# Predict the demand for Yellow Cab in New York City

Duc Hoan Nguyen
Student ID:
Github repo with commit

December 19, 2023

## 1 Introduction

Nowadays, the efficient functioning of transportation systems play a crucial role in the development of a fast-paced urban city. Expertly planned and executed transportation systems have a remarkable capacity to enhance productivity while concurrently alleviating issues of congestion and pollution[1]. Meeting the demand for ride-hailing service has always been a controversial issues for New York City. At the same time, the growth of ride-hailing services has had and may have negative impacts on city transportation and the environment.[2].

In this paper, we will show two different Machine Learning model estimating the hourly demand for yellow cab taxi rides in each neighbourhood around New York City. In addition to helping traditional taxis companies and drives to increase profits from learning more about demands, it can help ride-hailing services providers, for example Uber and Lyft, to better their decision making when it comes to matching the drivers with customers.

## 2 Dataset

The dataset we will use for this paper will be the trips records of Yellow Cab Taxis Trips from August 2022 to October 2022 and the weather data in the same time frame. We use this time frame because after inspecting the history weather data, we can see that this time frame covers a wide range of different weather.

### 2.1 NYC Taxi and Limousine Commission (TLC) Trip Records

The trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The records were collected and provided to the **NYC Taxi and Limousine Commission (TLC)** by technology service providers[3].

The yellow taxi cab (medallion taxis) has always been an iconic image synonymous with the hustle and bustle of New York City[4, 5]. We only choose medallion taxis trip records instead of all the traditional provided taxis (including green taxis (boro taxis)) since the green ones are only allowed to pick up passengers in Upper Manhattan, the Bronx, Brooklyn, Queens (excluding LaGuardia Airport and John F. Kennedy International Airport), and Staten Island, which is relatively small compare to the area that its yellow companions can provide. Medallion taxis are able to pick up passengers anywhere in the five boroughs, hence, its records is suitable when we want to inspect the hourly demands of every region in NYC[5].

We also do not use the data from high-volume for-hire vehicle bases (like Uber or Lyft). The reasons is that yellow cab taxis can presents the unplanned demands (demand for a ride when the passenger is in a urgent or cannot use other means to find for a ride) for ride-hailing services, which can only be shown by the traditional taxis. Moreover, with the partnership between Uber and the taxi industry from the middle of 2022, the data can also represents a part of demands from the mobile apps[6].

### 2.2 NYC Weather Dataset

The external weather dataset is obtained from an open-source weather API and offers free access for non-commercial use: **Open-meteo**. This data source can provide us with history of many different aspects of weather all over the world. However, we only use temperature, apparent temperature, the temperature equivalent perceived by humans, caused by the combined effects of air temperature, relative humidity and wind speed[7], (both in Celcius degree), rain (mm), and wind speed (km/h). These weather elements can represents the weather and how humans feel about the weather at that time.

A basic summary of data.

| *Datasets* | *No. of Features* | *No. of Instances* |
|---|---|---|
| **TLC Yellow Taxi Trip Record Data August 2022** | 19 | 3152677 |
| **TLC Yellow Taxi Trip Record Data September 2022** | 19 | 3183767 |
| **TLC Yellow Taxi Trip Record Data October 2022** | 19 | 3675411 |
| **TLC Yellow Taxi Trip Record Data Total** | 19 | 10011855 |
| **Open-meteo NYC Weather** | 6 | 2208 |

Table 1: Dataset Shape

## 3 Preprocessing

Despite only using a small number of features from both of the datasets, we will still eliminate the instances with inconsistencies and unreasonable values.

### 3.1 Data Wrangling

#### 3.1.1 Weather Dataset

Since this dataset is already robust, there is not much aspects that we have to take care about. We only have to split the time feature in this dataset and convert into two separate features: *time* and *date*.

#### 3.1.2 NYC Taxi and Limousine Commission (TLC) Trip Records

There are several steps that we need to do with this dataset.

- **Null value in the dataset**
  Since the NaN values are not in the columns that we are considering for our model and the missing percentage is less than 5%, we do not remove or fill these cells.
- **Unreasonable numerical values**
  We remove instances with these features having value less than or equal zero: *"passenger_count", "trip_distance", "fare_amount", "tip_amount", "tolls_amount", "total_amount"*.

