Understanding and Predicting the Tipping Behaviour of NYC Yellow Taxi Passengers

Wednesday 17th August, 2022

1 Introduction

Tipping, while not as prominent in Australia, is considered a social norm in many places throughout the world and often accounts for a significant chunk of the salary of workers employed in tipping industries (e.g. taxi, food, entertainment, etc.). In fact, tipping alone generated $47 billion USD in the food industry in 2011 [1].

In this report, we will be looking to understand the tipping behaviour of New York City (NYC) Yellow Taxi passengers while attempting to uncover the factors that influence them, for the purpose of increasing the tips that Yellow Taxi drivers receive.

2 Data sets

For this investigation, we used two data sets from 2018. We will refer to them as the "Taxi Data" and the "Weather Data". We chose 2018 as we were looking to investigate an average year for the taxi industry and the end of 2019 onward had been affected by COVID-19. We chose only 8 months due to memory and other hardware restraints.

2.1 Taxi Data

This data is published yearly by the NYC Taxi Limousine Commission (NYCTLC) [2] and it records statistics related to each trip taken by different taxis (Yellow, Green, FHV, HVFHV) in NYC. In this report, we will be focusing on the data for Yellow Taxis from January 2018 to August 2018. This data set contains 19 different features describing just under 70 million taxi trips. The NYCTLC also provides a shapefile describing location around the city. This will be used for geospacial visualisation.

2.2 Weather Data

In addition to the Taxi Data, we have chosen a data set outlining the hourly conditions of NYC weather to explore the tipping behaviour of customers. Anecdotally, weather conditions such as rain can often impact the mood of people. This could influence the likelihood of passengers leaving tips for their drivers.

The Weather Data is published by the National Oceanic and Atmospheric Association (NOAA) National Centers for Environmental Information [3]. It contains 93 features describing 9085 hours of weather around JFK Airport in 2018.

3 Preprocessing

Before conducting any analysis or modelling, the data first had to be cleaned of outliers and discrepancies.

3.1 Cleaning

In general, the more features we have to train our model on, the greater the accuracy of our model will be (too many may lead to overfitting). However, we can confidently say many of the features listed in our data sets have no casual relationship with the tipping behaviour of customers so they were removed. Furthermore, some features that may have been useful had to be excluded as they were not available for enough instances (e.g congestion surcharge for the Taxi Data and daily snow depth for the Weather Data).

Ultimately, the following features were selected from the Taxi Data:

- Number of Passengers
- Trip Distance
- Pick-up Location ID
- Drop-off Location ID
- Tip Amount
- Toll Fee
- Total Amount
- Day and Time
- Extra Fees
- MTA Tax
- Improvement Surcharge

And the following from the Weather Data:

- Wind Speed
- Visibility Distance
- Air Temperature
- Precipitation Depth

3.2 Null Values

Unfortunately, the Weather Data contained many null values (missing values). We had 251 null values for Wind Speed and Visibility Distance, 252 for Air Temperature, and 2056 for Precipitation Depth. Removal of these instances seemed unreasonable as this would mean removal of almost 23 percent of our data just from Precipitation Depth alone. The solution to this problem involved interpolation. This entails estimating unknown values between known values. We used forward linear interpolation with a limit of 2 (if more than 2 null values are in consecutive order then none will be interpolated) to fill all missing values for Wind Speed, Visibility Distance, and Air Temperature. There were still 79 null values remaining for Precipitation Depth. Although interpolating hourly weather data is usually safe, estimating more than 2 consecutive missing values may bring unnecessary risk. Thus, the remaining 79 instances were removed.

3.3 Feature Engineering

We looked to further refine our selected features into features with better predictability power as well as fabricating new ones. A **Month** feature was created as the month may have an impact on tip amounts. We also created a **Supplementary Fees** feature which is the summation of the extra fees, MTA tax and improvement surcharge. We removed the features it was based on. The thinking behind this was that customers who are charged a greater supplementary fee on top of their fare amount will be likely to tip less. We also created a **Weekend** feature which denotes whether a trip was on the weekend or not. It would not be far-fetched to predict that passengers on the weekend tip a greater

amount. The final constructed feature is **Time Duration**. We suspected this may be a relationship worthy of investigation.

