Taxi Tipping Habits in New York City

John Melleuish
Student ID: 1171367
Github repository

August 21, 2022

1 Introduction

Tipping in the United States is often seen as customary especially in the hospitality industry. Not tipping your server would spark disagreements and judgemental looks amongst the party you're with. But is it the same with taxi drivers? Do people see tipping as compulsory? If so, what external factors influence someones level of generosity?

This report will investigate from a taxi drivers perspective what kind of environment they can put themselves in, in order to maximise their chances of receiving a gratuity.

Two datasets were used in this analysis; New York City (NYC) yellow and green taxis dataset (184397591 rows x 19 columns) as well as weather data from LaGaurdia Airport (28361 rows $\times$ 102 columns). Both Yellow and Green NYC taxis were chosen for this investigation due to the relevant datasets containing almost identical variables captured, including tip amount. Ride Share data (Uber, Lyft, private limousines etc.) was available, however they did not make the amount tipped per ride available and thus couldn't be use.

The timeline of this analysis was taken from January 2018 through to December 2019. This period was chosen as it is the most recent data not effected by the COVID-19 pandemic. Although data from 2020 on wards is newer, it has a whole range of factors that effect the reliability of the conclusions that will be drawn from it. As we move out of the pandemic, taxi drivers will find data without a pandemic asterisk far more valuable as people's decisions around tipping would have no doubt changed.

2 Preprocessing

2.1 Taxi Dataset

This report primarily uses data from the NYC Taxi and Limousine Commission [1]. It is comprised of 184,397,591 instances with 19 variables that were measured for every trip taken taken. Due to this report being an analysis into tipping habits, not all of these variables were relevant to the investigation. 9 variables in total were concluded to have a potential to effect someones likelihood to tip. These included:

- Pick-up Location
- Drop-off Location
- Passenger Count
- Trip Distance
- Trip Time
- Fare Amount
- Surcharge Amount
- Total Amount
- Time of Day

The dataset was then cleaned up via invalidation of records. For each variables, a common sense test was performed for which values were plausible to have. For example, the passenger count variable

had many records in the hundreds which is clearly an error and thus we used the legal maximum of 5 occupants as the maximum and 1 occupant as the minimum. Similar rules were applied for the rest of the variables. Once complete 7,882,213 instance remained (4.27% of the original dataset).

2.2 Weather Dataset

A weather dataset was also used to support this analysis as it is believed weather can have a strong effect on someones mood and potential tip generosity. The data was taken from The National Centers for Environmental Information [2] at the weather station located at LaGaurdia Airport. This Airport was chosen as it is the most centrally located airport in NYC and thus would capture the average weather on each day more accurately.

The dataset came with 28,361 instances with 100 variables of highly precise measurements of which most were considered insignificant to the rate of tipping. 5 of the most common weather variables were chosen as they are widely available on a daily basis and understood to most, these include:

- Temperature
- Dew Point
- Pressure
- Wind Speed
- Wind Direction

The dataset was then cleaned up through invalidating records that fit outside plausible measurements. This was done by finding what the maximum and minimum recorded in NYC history were and setting them to be the upper and lower bounds that the data must fit inside. Once complete, 4841 instances remained (17.06% of the original dataset)

The two datasets were then joined through an inner join on their mutual data variable, the dataset produced consisted of 3,435,336 instances with the 16 variables discussed above.

3 Variable Analysis

3.1 Pickup Location (Geospatial Visualisation)

In the dataset there were 265 taxi zones across NYC. Figure 1 shows the distribution of the average tips given per taxi ride across the city:

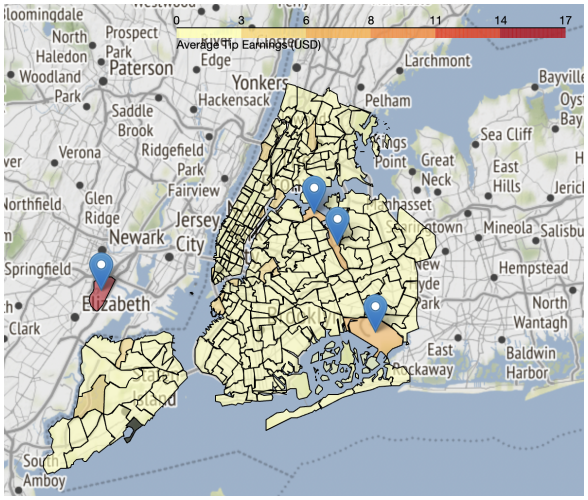

Figure 1: Distribution of Tips across NYC

As visible, there are 4 major outliers in this dataset highlighted with a pop-up icon, with the rest being fairly uniformly distributed under $3 per trip. These outliers are the locations of the 3 major

airports in NYC as well as Flushing Meadow Park (home of the US open and NY Mets baseball team). However, of these, Newmark Airport has by far the greatest average tip value. This is due to it being the furtherest airport from the centre of the city and thus people tend to have longer taxi trips (potentially greater tips) when being picked up from this airport.

Flushing Meadows is an interesting result from this analysis due to it being the only outlier, which is not an airport. Due to it being one of NYC most prominent parks, host to a wide range of events and activities, this result fits with our above analysis of the airports as placing people from all over NYC are likely to visit (hence longer trip, greater tips).

3.2 Fare Amount

In popular culture, the most tips are worked out to be a percentage of the raw bill. Can this conclusion be the same for taxis? Figure 2* shows the relationship between the fare amount and how much someone tipped. The 3 common tipping percentages (15%, 20%, 33%) were added to the graph to help visualise potential relationships.

* note: only 10% of the data is displayed in order to see the relationships clearer.

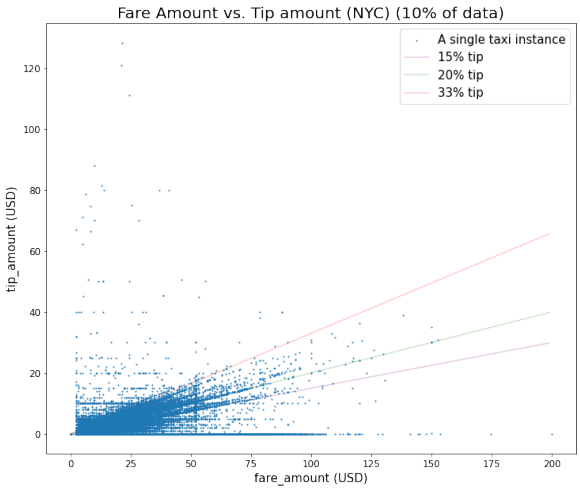

Figure 2: Fare Amount vs. Tips

As displayed, it is evident that the patterns of people tipping are correlated with the fare amount. Looking horizontally, it can be seen that it is quite common for someone to tip a whole number amount most notably $0. However, a great percentage of tips (51.99%) fall in the range between 15 and 33% (78.92% if you remove those that didn't tip).

Interestingly, there is a greater variance in tipping amount for small fares. It appears as though that the higher the fare amount is, the more likely someone is to either tip between 15 and 33% or not tip at all.

The most interesting part of this figure is $0 tip amount with 34.13% of instances recorded no tips to the driver. Although majority of those that didn't tip had fare amounts less than $50, many people still did not tip the driver well above this threshold. This goes against the assumption of America being a culture of compulsory tipping.

3.3 Time of Day

Human moods tend to fluctuate throughout the day with base on our circadian rhythm [3], thus it is no surprise that tipping generosity may fluctuate throughout the day. Figure 3 shows the ratio of people tipping to those that don't throughout the day. Figure 4 shows the average amount someone will tip and the average amount someone will tip excluding records with the tip amount being zero dollars.