- **Remove Location ID with value outside of our given region**
  Since there are only 263 given regions in the dataset, we will remove instances with either *"PULocationID"* or *"DOLocationID"* not in the range of 1 to 263.

### 3.2 Feature Selection and Data Aggregation

Because we are predicting demands based on region and time, we only care about the pickup location and time (not the date). However, since we need to join the trip records with the weather data, we still keep the date in our dataset.

After that we join two dataset by the same hour and date. The final dataset features are:

- *PULocationID*
- *time*
- *date*
- *temperature_2m (°C)*
- *apparent_temperature (°C)*
- *rain (mm)*
- *windspeed_10m (km/h)*

## 4 Analysis and Geospatial Visualisation

We will use the number of trips as a indicator for the demands for taxis in NYC in each region.

### 4.1 Demands for Taxis in Regions

| | **Zone** | **Borough** | **No. Trips** |
|---|---|---|---|
| 1 | JFK Airport | *Queens* | 493822 |
| 2 | Upper East Side South | *Manhattan* | 430422 |
| 3 | Midtown Center | *Manhattan* | 379641 |
| 4 | Upper East Side North | *Manhattan* | 372216 |
| 5 | Penn Station/Madison Sq West | *Manhattan* | 318075 |
| 6 | Midtown East | *Manhattan* | 309235 |
| 7 | Lincoln Square East | *Manhattan* | 294593 |
| 8 | Times Sq/Theatre District | *Manhattan* | 290632 |
| 9 | Murray Hill | *Manhattan* | 286570 |
| 10 | Clinton East | *Manhattan* | 277065 |
| 11 | LaGuardia Airport | *Queens* | 268244 |
| 12 | Midtown North | *Manhattan* | 262642 |
| 13 | Upper West Side South | *Manhattan* | 253440 |
| 14 | Union Sq | *Manhattan* | 250918 |
| 15 | East Village | *Manhattan* | 238198 |