The data dictionary for the Taxi Data tells us that only tips left by credit card are reported. It would make little sense to include trips that were not paid by credit card. This left us with 48.2 million trip reports.

Upon inspection of the Weather Data, we found conflicting data for some instances with the same time and date. We removed the inconsistent data by only keeping the first instance of each time and date. This resulted in the removal of 14 data points. The Weather Data reported the data at odd intervals. It was not on the hour and contained random instances (e.g instances 15 minutes apart). This made it somewhat difficult to join the two data sets together. Also, to model the relationship of the above weather features and tipping amount, it is sufficient to compare the daily weather conditions rather than the hourly conditions. We decided to aggregate the Weather Data by day and use the average of the values listed above.

The feature set selected undoubtedly possesses some degree of multicollinearity but we will address this later on.

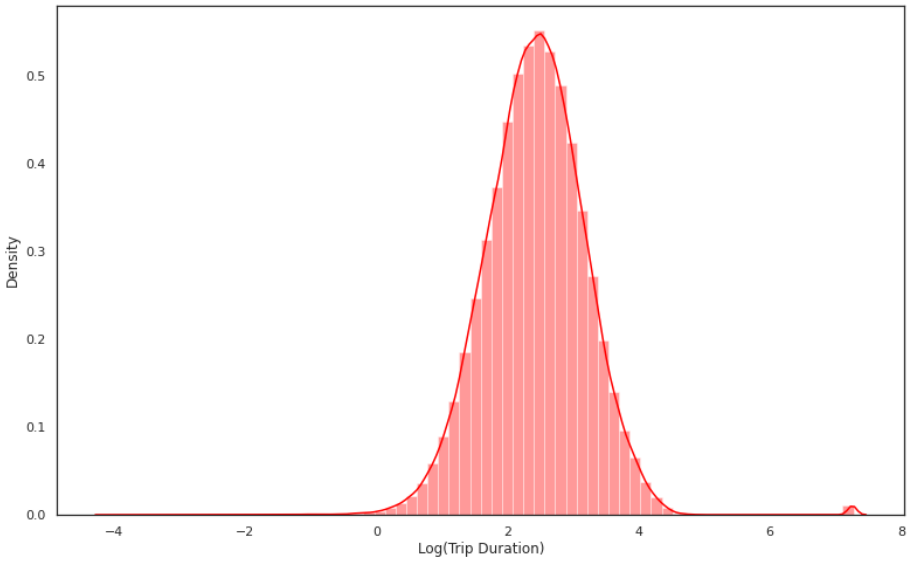

Figure 1: Histogram of the Log of Trip Duration

3.4 Outlier Removal

We created rules based on the data dictionaries for both data sets to remove data points that clearly showed discrepancies to what was expected. In summary, we removed the following instances:

- Pick-up Location ID above 263 or under 1.
- Drop-off Location ID above 263 or under 1.
- Zero trip distance
- Negative trip duration

- Zero total amount paid
- Trips that had pick up or drop off time before 2018-01-01 00:00:00
- Trips that had pick up or drop off time after 2018-08-31 23:59:59

Following this, we were left with 46.9 million trip reports.

To handle outliers in tip amounts we removed any data points that tipped more than 15 percent of $723.9 USD. $723.9 was the highest total amount (not including tips) cost of any trip and 15 is the average percentage of the total amount that is usually expected [4]. It is unrealistic to presume that any average tip will be higher than this.

Two features that still contained many outliers were the trip duration and tip amounts. We plotted the histogram of the trip duration and recognised that it followed a log-normal distribution, as can be seen in Figure 1. This allowed us to remove any data points that were more than 2 standard deviations away from the mean resulting in the removal of 90110 instances.

4 Analysis

In this section, we will observe and analyse the univariate and multivariate distributions and correlations in our data. Due to time and hardware constraints, we took a random sample of 10 percent from our population to conduct analysis on.

4.1 Tip Amounts

To gain a better understanding of our data, we plotted the distribution of Tip Amount.

