# Uber and Yellow taxi trip cost prediction in NYC

Aobo Li
Student ID: 1172339
GitHub repository

August 21, 2022

## 1 Introduction

With the emergence of various types of high-volume for-hire vehicle apps, competition between self employed workers with these apps and traditional taxi drivers has become increasingly fierce. As customers, we would like to choose the better one between the two, mainly in terms of trip costs. As companies, decisions need to be made to find maximum profit, or to just survive. This is especially important under the current inflation surge across the globe.

This report would take the perspective of customers, with the objective of forecasting trip cost for Uber(HVFH) and Yellow taxi in New York City, while also provide feedbacks to service companies. Linear regression and Gradient Boosted Tree regression will be used to predict trip cost, with the use of Pearson Correlation coefficients to select related features. Results are compared to find the better service.

### 1.1 Dataset

This report primarily uses data published by the NYC Taxi and Limousine Commission(NYCTLC)[1], recording different types of taxi service data and HVFH data provided by the company. Features in these datasets are explained in their relevant dictionary file provided on the website. Shapefile of New York's different zones is also provided for visualisation purpose.

Daily Weather and Sports event data is obtained for the given time period through web scraping. Both data is publicly available from world-weather.info[2] and mustseennewyork [3].

### 1.2 Data Range selection

Period of the dataset that is analyzed is selected from February 2019 to September 2019, prior to the Covid Pandemic. This specific time period is chosen because, even though Covid still exists now, after two and a half year of epidemic, society has developed a sophisticated action plan to deal with the pandemic, and citizens are returning to their normal life. Trip records prior to the Covid Pandemic is therefore more relevant to current daily life (Inflation is not considered when predicting costs).

# 2 Preprocessing, Analysis, and Geospatial Visualisation

## 2.1 Preprocessing

Although datasets' format provided by NYCTLC is neat and tidy, outliers is still detected in the dataset. Some calculation are also needed so that Yellow taxi and HVFH(Uber)'s data can match. This section will provide detailed data preprocessing, analysis and Visualisation process.

### 2.1.1 Variable calculation

Uber(HVFH) data provided contains base passenger fare, tolls and etc., which does not contain a customer's total cost. Total cost will need to sum these columns up. Yellow taxi's dataset does not contain total trip time, therefore trip time is calculated.

### 2.1.2 Outlier detection & data selection & feature selection

Outliers are detected when inspecting the dataset, HVFH dataset contains several transport mobility provider, select Uber data only. Instance that do not match requirement are filtered.

- Trip distance less or equal 0 and less than 100 miles
- Total cost less or equal 0 and less than $300
- License number of "HV0003" only (HV0003 indicates Uber)
- Trip time over 0 seconds and less than 2 hours
- Correct monthly data

Features of the dataset is then selected.

Uber data:

- pickup_datetime
- PULocationID
- DOLocationID
- trip_miles
- total_amount
- trip_time

Yellow taxi data:

- tpep_pickup_datetime
- passenger_count
- trip_distance
- RatecodeID
- PULocationID
- DOLocationID
- total_amount
- trip_Time

Shared features from Weather and Sports data:

- Day (0 indicate weekday, and 1 indicate weekend)
- Tmid(median temperature)
- Event (1 indicate a sport event, 2 indicate none)
- Weather (categorical variable obtained from weather data, indicate different type of weather)

After removing all outliers from Dataset, the dataset is then joined with Weather data and Sport events data. It is expected that the day of the week correlates with number of trips, and potentially influence trip costs as demands increase. Null values is created in Events when joining dataset, replace null values in Event column as 2, since it indicates no sport events.

Null values are then removed from the dataset, since the volumn of data is very large, removing null values should have little influence to the distribution of data.

After pro-processing with the raw dataset, total raw dataset contains roughly 150 million records, where Uber almost doubles Yellow Taxi's operational trip records (Figure 1).