Table 2: Regions with the highest Number of Trips

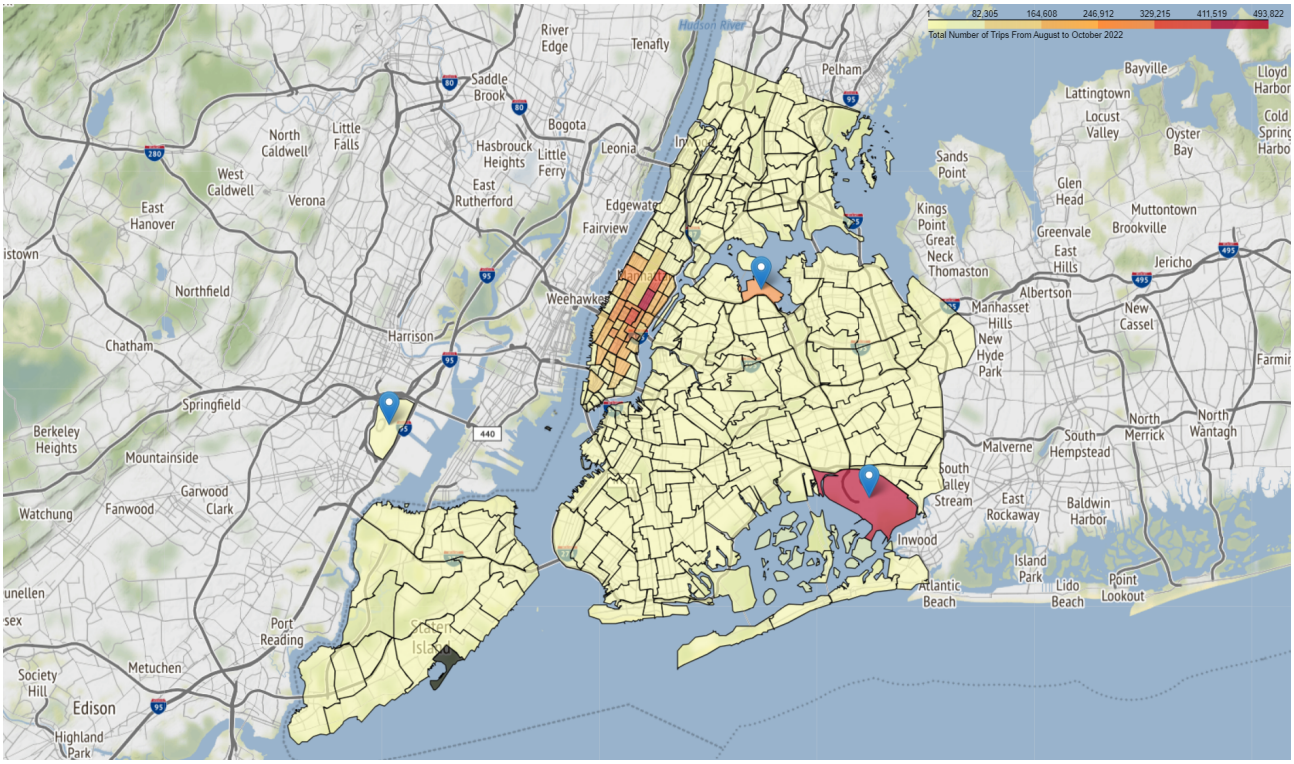


Figure 1: Total number of Trips from August to October 2022

Figure 1 describes the distributions of the total number of trips in regions from August to October 2022. Table 2 gives us 15 regions with their borough having the highest number of yellow cabs trips over 263 regions in New York City.

From Figure 1, we can see that there is a strong correlation between number of trips and the Pick-up Location ID. Based on Table 2, we can see that the *Manhattan* borough has the highest demands for medallion taxis with 13 of its zones has the highest number of trips. The other 2 zones in the list is in the *Queens* borough. However, these two zones appearance in the top 15 lists may be due to the fact that they both have an airports, the LaGuardian and the JFK Airport. The JFK Airport zone has the highest number of yellow cab trips, with a total value of nearly 500,000 trips over the three months period. The total number trips from the zones in the Table 2 is equal to more than 50% of the total of 9,115,448 trips from all of the New York City.

### 4.2 Hourly Demands for Taxis

Figure 2 describes the number of hourly trips from August to October 2022. We can see that there is a clear relationship between the time of the day and the demands for yellow cab taxis in NYC. According to the figure, NYC has the highest demands for medallion taxis at 6 p.m. with a total of 639557 trips, while it has the lowest demands at 4 in the morning with only 48177 trips. Despite the large difference between the peak and low point in demands, the number of trips in more than half of the day (from 10 a.m. to 10 p.m.) is still high with value larger than 40,000 trips per hour.

### 4.3 Trend in the Demand for Taxis

Figure 3 indicates the number of trips daily from August to October 2022. There is no clear trend in the number of trips over the time period. There is a clear decrease in the number of demand around

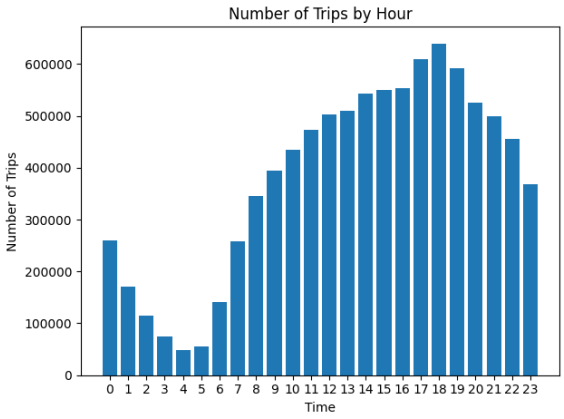


Figure 2: Total number of Trips from August to October 2022 based on time of the day

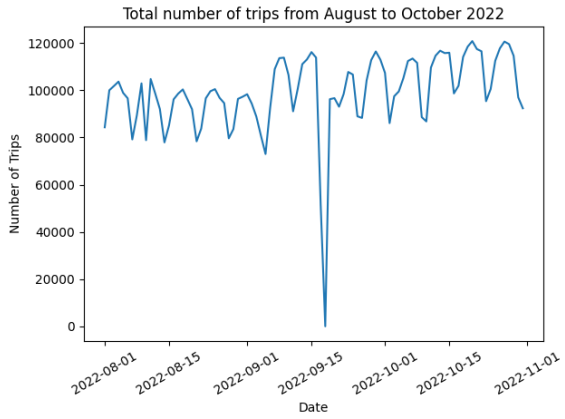


Figure 3: Trend in the number of trips in the time period

the middle of September 2022. However, this can be caused by our preprocessing steps may have removed all the instances on a single day, leading to a zero value in demand on that day. Overall, there is an increasing trend in the demands for yellow cab taxis, starting from a peak in August of just over 100,000 trips per day to a high point of more than 120,000 trips per day in late October.