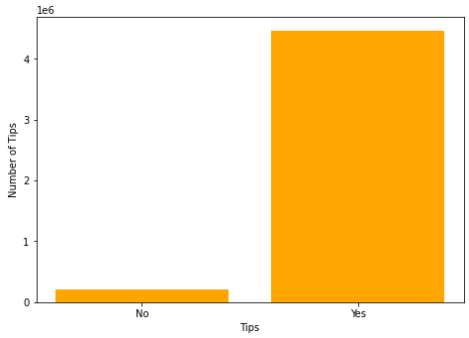

Figure 2: Distribution of Tip vs No Tip per Trip

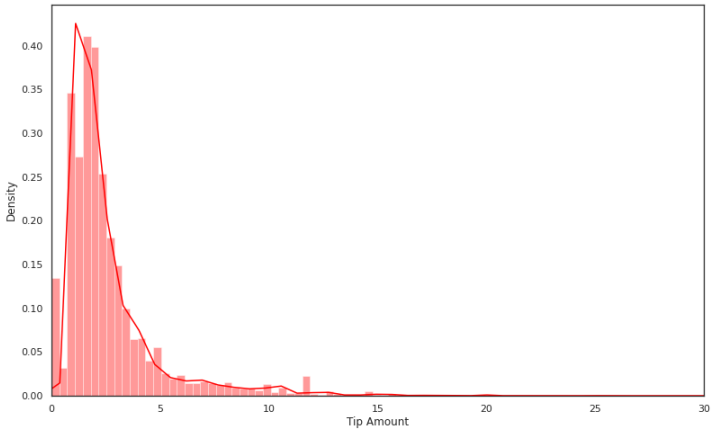

Figure 3: Histogram of Tip Amount per Trip

We can observe that an overwhelming 95.4 percent of passengers tipped their driver from figure 2. With tips so common in NYC Yellow Taxi trips, it is apparent that understanding the tipping behaviour of customers to maximise tips can have a substantial affect on the salary of taxi drivers.

The histogram of Tip Amount per trip shows us the distribution of Tip Amount in Figure 3. Most instances are clustered around the mean of $2.64. The distribution has a positive skew with a long tail reaching far from the mean. The distribution initially seems Gaussian with the left tail cut-off. This lead us to test the hypothesis that the population this sample came from has a normal distribution. However, upon conducting the Anderson–Darling test on our sample, we obtained a test statistic of

45074.825 which is much larger than the critical value of 0.787. Thus, we can reject the null at the 5 percent significance level and conclude that the sample does not come from a normal distribution.

The formula for the Anderson-Darling Test is:

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[\ln F(X_i) + \ln 1 - F(X_{n-i+1})] \tag{1}$$

Where n = the sample size, F(x) = CDF of the distribution (normal in our case), i = ith sample, calculated when the data is sorted in ascending order.

4.2 Correlation Analysis

We created a correlation heat map, in order to gain an understanding of which features influence the movement of Tip Amount and to visualise collinearity in the data.

