NYC Taxi Tipping Behavior

ABSTRACT

New York City (NYC) is the most populous city in the United States. As such, NYC features one of the most robust and heavily used transportation systems in the country. As part of the NYC OpenData program, the NYC Taxi & Limousine Commission (TLC) releases monthly trip records for the city's yellow and green taxis dating back to 2009 that contain millions of trips in any given month. The dataset provides a wealth of information on the use of taxis throughout the city including key attributes such as drop off/pickup times, fares, and trip distances. In analyzing the domain, our team is looking to identify key trends and features from the dataset that can help accurately predict tips received on any given trip through the use of machine learning. Due to the sheer volume of taxi trips conducted in a given month, this task is best approached through machine learning to identify trends and patterns within the large dataset that are otherwise likely to remain unnoticed through a cursory analysis. This study can help the taxi industry identify where to focus their resources and allow the taxi drivers earn a higher tip. As an alternative, this study could be of use to For Hire Vehicle (FHV) drivers, such as Uber and Lyft, in identifying the best locations and times to maximize profits while on duty.

1. Insights

During the course of our analysis, we are seeking to identify any key features that are best associated with predicting the tip a taxi driver will receive. Our primary focus will be on features such as time of day, location, trip duration, passenger count, and total fare. In analyzing these features, we hope to identify several key insights in regards to available taxi trip data including:

1. What day of week and/or time of day is likely to generate the highest tip?
2. Are certain pickup or drop-off locations more apt to produce consistently higher tips?
3. What impact does the passenger count, if any, have on the tip?
4. Does weather have any effect on the amount a driver is tipped?

2. Dataset

For our study, we will be using the New York Taxi & Limousine Commission dataset. The dataset is publically available, partitioned every month from 2009 - 2016, and contains data from yellow and green taxi cabs. We will analyze the data from both yellow and green datasets from July 2015 - June 2016. This timeframe will provide a large enough dataset to identify key insights and predictions over several months while remaining within the capabilities of our available computing resources. In addition, we wanted to add the National Oceanic and Atmospheric Administration (NOAA) weather data from July 2015 - June 2016 as another feature to analyze in our study. In the case of using NYC taxi data to predict the tip percentage, using a data-driven approach is the best way to analyze the problem. In a single year, the NYC taxis logged over 140 million records and associated data points. Processing this much data lends itself to a data-driven approach where machine learning can be used to process, extract, and analyze the data in a fashion that allows for actionable results in a timely fashion.

3. Characteristics

The NYC taxi dataset from July 2015 - June 2016 will have a total of 159,733,022 data points. The data from this period totals approximately 24.84 GB in size. The months for the yellow taxi dataset ranges from 10.9 million to 12.3 million and total out to 140,738,355 data points, whereas the months for the green taxi dataset range from 1.4 million to 1.6 million and total out to 18,994,667 data points. The dataset contains structured data and can easily confine to a schema. The NOAA weather data will have 365 data points and is 11 KB in size.

4. Features

Table 1. Yellow Taxi Dataset

Field	Type	Description
VendorID	Categorical /Numerical Code	Code to determine which Taxi company
tpep_pickup_datetime	Date	Date and time when taxi fare starts
tpep_dropoff_datetime	Date	Date and time when taxi fare ends
passenger_count	Numerical	Number of passengers
trip_distance	Numerical	Trip distance in miles
pickup_longitude	Numerical	Longitude of pickup
pickup_latitude	Numerical	Latitude of pickup
RatecodeID	Categorical /Numerical Code	Code to determine the rate type
store_and_fwd_flag	Boolean	Flag that determines if data was stored in memory
dropoff_longitude	Numerical	Longitude of drop-off
dropoff_latitude	Numerical	Latitude of drop-off
payment_type	Categorical /Numerical Code	Code to determine how passenger paid
fare_amount	Numerical	Fare calculated by time and distance
extra	Numerical	Miscellaneous extras and surcharges
mta_tax	Numerical	Tax fee based on rate
tip_amount	Numerical	Tip amount from credit cards, cash not included
tolls_amount	Numerical	Total amount of all tolls paid in trip
improvement_surcharge	Numerical	Surcharge for assessed trips at the flag drop
total_amount	Numerical	Total amount, does not include cash tips

The Yellow Taxi dataset contains nineteen fields.

Table 2. Green Taxi Dataset renamed/extra field

Field	Type	Description
lpep_pickup_datetime	Date	Date and time when taxi fare starts
lpep_dropoff_datetime	Date	Date and time when taxi fare ends
trip_type	Categorical /Numerical Code	Code to determine if trip was street hailed or dispatched

The Green Taxi dataset is similar, but has two renamed date fields and one new field, which totals to twenty.