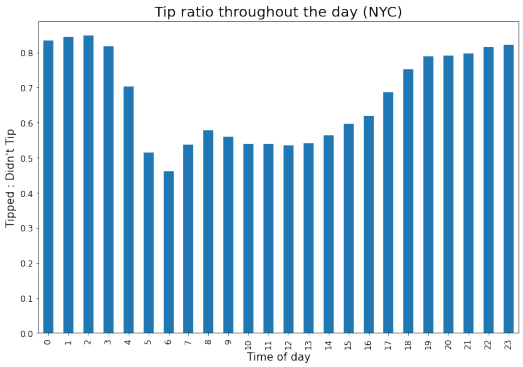

Figure 3: Tipping Ratio

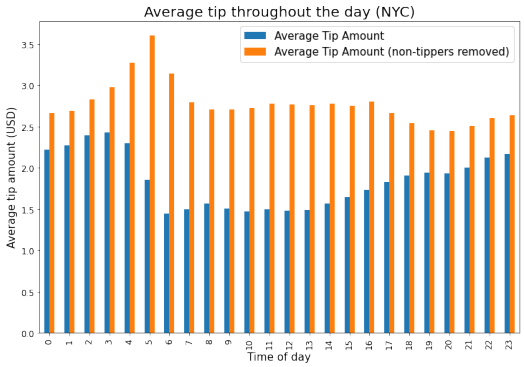

Figure 4: Average Tips

Figure 3 shows that a higher proportion of people tip at night compared to during the day. This can be hypothesised as people during the day are often only travelling short distances to and from work, for lunch or meetings. Whereas at night, people will often be in the spending mood as they go out to club, bars and restaurants and are expected to tip more.

Figure 4 shows a far more interesting relationship. In blue are the average tip amounts through the day which correlated fairly closely to Figure 3. However, when you remove those that didn't tip from the data, you see a far more uniform distribution with the expectation of the hours surrounding 5am. This means that people taking a taxi in the early hours of the morning are not very likely to tip but when they do tip, they tend to tip far above normal levels. This is believed to be due to a large portion of daily commuters not often tipping and party goers being very generous towards drivers for getting them home.

3.4 Temperature

The temperature in NYC can be measured on the extremes. In the winter, they experience snow storms and in the summer, temperatures reaching above 35C (95F). Figure 5 shows the likelihood of someone tipping a driver based on various temperatures.

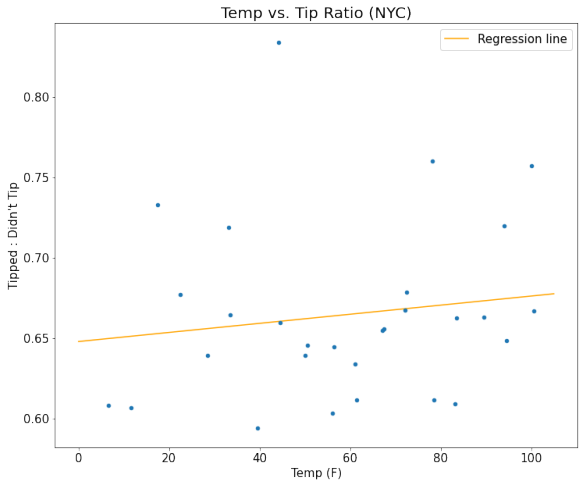

Figure 5: Temp vs. Tips Ratio

Figure 5 shows a minimal relationship between the two variables. There does appear to be a slight positive correlation but the variance is too large to make any definitive statement. This result was quite surprising as I thought people in extreme temperatures would've tipped more often.

4 Modelling

4.1 Feature Selection

When fitting a machine learning model, it is important to only use the variables that have a significant correlation with the largest variable. Starting with 11 variables, we normalised the variables in order to not unfairly weight any variable more than another. We then recursively eliminated the least significant variable in the model until we had 5 variables remaining. Those were:

- Fare Amount
- Trip Distance
- Time Taken
- Hour of the Day
- Temperature

It is important to note that for each of the models, all instances with zero tip amounts were removed in order to fit the models better.

4.2 Neural Network

A neural network is a type of supervised machine learning in which, through hidden layer, learns to distinguish between instances in order to predict the outcomes of unseen data. This neural network used 5 input variables chosen from the feature selection method outlined above, two hidden layers comprised of 5 nodes each, and an output layer of a single prediction node (tip amount).