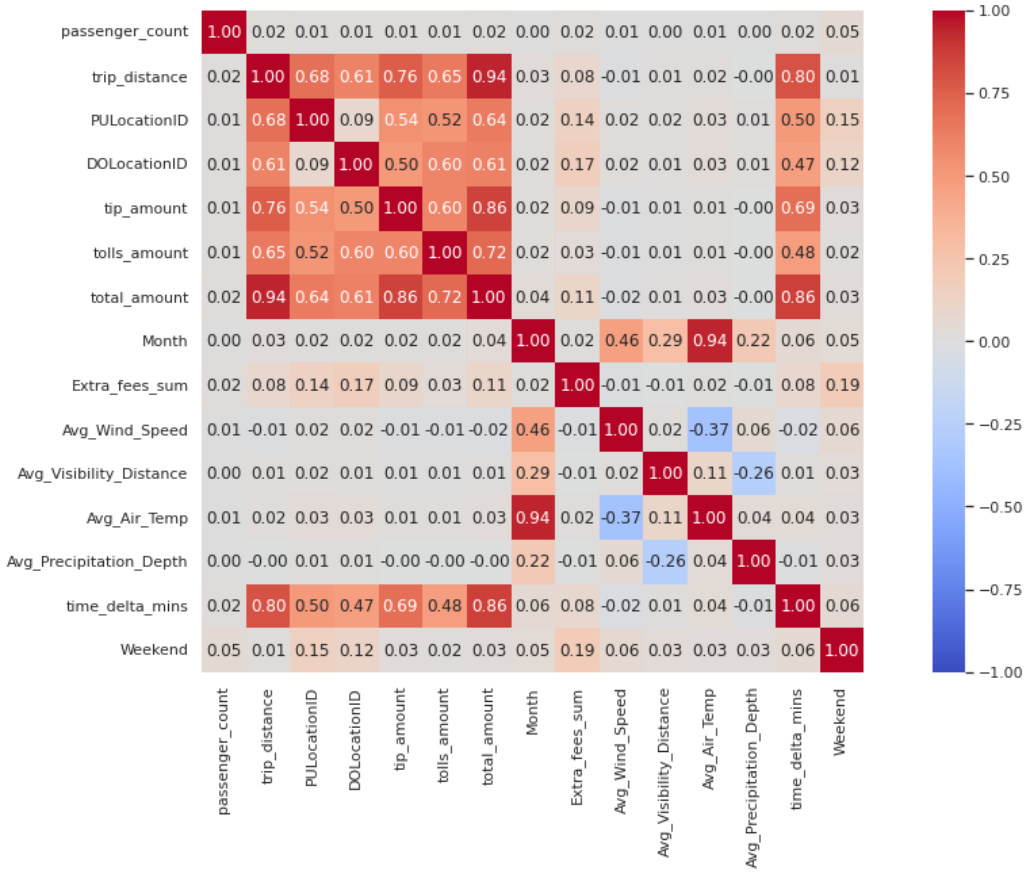

Figure 4: Correlation Map of Selected Features

As we can observe from Figure 4, Tip Amount has close to no correlation with any features from our Weather Data. The correlation with the Supplementary Fees feature is also negligible. This may seem somewhat surprising at first but we will discuss why this may be the case in the "Discussion" section.

The features that have considerable correlation with Tip Amount are Trip Distance, Pick-up Location ID, Drop-off Location ID, Tolls Amount, Total Amount, and Time Duration. We will discuss which of these features to use for our final model in the "Feature Selection" section.

4.3 Visualisation of Pick-up and Drop-off Locations

Pick-up and Drop-off Location ID having a strong correlation with Tip Amount is something that we might not have surmised at first. We will attempt to visualise this relationship. As can be seen