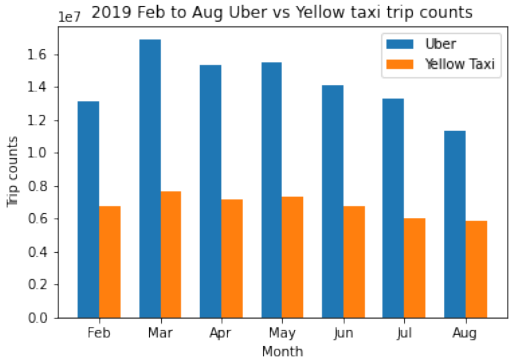


Figure 1: Uber vs Yellow taxi Monthly trip comparison

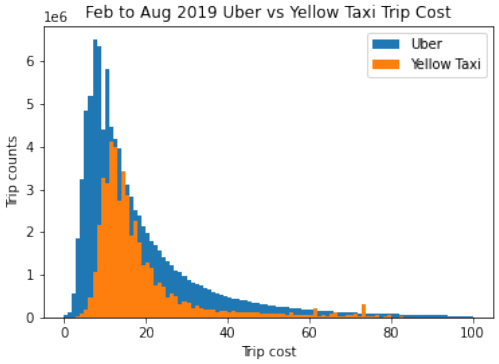


Figure 2: Uber vs Yellow taxi total trip cost

## 2.2 Visual Analysis

### 2.2.1 Total Trip cost analysis

Initially, the distribution of trip costs were believed to follow a normal distribution. However, based on 2019 February to August data, a right skewed distribution of trip costs is observed (Figure 2). This indicates that customers are more likely to spend less money on ridesharing services, with the majority spend less than $30 (Figure 2).

### 2.2.2 Geospatial Analysis: Average Trip cost on Pick up Location ID

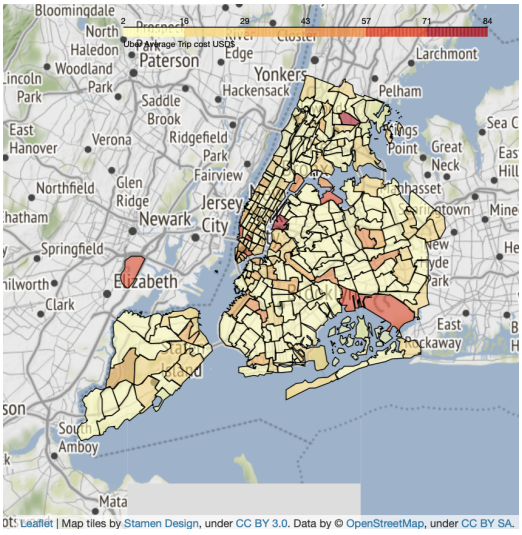


Figure 3: Uber Average Cost per location

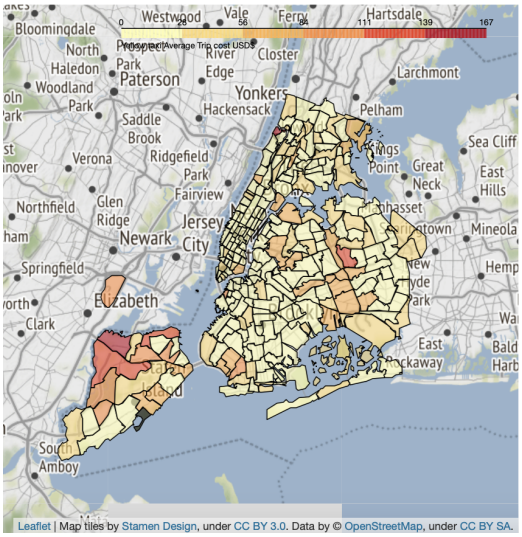


Figure 4: Yellow taxi Average cost per location

From Figure 3 & 4, Uber and Yellow taxi's average trip cost with pick up location seems to be very similar. However, the scale of the color allocated to each location is different. Uber scales from $2-16 per color(Figure 3), whereas Yellow taxi's scale is from $0-28(Figure 4). This is consistent with Figure 2's information, where Yellow taxi generally has a higher cost than Uber in these months.

### 2.2.3 Feature selection

Pearson correlation is used to find the most correlated features against total amount. From the graph, we can see that not all the selected features has strong connections to the total amount. Uber's major features are trip miles, total time and drop off location (Figure 5). Yellow taxi's most influential features are trip miles, total time and ratecodeID (Figure 6). All of the features mentioned above has correlation over 0.1, and these features will be taken as features of the final model.

