Taxi Tipping Behaviour Analysis in NYC

Chaitanya Chandrika Raghuvanshi
Student ID: 1117645
Github repo with commit

August 29, 2022

1 Introduction

This report focused on the tipping behaviour of New York City taxi passengers with the aim of finding ways to increase the profitability of taxi drivers' income. Furthermore, looking at the tipping behaviour would give drivers an insight to customer satisfaction and how they can improve their services. NYC Yellow taxi was considered for the model as they are the only type of cabs that can be street hailed by passengers from any part of the city. The report also considered how NYC weather effects the tipping behaviour of its passengers. Furthermore, the median household income for each borough was also considered to see if some boroughs would be more profitable than others. Time's role in the tipping behaviour of passenger as we analysed the pickup hour and the day of the week.

1.1 Dataset

The monthly dataset was taken from TLC data [1]. Year 2018 and 2019 were selected as anything after was impacted by COVID-19. However, now that the world is going back to pre-pandemic stage and the restrictions are mostly lifted, 2018-19 would help our model not be impacted by COVID-19 exceptions. Furthermore, only 1-6 months were chosen for both the years as the model focuses on weather instead of climate. Therefore, we wish to keep the climate the same so that it doesn't affect our predictions.
The daily summaries of New York City weather was taken from NOAA [2]. Various extreme weather conditions such as hail, snow, storm was combined to form a bad weather attribute ranging from 1 to 6 along with amount of snowfall and snow depth. The reasoning behind taking these features was that passengers may appreciate the efforts taken by taxi drivers to drive in such conditions, thus tipping more. Another external dataset taken into account was regarding the median household income for people living in New York from United States Census Bureau[3]. Each borough was given an income order rank which would act as an ordinal feature of the dataset. The hypothesis under consideration was that customers being picked up from boroughs with higher income would give higher tip amounts.

1.2 Overview of Methodology

We start of by briefly discussing the datasets involved and the reasoning behind their range selections. Secondly, we mention the features needed for our modelling along with their pre-processing and feature engineering. Thirdly, the report showcases Random Forest and Neural Network Machine Learning Models and a predictive analysis of the same. Lastly, we give recommendations to target audience, that is, taxi drivers and provide a brief discussion on the outcomes.

2 Preprocessing, Analysis and Feature Engineering

2.1 Tip Amount

- For considering tip amounts, we needed to only consider payment type 1, that is, the one with credit card.
- The tip amount is relative to the total amount paid by the customer; therefore, a more accurate dependant variable would be percentage of tip amount relative to total amount.
- It was found that the max value was much higher than the mean value. Therefore, outlier anlaysis was performed. Only 107,283 rows out of 37,568,814 rows contained tip amount percent greater than 35 percent, therefore we dropped them as they would negatively impact model performance.
- Figure 1 shows that the majority of the tip amounts are between 15 and 20 percent. To make the problem and response variable easier to understand, we create three bins to classify the tip amounts as low(0), average(1) and high(2).

