results.

Further investigation is required to understand and improve the over-estimation and under-estimation in certain locations as can be seen from figure (5). More external factors apart from the impact of Covid-19 factors should be taken into consideration such as weather, major sporting events, concerts, public transport accessibility, etc. to further build and improve the current model.

This model can be used by the taxi service providers to predict the demand for their services in the desired location for any given day or month. Businesses can use such a model to modulate the base prices of the trips depending on the forecasted demand. Drivers can also use this model to decide their working locations to get more number of trips depending on the high demand locations.

7 References

References

[1] *How the Pandemic Hammered a Yellow Cab Industry Already in Crisis*. URL: https://www.ny1.com/nyc/all-boroughs/news/2020/08/08/nyc-yellow-cabs-taxis-coronavirus-impact-on-drivers3. (accessed: 20.08.2022).

[2] *Alliance Of Uber And New York City Taxis Joins Two Pandemic-Hit Services*. URL: https://www.forbes.com/sites/katharinabuchholz/2022/03/25/alliance-of-uber-and-new-york-city-taxis-joins-two-pandemic-hit-services-infographic/?sh=68f879cb4563. (accessed: 20.08.2022).

[3] *Yellow taxi and Uber trips dataset*. URL: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. (accessed: 15.08.2022).

[4] *Covid -19 dataset*. URL: https://health.data.ny.gov/. (accessed: 16.08.2022).

[5] *Borough boundaries geospatial dataset*. URL: https://opendata.cityofnewyork.us/. (accessed: 20.08.2022).