Overall, we can still see that there has been an increasing demand for yellow cab rides in New York City over the three-month period.

# 5 Modelling

We use a Linear Regression Model and an ensemble model based on Decision Tree, Gradient Boosting.

## 5.1 Linear Regression

Since our response variable are of numerical type and our features are both numerical and categorical, Linear Regression model is a suitable model for this task. Our assumptions in this case is that our features are independent, meaning that there is no relationship between the time, locationID, and the weather data.

## 5.2 Gradient Boosting

Gradient boosting refers to a class of ensemble machine learning algorithms that can be used for classification or regression predictive modeling problems. Gradient Boosting is one of the best-performed boosting algorithms on a wide range of tabular datasets[8]. Hence, this model is suitable for our dataset, which consists of both numerical and categorical features.

# 6 Results

We use the dataset obtained similarly but in a different time frame, from March to May 2023, as the test dataset. We also do all the similar preprocessing steps on this dataset as the train dataset.

Since we are dealing with numerical variables, we use Root Mean Squared Error as our evaluation metrics. Also, we calculate the error based on regions.

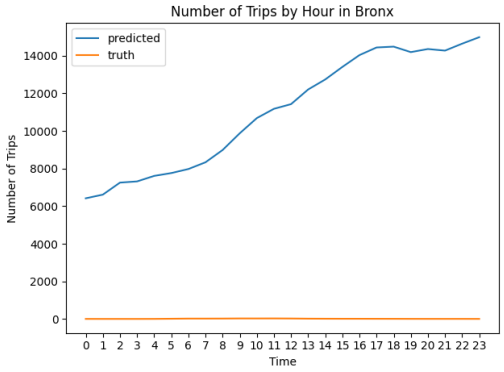


Figure 4: Linear Regression

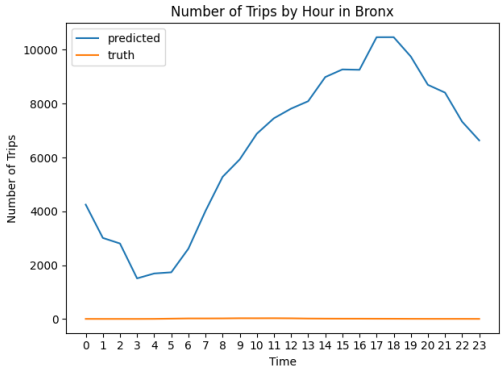


Figure 5: Gradient Boosting

Figure 6: Predicted vs True Number of hourly trips in Bronx

We also plot the Predicted against True value for 2 boroughs *Bronx* and *Manhattan*.

| **Borough** | **Linear Regression** | **Gradient Boosting** |
|---|---|---|
| *Bronx* | 11435.74 | 6968.67 |
| *Brooklyn* | 10595.23 | 5689.77 |
| *EWR* | 8590.05 | 2150.29 |
| *Manhattan* | 5967.08 | 2243.85 |
| *Queens* | 10521.24 | 6269.12 |
| *Staten Island* | 9935.67 | 4857.76 |

Table 3: Root Mean Squared Error

## 7 Discussion

From the Table 3, we can see that our error is quite large. In both case, the borough *Manhattan* has the lowest error rate in the linear regression model, and *EWR* has the lowest error rate in the other model. On the other hand, the borough having the highest error rate in both models is *Bronx*. We can see that *Manhattan* has a significant lower value of Root Mean Squared Error compare to other region. This can be caused by the difference between the number of instances between *Manhattan* and others boroughs. The more instances that we have in each borough, the more information that the model can learn about that region.

Between two models, we can see that Gradient Boosting model has a better performance compare to the Linear Regression model. The difference in performance may be caused by the non-linearity relationship between features and features or features and response variables. This aspect may caused the Linear Regression model to under-performed and cannot predict well. In contrast, Gradient Boosting can even model non-linear relationship between features and response variables, hence, it can perform better than the Linear Regression in this case.