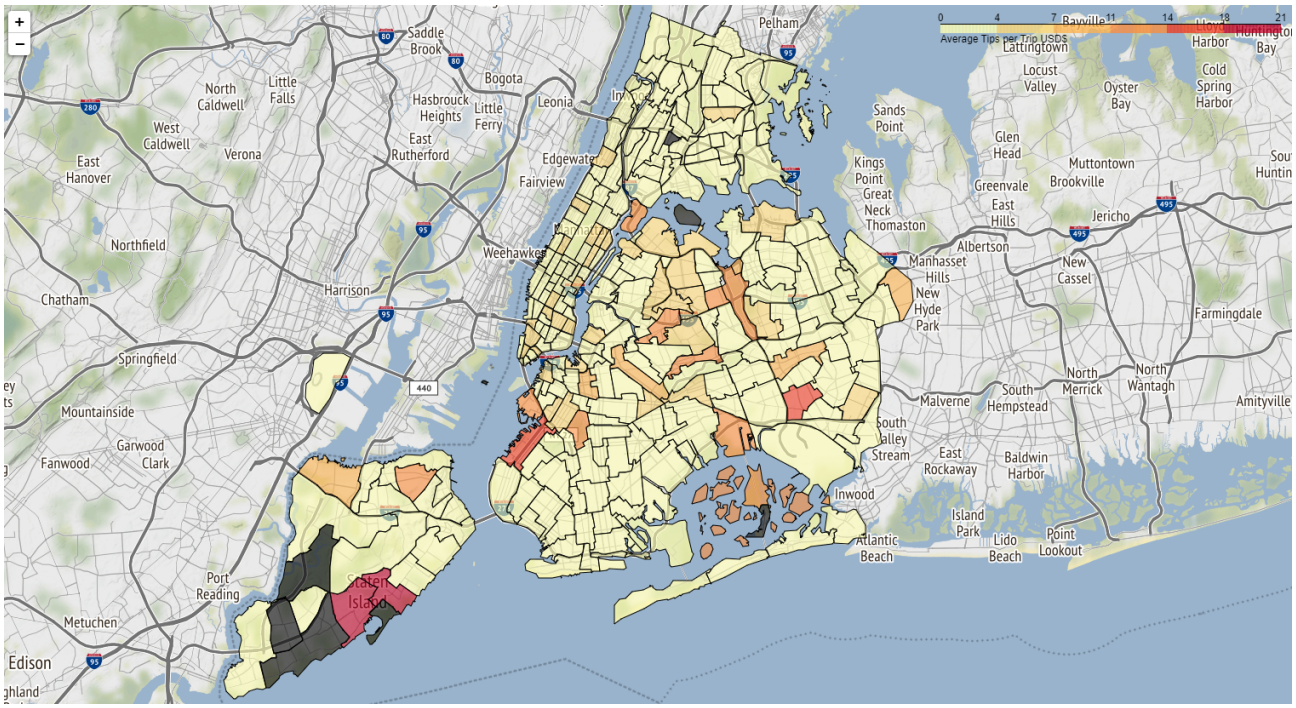

Figure 5: Average Tips per Trip by Pick-up Location (USD)

from Figure 5, some pick-up locations provide drivers with much higher tips per trip than the mean of $2.64. This could potentially be used by taxi drivers to emphasise picking up passengers from the high tipping areas. Figure 6 shows us the top 10 hot-spots for tips. This can further help taxi drivers

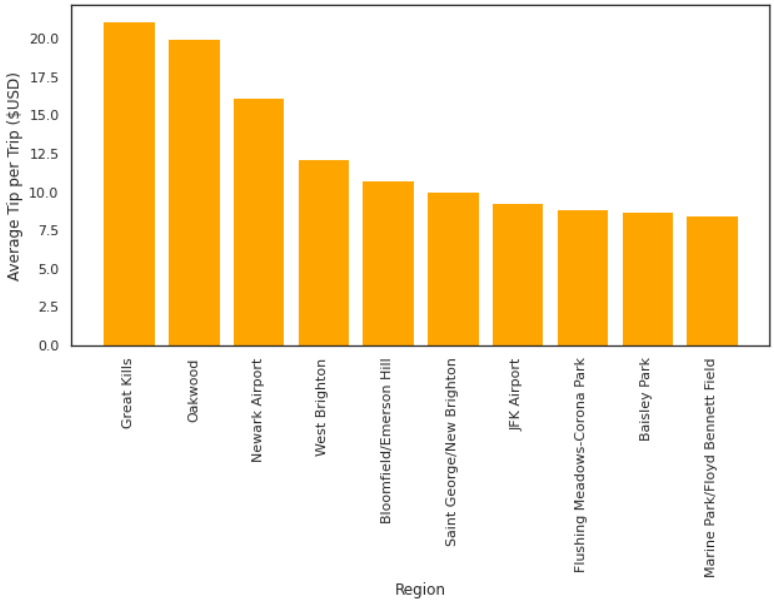

Figure 6: Top 10 Tipping locations per Trip

plan their routes through high tipping pick-up locations.

4.4 Feature Selection

The features that have a high correlation with Tip Amount, also seem to have a high degree of correlation amongst each other. This is referred to as multicollinearity and it can impact our regression results by changing the weight of each coefficient. This can reduce the precision of the estimated coefficients and the accuracy of our model. We used a method to detect and minimise multicollinearity called VIF (Variable Inflation Factors).

Table 1: VIF of Selected Features

Features	VIF
Trip Distance	14.782
Pick-up Location ID	4.742
Drop-off Location ID	4.386
Tolls Amount	2.463
Time Duration	12.353
Total Amount	41.857

Table 2: VIF of Selected Features without Total Amount

Features	VIF
Trip Distance	6.123
Pick-up Location ID	4.324
Drop-off Location ID	4.148
Tolls Amount	1.814
Time Duration	6.800

A VIF exceeding 10 is indicative of high multicollinearity between that feature and others. In Table 1, we can see that 3 features are above this limit. Upon removing Total Amount from the feature set, all of our features fall below the threshold, as can be seen in Table 2. We will use the remaining features in our model.