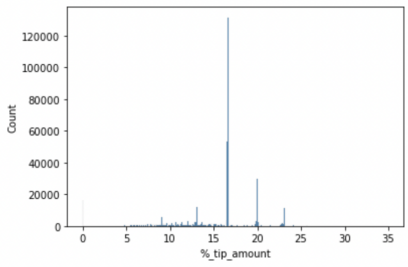

Figure 1: Relative Tip Amount Distribution

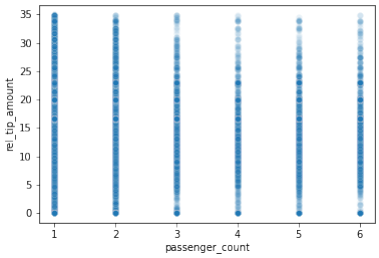

Figure 2: Distribution of relative tip amount based on passenger count

2.2 Passenger Count

- O and 192 passenger counts were unexpectedly found in the dataset. For 0, either the driver entered incorrect details, the passenger did not sit in the car for the booked ride or there were good being transferred. 192 is practically not possible, therefore we remove both of these counts.
- The maximum number of taxi passengers allowed is 5 with the exception of an extra passenger under the age of 7. Therefore, in total 6 passengers can sit in a car [4]. Therefore, we also remove passenger counts of 8 and 9.
- In Figure 3, since the line is getting dotted and blurred more on increasing passenger count, we can conclude that rides with lesser passenger counts leads to increased tipping amounts. This maybe due to the fact that trips with higher passenger counts would be shared rides which is usually taken by people trying to save money rather than spending more on tips.

2.3 Trip Distance

- For trips shorter than 0.05 miles, the duration of trip is mostly in seconds and the difference in pick up and drop off location is negligible. Therefore, it is most likely the trip was canceled or the driver refused the trip. We won't consider this scenario in our model for predicting tip amounts to keep it simple. Furthermore, getting tips even when the ride was not completed can be a confusing outcome for drivers.

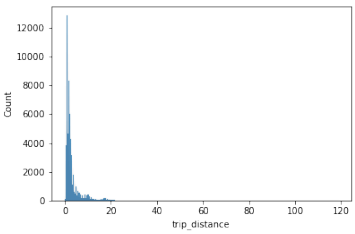

Figure 3: Trip distance distribution

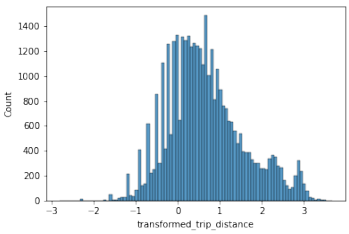

Figure 4: Transformed trip distance distribution

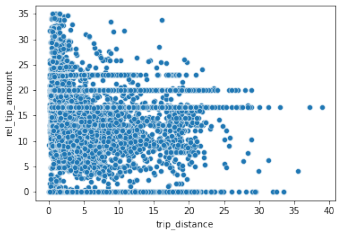

Figure 5: Tip Amount for given Trip distances

- Looking at Figure 4 above, we can clearly see most of the trip distances are less than 20 miles. The length of New York City from northeast to southwest is 35 miles[5], and since we are looking for rides within New York City anything above 40 miles can be considered an outlier. Out of 36,800,139 only 4021 rows have a trip distance of over 40 miles. Therefore, we just remove them from our dataset.
- The data is also highly skewed to the right, therefore we perform a log transformation as shown in Figure 5. A more normal distribution would make predictions for our model easier.
- We can see from Figure 6 above that most of the above average tip amounts are for shorter trip distances, while long distance rides mostly have below average tip amounts. Therefore, we hypothesise that an increased trip distance results in decreased tip amount.

2.4 Pickup Zone

- Figure 6 shows us the average tip amount for each pickup zone based on LocationID. We can see from the figure tat majority of the above average tips are in zones of Manhattan and surrounding areas. The tip amounts are very high around the islands near JFK airport. This could mean that tourists staying in the nearby islands of the airport are giving higher tips.