## 8 Recommendations

Having a look at Figure 6 and Figure 9, we can see that because of the small number of instances from borough *Bronx*, we can see that both model cannot perform well on this borough. However, for the

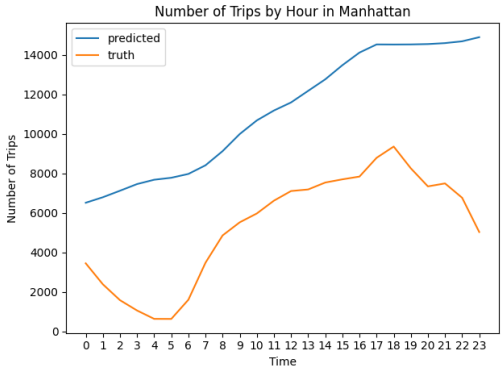


Figure 7: Linear Regression

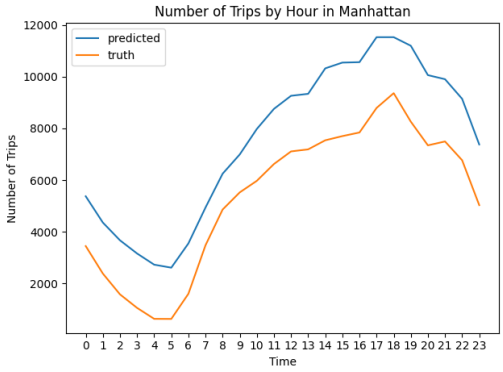


Figure 8: Gradient Boosting

Figure 9: Predicted vs True Number of hourly trips in Manhattan

*Manhattan* borough, we can see that the gradient boosting model perform very well for this dataset. Based on Figure 8, the shape of the predicted value line is quite well aligned with the truth value line. Therefore we can only give recommendations for the *Manhattan* region using the Gradient Boosting model.

Our model predicts that the highest demand for the *Manhattan* region in the day is at 6 p.m. Moreover, from 4 p.m. to 8 p.m., there is a high demand for yellow cab rides. Our suggestion for drivers is that drivers who located in the *Manhattan* borough should work in the afternoon shift if they want to increase their work efficiency. Also, taxi company should allocate more drivers to the *Manhattan* borough to meet the increase in demand around the same time period.

In addition to traditional taxi company, we also suggest that Ubers, who is currently teaming up with Waymo to offer rides in fully autonomous cars[9], should take into account the surge in demand around the rush hour period when designing their cars allocation algorithms.

## References


[1] State Government of Victoria. *The case for good design: Transport*. https://www.ovga.vic.gov.au/case-good-design-transport-guide-government.

[2] NACTO. *Ride-hailing services: Opportunities Challenges for Cities*. https://nacto.org/wp-content/uploads/2016/06/Policy-Ride-Hailing-Services-2016.06.pdf.

[3] NYC Taxi and Limousine Commission (TLC). *TLC Trip Record Data*. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

[4] Sabrina Romano. *The History of the NYC Taxi Cab*. https://untappedcities.com/2014/06/11/vintage-nyc-photos-the-past-present-and-future-of-the-nyc-taxi/.

[5] Wikipeida: The Free Encyclopedia. *Taxis of New York City*. https://en.wikipedia.org/wiki/Taxis_of_New_York_City.

[6] Aarian Marshall. *New York Taxi Drivers Hated Uber. Now They're Going to Help It*. https://www.wired.com/story/uber-new-york-taxi/.

[7] Wikipeida: The Free Encyclopedia. *Apparent Temperature*. https://en.wikipedia.org/wiki/Apparent_temperature.

[8] Jason Brownlee. *How to Develop a Gradient Boosting Machine Ensemble in Python*. https://machinelearningmastery.com/gradient-boosting-machine-ensemble-in-python/?fbclid=IwAR3lw95LsQFGX4f-w3ptwHkAgO6RLlg9Q8-YOE-4NZLjKGduFZiWjIrk3Ik.

[9] Kristin Houser. *Fully self-driving cars are finally available on Uber*. https://www.freethink.com/transportation/autonomous-cars.