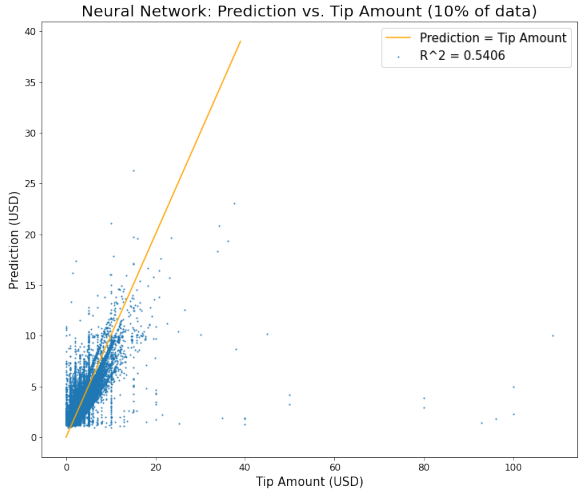

Figure 6: Neural Network

Each layer of the network had a RELU activation function, chosen as it weights all values less than zero as zero and then a one to one for positive values, this is an important feature to have as we are trying to predict a non-negative amount.

The model was also configured during training on a gradient descent optimiser and a mean squared estimator as the loss function. This was because it is a fast method and works best for continuous variables.

Once the network was trained (through 10 epochs), and then tested against the 20% holdout set, it produced a result of $R^2 = 0.5406$. Figure 6 is the resulting graph of the predictions made by the neural network compared to the ground truth values in the dataset. This result is quite impressive considering have many personal factors (that were not measured) go into someone's likelihood to tip.

4.3 Linear Regression

Linear regression is used to model the effect of multiple predictor variables on the target variable. This linear regression used the 5 variables chosen using feature selection to predict the amount an individual will tip. The following is the result of the regression analysis:

$$Tip = 1.41 * Fare + 0.17 * Distance + 0.02 * Time - 0.01 * Temp + 0.03 * Hour$$

The above shows that Fare amount is the most important variable in determining the tip amount. It is important to note that both Distance and Time are highly correlated and thus the "Times" coefficient is not a true reflection of its actual importance on Tip amount

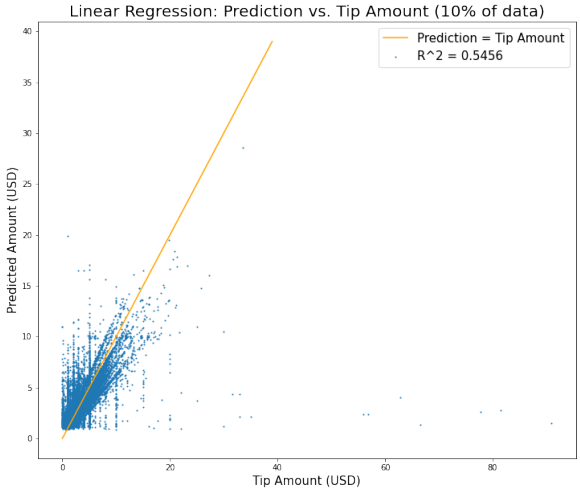

Figure 7: Linear regression

Figure 6 shows the relationship between these predicted values and the ground truth. A line has been added to aid in seeing where the prediction value perfectly matched the data. This result of $R^2 = 0.5456$ (like the neural network) is a lot better than expect due to the not knowing anything about the individual who got in the taxi.

4.4 Passive Aggressive Regression

Passive aggressive regression is a form of unsupervised machine learning, it works by feeding the data through in batches, allowing it to adapt to new data as it is made available. This method, although not meant for machine learning on stand alone data, was chosen as it has the potential to be used to make real time adjustments to its regression, if fed new taxi instances (perfect for taxi drivers on shift).