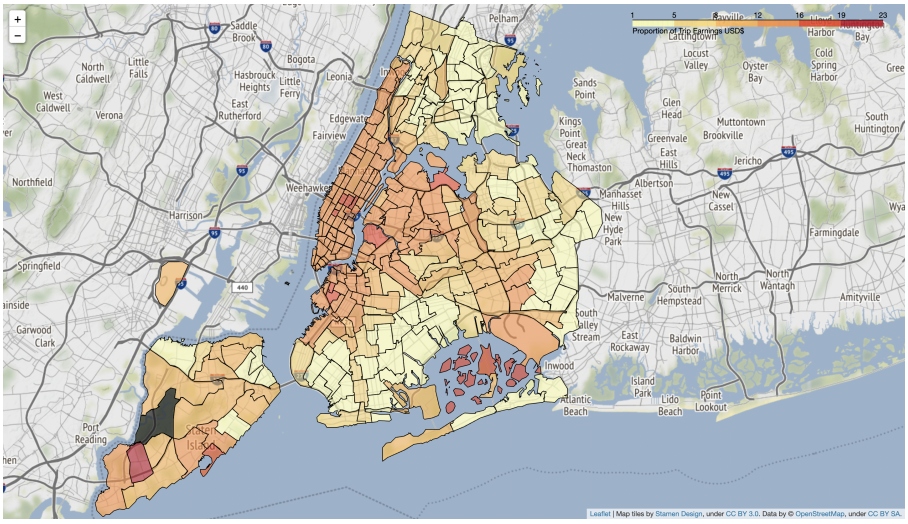

Figure 6: Average Tip Amounts for Major Pickup Zones

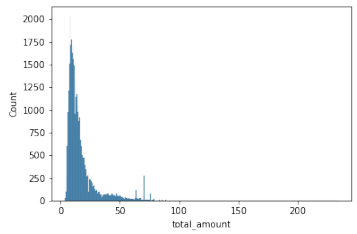

Figure 7: Total amount distribution

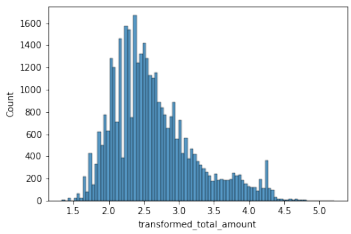

Figure 8: Transformed total amount distribution

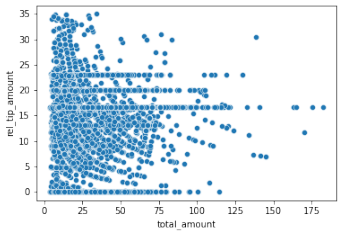

Figure 9: Tip Amount for given Total Amount

2.5 Total amount

- We consider total amount as it comprises of fare amount, toll amounts and other surcharges. The minimum charge for a yellow taxi is 2.5 dollars so we filter out the fares smaller than this [5]. Seeing Figure 7, a fare of above 200 seems like an outlier as it would be to much for just a city ride. So we filter them out as outliers.
- We use a log transformation to scale the data (as data is sparse) and adjust the skewness as shown in figure 8.
- From Figure 9, we can clearly see there is a decrease in tip amount as the total amount increases. However, the trend is not linear so we need to make sure our model can take non linear relationships as well. This trend can be due to the fact that the customer might be hesitant to spend more after paying high amounts for other charges.

2.6 Pickup Hour

- We can see in Figure 10, above average tip amount is higher for evening rides. These taxis were picked up around the ideal time when people leave office or go out for dinner.

2.7 Borough Median Household Income

- As shown in Figure 11, majority of the tips were given by passengers from Manhattan, the borough with highest median household income. The tip amount gets lower and lesser as we move to boroughs with lower median household incomes.

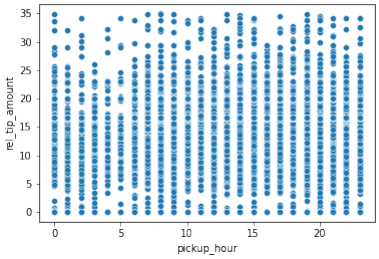

Figure 10: Tip Amounts based on Pickup Hour

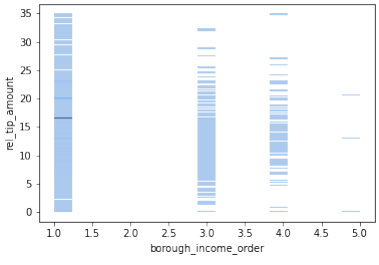

Figure 11: Distribution of relative tip amount based on median household income of boroughs

2.8 Pickup Day of Week

- We cannot see much significant difference based on day of week as seen in Figure 12. However, there is a slight difference between weekdays and weekends, with weekdays being higher tip amounts. Other factors such as traffic, mood of passenger for a particular day contribute to tip amount as well and our model fails to take these into consideration. Further research can be done by adding traffic data for each day to see how it affects the tip amount.