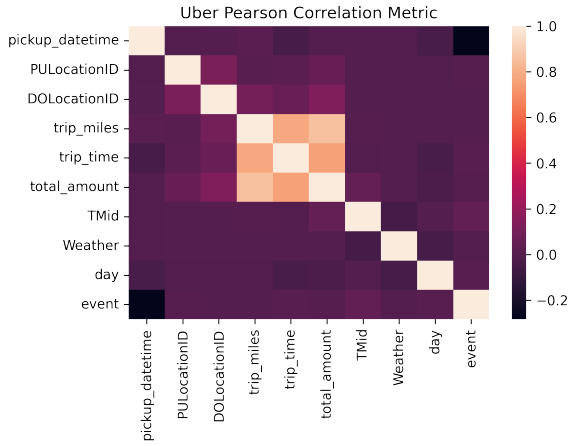


Figure 5: Uber correlation

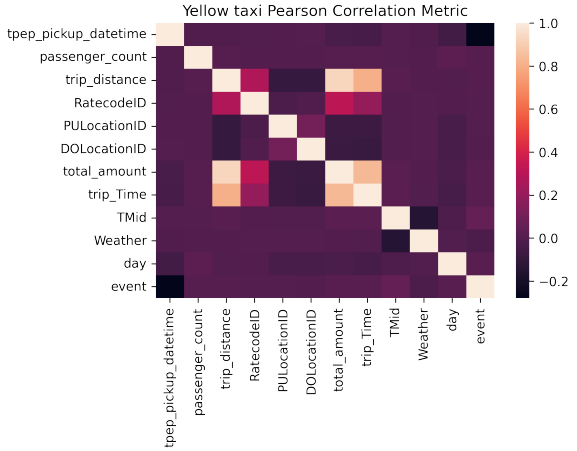


Figure 6: Yellow taxi correlation

Surprisingly, Weather and Sport event data does not seem to he highly correlated with the total amount. This may be due to the fact that weather data is too general, where only daily median temperature is used and different sport is not allocated as different features, but sport event as general.

# 3 Modelling

## 3.1 Models

Since the response variable(total amount) is a continuous variable, regression is chosen to find variable relationship to build a model. Linear regression and Gradient Boosted Tree regression are chosen to fit the training data, each uses the features selected 2.2.3. Each model learns training data(February to August) with the most correlated features. R squared value and Root Mean Square Error are selected to evaluate the model. Even though increasing the number of features will reduce sum of square and increase R squared value, however it does not imply the model is better. Using those uncorrelated features are prone to overfit the model with training data, making it less accurate to predict future data. Therefore, in order to make the model more generalized, only features with correlation coefficient of 0.1 are selected for model training.

### 3.1.1 Linear Regression

Linear regression learns the coefficients of features based on training data, through the form of:

$$f(x) = \beta_0 + \beta_1 x_1 + ..... + \beta_n x_n \tag{1}$$

where $\beta$ is the coefficient and x is feature values, and use the learnt model to predict on future data.

### 3.1.2 Gradient Boosted Tree Regression

As taught in COMP30027, gradient boosting works by using the model to predict on previously false predicted instances, and update the model. Gradient Boosted Tree Regression is essentially an ensemble decision tree. Decision tree can handle continuous outcome by building several decision trees(sequential classifiers) until an appropriate model is build.

## 3.2 Results & Error Analysis

| | R-squared | Root mean square error |
|---|---|---|
| Uber Linear Regression | 0.7780726713191319 | 10.198991462957203 |
| Uber Gradient Boosted Tree Regression | 0.7855092148085336 | 10.026656918656347 |
| Yellow taxi Linear Regression | 0.8766610428084426 | 5.157113068859429 |
| Yellow taxi Gradient Boosted Tree Regression | 0.9257194230193536 | 4.002157276259034 |

Table 1: Uber & Yellow Taxi Evaluation metrix

Based on the results of Table 1, it is shown that Gradient Boost Tree Regression as a more complex model, has better performance in predicting the September's record with both Yellow taxi and Uber data.

R-squared value indicate the amount of data points that can be explained by the regression model. 78% of Uber and 92% of Yellow taxi's prediction can be explained by the selected predictor features. With $4 of the average of squared differences between Yellow Taxi's prediction and the actual total costs. This is shown in Figure 7 and Figure 8, where the predictions generally matches True trip total costs.