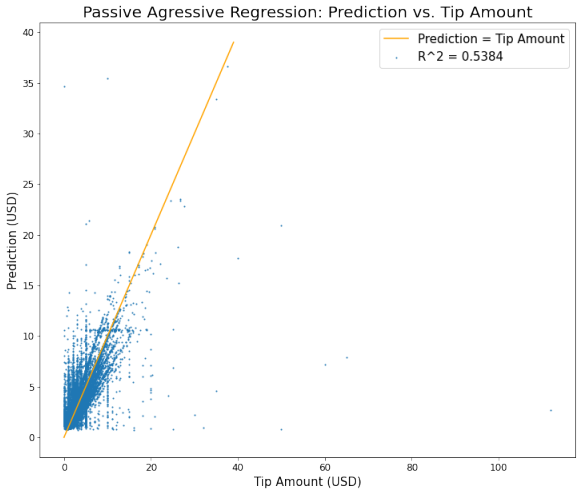

Figure 8: Passive Aggressive Regression

Through a process of iterative search it was determined that the regularisation parameter was optimised to be a low as possible and thus was chosen to be 0.0001. The result of training saw the following coefficients:

$Tip = 1.68 * Fare + 0.06 * Distance + 0.02 * Time - 0.01 * Temp + 0.03 * Hour$

These results are quite similar to the coefficients found in the linear regression, just with greater weighting on the " Fare " variable. Figure 7 shows that like the previous models, the passive aggressive regression model works quite well with $R^2 = 0.5384$, thus concluding this model to be relatively accurate in predicting tip amounts.

Overall the 3 models had a fairly similar accuracy of an approximate r-square value around 0.54. This shows how well just 5 variables can predict how much someone will tip. The best model was the linear regression, concluding that fare amount was the most important factor that goes into someones tipping decision (out of the ones tested). This result is not surprising, as laid out in the " Fare Amount " section, it is what is commonly seen to be the norm in popular culture (tip the a percentage of the bill).

The passive aggressive model was the worse performing model, although not by much. It is believed to not have performed as well due to it being built to constantly be updated as new data is made available.

If a model was to be used by taxi drivers on the job, it would be the passive aggressive regression model that would work best, for the reasons outlined above. .

5 Recommendations

This report was aimed at taxi drivers who wish to increase the amount of tips they receive (and thus overall take home pay). Based on the above analysis and modelling, it is strongly recommended that drivers:

- Do majority of their trips in and around airport

From the geospatial visualisation, it was seen that the extreme cases of high tip averages (up to $16) came from the three main airports in NYC as compared to the rest of the city for which each zone commanded less than $3 per trip.

- Drive between the hours of 10pm and 4am

From the "Time of Day" analysis section, it could be seen that at night people were far more likely to tip. It was shown that if you remove those that don't tip, the average tip amount throughout the day is relatively uniform (with the exception of the hours around 5am) and thus looking at the times where people are more likely to tip will boost overall tips given in a shift. This correlation is also supported in the modelling section with "Hour of the day" being a variable determined to be one of the most important (of the ones tested) in determining Tip Amount.

6 Conclusion

Overall, it was determined that "Fare Amount", "Trip Distance", "Time Taken", "Hour of the Day" and "Temperature" were the most important variables tested in predicting the "Tip Amount". Airports, as well as Flushing Meadows generated the most tips per trip and night time driving appeared to lead to people tipping more often. Data on the individual taking the trip is highly likely to boost the accuracy of the models and thus give a greater picture as to what goes into someones likelihood of tipping.

7 References

1. TLC Trip Record Data - TLC. (2022). Retrieved 21 August 2022, from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

2. National Centers for Environmental Information (NCEI). (2022). Retrieved 21 August 2022, from https://www.ncei.noaa.gov/access/search/data-search/global-hourly?bbox=40.798,-73.897,40.754,-73.853startDate=2018-01-01T00:00:00endDate=2020-01-01T23:59:59

3. Walker, W.H., Walton, J.C., DeVries, A.C. et al. Circadian rhythm disruption and mental health. Transl Psychiatry 10, 28 (2020). https://doi.org/10.1038/s41398-020-0694-0