2.9 Snowfall

- Its important to note that our snow features contains very sparse data, with most of them having value of 0, therefore its harder to analyse the data visually. However, we can see people tipping around 35 percent when snow depth is 8. But, when snowfall amount is higher, the tip amount is still low or average. Again this behaviour can't be explained by our model as we need to take other conditions such as traffic and passenger mood into account.

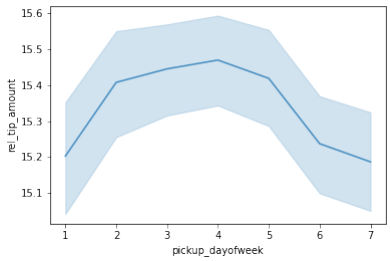

Figure 12: Relative Tip Amount Distribution based on Day of Week

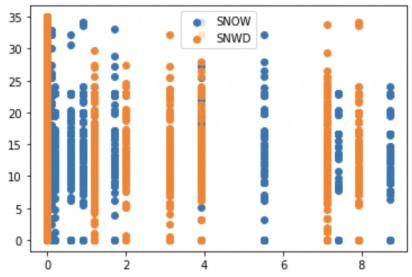

Figure 13: Distribution of relative tip amount based on snowfall

2.10 Weather Conditions

- We consider fog, freezing fog, thunder, pellets, hail, glaze,smoke and drifting snow as the weather conditions that can affect the driver's tip.
- The Heatmap in Figure 14 shows how correlated these conditions are. We can see that pellets, glaze, fog and freezing fog have high correlations so they might not be as significant or different than the rest of the features.

2.11 Correlations among Ordinal and Numerical Features

- Figure 15 shows us correlations among ordinal and numeric features of the dataset. We can see that the trip distance and trip amount are highly correlated which intuitively makes sense. Therefore one is likely to be much more significant than the other.

3 ANOVA Parameter Relevance Model

In order to check if our features are significant for our model, we performed an ANOVA test as seen in Figure 16. The results strongly support our initial analysis as:

- Pickup Hour is much more significant than pickup day of week when it comes to analysing tip amounts.

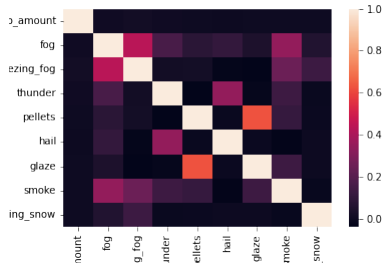

Figure 14: Correlation Heatmap for weather conditions

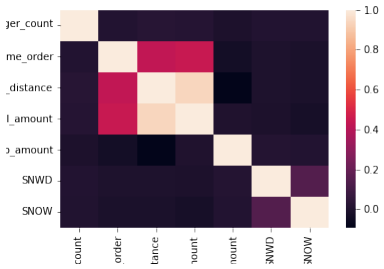

Figure 15: Correlation Heatmap for features

- Transformed total amount has a much smaller p value than transformed trip distance which again means as Transformed total amount is highly significant, transformed trip distance is not as much due to the correlation between the two.
- Glaze is the least significant as we saw it highly correlates to pellets.