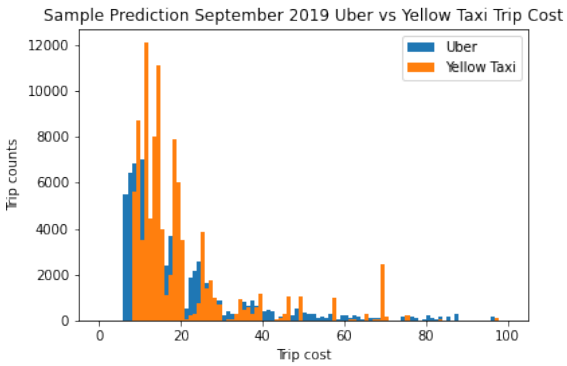


Figure 7: Model Prediction

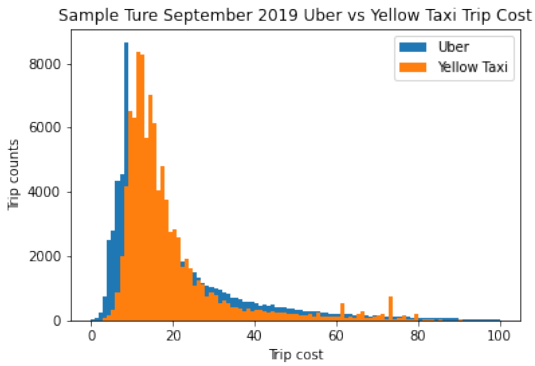


Figure 8: True cost

However, some of the errors observed in the model are worth noting of.

First, prediction of Uber and Yellow Taxi 7 both shows a pattern where Uber has no prediction with less than around $5 and Yellow taxi has no prediction below roughly $9. As it is a ensemble decision

tree model, it can predict continuous outcomes, however, decision trees will make more of an interval of prediction based on number of trees, as it is shown in the prediction histogram plot with 8 major intervals.

Second, Uber's Gradient Boosted Tree Regression has a very high root mean square error and a relatively low R-squared value. Possible reason for this outcome is due to Uber's different type of ride. Uber provide many different type of rides, including Uber comfort, UberXL, Uber Green and etc[4]. With different type of ride, single trip's total cost could vary significantly, with a similar trip distance and trip time. This data is not provided by NYCTLC, Uber's ride type relation with trip's total amount costed cannot be accessed, making the model less accurate, and lead to this high Root Mean Square Error value.

### 3.3 Yellow Taxi & Uber trip cost comparison

Since Gradient Boosted Tree Regression performs better, a new model is build to compare average trip cost difference between Uber and Yellow Taxi. Features in the new model are selected as trip distance and trip time for both dataset. The prediction will be based on September Yellow taxi's data, each model will learn based on their training data, and predict on yellow taxi's data since features are identical.

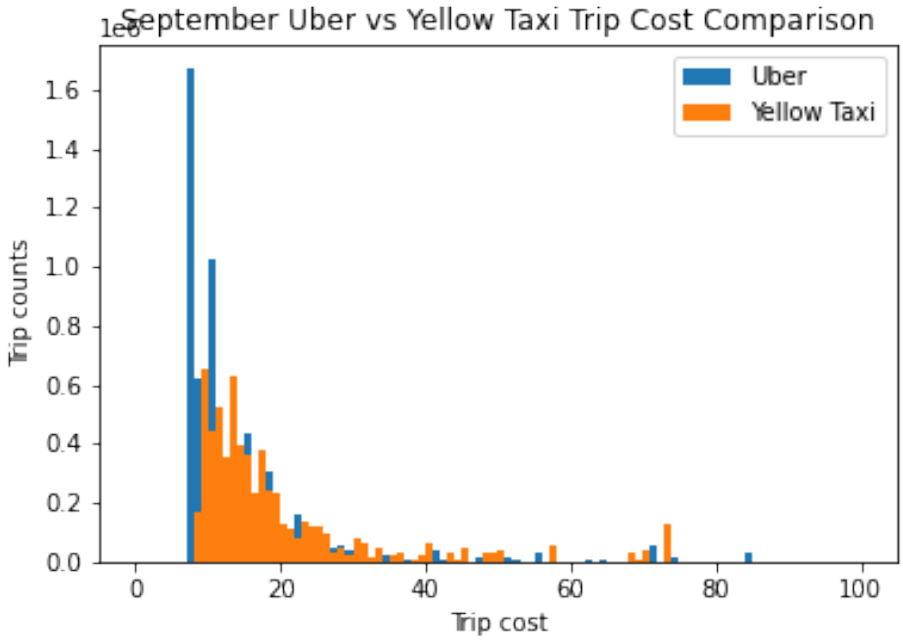


Figure 9: Uber & Yellow Taxi prediction comparison

| | R-squared | Root mean square error |
|---|---|---|
| Uber Gradient Boosted Tree Regression | 0.7309230325264138 | 7.6171898402535625 |
| Yellow taxi Gradient Boosted Tree Regression | 0.8579657875940216 | 5.534176437242933 |

Table 2: Uber & Yellow Taxi prediction comparison Metric

With a reduced feature, both model's R-squared value is lower than the previous model, however it can still be used to compare Uber and Yellow taxi's trip cost (Table 2). Based on Figure 9, we can see

that Uber's trip cost are generally lower than Yellow taxi even with the same prediction data. Yellow taxi's average trip cost is also $3.28 more than Uber's average cost by calculation.