5 Modelling

In order to predict future tip amounts, we will be utilizing two machine learning techniques, linear regression and random forest regression. We have chosen these as they are trusted techniques that work well with large data sets. We must state some assumptions to use linear regression: the relationship with our features and Tip Amount are linear, we have removed most of the multicollinearity, homoscedasticity, and independence of taxi trips.

Before we can model our data, we must first deploy one-hot-encoding on our categorical data (Pick-up and Drop-off Location ID). This is the process of converting categorical data into a series of binary vectors. Following this step, a transformer should be used to combine the columns of features into a single feature vector, which can be easily used to train machine learning models.

5.1 Model Comparison

We randomly split our data into a training and testing set, then trained the models over a set of hyperparameters. We measured the performance using the R squared and R Mean Square Error (RMSE) of the training and testing sets to understand which model is more suitable. Table 3 shows us the performance of the linear regression. We usually expect a higher R squared on the training data, however, this value is much higher than the one of the testing set. This may be indicative of overfitting. Further testing is needed to confirm. We can observe from Table 4 that the random forest regression has all around worse performance than the linear regression. Based on this, it would be wise to pick the linear regression as our final model.

The linear regression model can be used in the future by taxi companies to predict the tipping behaviour of customers. A greater priority can be given to trips that are predicted to generate more income from tips.

Table 3: Linear Regression

R^2	RMSE	Data set
0.837	1.466	Training Data
0.646	1.472	Testing Data

Table 4: Random Forest Regression

R^2	RMSE	Data set
0.613	1.523	Training Data
0.604	1.544	Testing Data

6 Discussion

We initially estimated that weather would have a moderate impact on the tipping behaviour of NYC Yellow Taxi passengers. Mood is often thought of as being associated with the weather, so it would seem reasonable for the weather to impact tipping behaviour. However, a 2008 study [5] shows that this may not be the case. In fact, the study reported that weather has zero impact on mood if the individual is not unhappy. This makes our findings more explainable, but the outcome is still somewhat disappointing.

Overall, our models were able to predict the Tip Amount of trips to an adequate degree, however, their performances can certainly be improved. More research should be done into regression models that can work with heavily interlinked feature sets. Steps were taken to combat multicollinearity, yet the remaining features are still clearly correlated - take Trip Distance and Time Duration for example. General Linear Models may be of interest to research in this regard. The chosen models can further be improved with greater and stricter hyperparameter (max number of trees for random forest regression).

The relationship with Pick-up and Drop-off Locations and Tip Amount should be pursued further using different external data sets. Data concerning socioeconomic factors such as gender, age, education and income may prove useful. Other relationships with Tip Amount should also be pursued in order to maximise the potential tipping behaviour of customers. The friendliness and loquaciousness of the driver, the cleanliness of the taxi, card-less tap and pay options inside the vehicle are all things that could be assessed, although finding the relevant data may prove challenging

7 Recommendations

Taxi drivers and taxi companies who are reading this report should look to maximise profits through tips by placing greater emphasis on providing service to customers in high tipping areas. The area between Queens and Brooklyn is a hot spot for tips and may continue being so in the future. Airports are also good options. They should also look to accept trips with the destination in a high tipping area. This will further maximize tips.

Steps may be taken to increase the time duration of the trip in order to increase tips further. This needs to be implemented in an ethical manner.

References

[1] Ofer H Azar. "Business strategy and the social norm of tipping". In: *Journal of Economic psychology* 32.3 (2011), pp. 515–525.

[2] NYCTLC. *TLC Trip Record Data*. URL: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

[3] NOAA. *Integrated Surface Dataset (Global)*. URL: https://www.ncei.noaa.gov/access/search/data-search/global-hourly.

[4] Charity Cab. *7 TIPS FOR TIPPING YOUR TAXI CAB DRIVER*. URL: https://charitycab.com/7-tips-tipping-taxi-cab-driver/.

[5] Jaap JA Denissen et al. "The effects of weather on daily mood: a multilevel approach." In: *Emotion* 8.5 (2008), p. 662.