	sum_sq	df	F	PR(>F)
PULocationID	8.247425e-22	235.0	4.417537e+03	0.000000e+00
pickup_hour	6.850595e-25	23.0	3.749127e+01	3.809417e-167
pickup_dayofweek	6.222471e-26	6.0	1.305393e+01	7.959697e-15
passenger_count	7.561839e-26	1.0	9.518251e+01	1.748145e-22
borough_income_order	1.931723e-24	1.0	2.431501e+03	0.000000e+00
tip_class	2.551554e+05	1.0	3.211696e+32	0.000000e+00
transformed_trip_distance	4.068071e-26	1.0	5.120569e+01	8.334578e-13
transformed_total_amount	4.549565e-25	1.0	5.726637e+02	1.887859e-126
SNOW	1.197683e-26	1.0	1.507550e+01	1.033152e-04
SNWD	3.205948e-25	1.0	4.035396e+02	1.054804e-89
TAVG	1.033779e-26	1.0	1.301240e+01	3.094799e-04
TMAX	1.996007e-25	1.0	2.512417e+02	1.459668e-56
TMIN	1.374292e-26	1.0	1.729851e+01	3.195155e-05
fog	2.628375e-26	1.0	3.308393e+01	8.834129e-09
freezing_fog	2.800007e-26	1.0	3.524430e+01	2.911154e-09
thunder	7.940756e-27	1.0	9.995201e+00	1.569626e-03
pellets	1.443220e-26	1.0	1.816612e+01	2.025032e-05
hail	6.124097e-26	1.0	7.708534e+01	1.644709e-18
glaze	2.253000e-27	1.0	2.835900e+00	9.218006e-02
smoke	2.600167e-25	1.0	3.272887e+02	4.055778e-73
drifting_snow	1.704101e-26	1.0	2.144989e+01	3.633320e-06
Residual	2.673148e-22	336475.0	NaN	NaN

Figure 16: Anova model for parameter relevance

4 Multi Layer Neural Network Model

Since we had a non linear classification problem, we chose a multilayer neural network classifier to model our predictions. 279 features were taken a sparse vector with 3 output layers. The accuracy received was 0.498, meaning the fit was average at best. Furthermore, the computation time was around 2 hours, therefore it is also expensive to carry out. This was due to the large training set of over 35 million records. The model failed to achieve required prediction goals. However, this report provided a thorough preliminary analysis of features, leaving researchers for various opportunities to expand the scope.

5 Recommendations and Discussion

The report aimed to find ways to increase taxi driver's profitability by analysing the tipping behaviour of NYC passengers. This was achieved by a thorough visual analysis of dataset features, their significance and prediction regarding future tipping behaviour. The recommendations for drivers include:

- Looking out for rides with lesser number of passengers as they are likely to receive higher tip amounts.
- Driving more around Manhattan JFK airport and the nearby islands to target the customers who are likely to pay higher tips.
- Do more rides during the evening when customers are likely to pay higher tips
- Focus more on smaller trip distances as people travelling longer distances are not as likely to give high tips. Furthermore, if the total amount is reasonably large passengers will hesitate to spend more on tips. This report can also be benefited by data science researchers to further the analysis by analysing traffic events.

6 References

NYC Taxi Limousine Commission. NYC Taxi Data. https://www1.nyc.gov/site/tlc/ about/tlc-trip-record-data.pag
Accessed: 2022-08-09.

@misc2022sensorlocation, title = Climate Data Online Search, author = NOAA, howpublished = https://www.ncei.noaa.gov/cdo-web/search, note = Accessed: 2022-08-02

@misc2022sensorlocation, title = American Community Survey 5-Year Data (2009-2020), author = United States Census Bureau, howpublished = https://www.census.gov/data/developers/data-sets/acs-5year.html, note = Accessed: 2022-08-04

@misc2022sensorreading, title = Passenger Frequently Asked Questions, author = NYC Taxi and Limousine Commission, howpublished = https://data.melbourne.vic.gov.au/Environment/Microclimate-Sen u4vh-84j8, note = Accessed: 2022-08-01

@misc2022sensorlocation, title = NYC By the Numbers, author = Walks of New York, howpublished = https://www.walksofnewyork.com/blog/nyc-by-the-numbers#:~:text=The%20total%20area%20of%20the,southwest%2C%20is%20about%2035%20miles