# 4 Recommendations

With the surging inflation in US (9.1% by July 2022[5]), trip cost will becoms more important as it relates to people's spending. Even though the models are very simple, it still demonstrates that Uber is cheaper than Yellow Taxi in New York City.

Recommandation For customers:
Uber on average is $3.28 cheaper than Yellow Taxi. Therefore Uber would be the better option for getting around within the New York City. Uber also offers more ride services than Yellow taxi, making it a more favourable competitor than Yellow taxi. Even though Taxi service are also implementing e-hail app for pre-ordered trips, it's main rival Uber is still beating Yellow Taxi in price. Yellow taxi would only be a better option if a trip is urgent and customer has to hail on street, as it is the only service that can respond to a street hail in all 5 boroughs. In short, only choose Yellow taxi if the customer needs an urgent ride, otherwise, Uber is the better option.

Recommandation For Companies:
Not much recommendation is needed for Uber, however Yellow Taxi is facing serious problems. Uber's operation has doubled Yellow taxi's trip, and continuing to take shares from Yellow taxi. With lower price and the ongoing economic recession, customers are more likely to choose Uber rather than Yellow taxi.

It's not feasible to simply reduce Yellow Taxi's fare as companies and drives are already suffering from Covid Pandemic[6]. Yellow taxi companies need to seek help from Government to keep the business operating, in the form of funding that cuts taxi fare but provide taxi drivers with subsities, or cash back from the government if a customer choose Yellow Taxi for their trip. Vice Versa, if New York wants to keep it's Iconic Yellow Taxi, funding is needed to support Yellow taxi companies from closing down. Nonetheless, it is foreseeable that Yellow taxi will be obsolete by time. Some Yellow Taxi will still remain, but may no longer operating primarily as taxi service, instead operating as a symbol of New York. Even if Yellow taxi can withstand the economic pressure, it's trip share are very likely to be taken by Uber or similar companies.

# 5 Conclusion

Considering the models are fairly simple, they still show the potential of providing more accurate predictions with more features. The external dataset collected in this report doesn't help with model prediction, mainly because the data is not precise enough. Different type of sports should be classified as several different features, while weather data should contain hourly statistics. Precipitation, wind speed is not considered in this project, but are likely to be influential on trip cost as it could influence individual's will of taking a taxi/Uber, and increase traffic during that specific time.

The investigation of this report is limited by the amount of features in the dataset, therefore it is highly recommended that future studies needs to use more external dataset relevant to the study, for example New York's Traffic data, as Uber and Taxi both charges more during peak hours due to high demand and increasing trip duration caused by heavy traffic. Data cleaning is fairly simple in this study, therefore additional data selection is also highly recommended. Even though Linear regression and Gradient Boosted Tree Regression may provide high accuracy, future studies are still

recommended to use a more complex model, as simple models cannot catch the complex interactions between different features.

## References


[1] City of New York. *NYC Taxi and Limousine Commission*. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: 2022-08-03.

[2] Foreca. *New York City Weather records*. https://world-weather.info/forecast/usa/new_york/february-2019/. Accessed: 2022-08-03.

[3] Must See New York. *New York City Sport events records*. http://www.mustseenewyork.com/new-york-top-sports.html. Accessed: 2022-08-03.

[4] Uber. *Uber Ride Type*. https://www.uber.com/us/en/ride/ride-options/. Accessed: 2022-08-20.

[5] Christopher Rugaber. *US inflation surges again in June, raising risks for economy*. https://apnews.com/article/inflation-economy-prices-consumer-74e1a5c9bced40460e4079f62e980095. Accessed: 2022-08-20.

[6] Jack Denton. *Where Did All the Yellow Cabs Go?* https://www.curbed.com/2021/05/nyc-yellow-cabs-taxis-disappearing.html. Accessed: 2022-08-20.