Table 3. Weather Dataset

Field	Type	Description
Date	Numerical	Identifies the date of weather data
PRCP	Numerical	Precipitation in inches
SNWD	Numerical	Snow Depth in inches
SNOW	Numerical	Snowfall in inches
TMAX	Numerical	Maximum Temperature in Fahrenheit
TMIN	Numerical	Minimum Temperature in Fahrenheit

The weather dataset has twenty-five fields, but we will only be focusing on the temperature and stormy weather conditions.

5. Target Prediction

The target of our machine learning algorithm is the tip on any given taxi ride. Due to the nature of tips, we do not believe it is feasible to predict the exact amount a driver will receive. Instead, we will focus our target predictions on a range of tip percentages (e.g. 0-5%, 6-10%, etc.). We are seeking to identify the best grouping of target ranges during our data analysis phase. However, we hope to remain within windows of five or ten percent. As for the results of our target prediction, we are optimistic that an average accuracy of 70% or higher is achievable when predicting the more common tip ranges of 15-30%. Other target ranges may prove to be wildly variable and thus not a reliable way to generate a steady income.

6. Algorithm

The problem is best cast as a supervised learning problem because we are attempting to classify what a rider will tip, a known value, based upon known data points. Further, since we are attempting to categorize the tip into different percentages the problem lends itself towards a classification model. To conduct our research, we will be focusing on using classification techniques to gather insights on customer tipping behavior. Decision trees, Naive Bayes classifier, Nearest Neighbor, and Linear classifier will be the primary focus to answer our machine learning questions.

7. Expectations

We expect to see a trend based on the customer tipping behavior based on conditions of the location, a number of passengers, the day of the week, and weather conditions.

8. Measure Improvements

Based upon the current goals of our analysis, measuring improvements can be determined through analysis of cash flow before and after the algorithm is implemented. An effective algorithm would allow for a taxi company or driver to focus resources where they are most likely to generate the highest revenue stream. If successful, the new focus of company resources will generate revenues greater than revenue prior to implementation on a consistent basis. The taxi company will want to focus resources based on the time of year, weather conditions, and locations from densely populated areas.

9. Identified/Extracted Features

9.1 Noise Reduction

To narrow down on useful features, we will remove VendorID, extra, mta_tax, tolls_amount, improvement_surcharge, store_and_fwd_flag, trip_type, ehail_fee, and RatecodeID because they will not help us in determining useful insights and only cause additional noise.

9.2 Trim Dataset

To clean the data, we will only keep the data when payment_type equals the code value of 1 (credit card) because it is the only class that ensures we are dealing data that has tipping data associated with it. Other payments types, such as cash, are not required to list any tip amount. We will remove any data where the total_amount and trip_distance is less than 0 to avoid any anomalies. To narrow the data to focus only in New York City, we will only use any data in the pickup and drop off coordinates between the Latitude (40.459518, 41.175342) and Longitude (-74.361107, -71.903083).

9.3 Date Features

To categorize the date features, we performed some feature engineering to parse the date fields (tpep_pickup_datetime, tpep_dropoff_datetime, lpep_pickup_datetime, lpep_dropoff_datetime) into three separate fields (trip_time, day_of_week, pickup_hour). The trip_time will filter out trips that are less than two minutes and more than three hours, day_of_week will categorize the days into codes, and pickup_hour will identify the hour of the customer pick. In addition, we added the month, year, and season to further categorize the date features that will allow us to identify any hidden insights during a particular part of the year.

9.4 Tip Percentage

To determine the customer tipping behavior, we created a field that would categorize the tipping percentage using the tip_amount and total_amount.

The calculation is:
percent = (tip_amount / (total_amount - tip_amount)) * 100

The percent will be categorized into 0%-5%, 5%-10%, 10%-15%, 15%-20%, 20%-25%, 25%-30%, 30%-35%, 35%-100%, and greater than 100%.

The tip_percent field will be used as the target feature in our training and test our algorithms.

9.5 Location Grouping

To further generalize the location data, the accuracy of the pickup and drop-off coordinates is rounded off to the third decimal place. Rounding to the third decimal place provides a rough accuracy of 110 meters between different coordinates. We felt this accuracy to be sufficient enough in grouping locations to roughly a block by block accuracy. In further feature engineering of the location values, we turned NYC into a grid matrix and assigned each lat/long coordinate within the grid a value. This feature engineering reduced locations values into a pickup_area and dropoff_area.

9.6 Merge Weather Data

To merge the data from the NOAA weather dataset, we added a temporary date field to the taxi dataset that matched the format of the date in the weather data. This date field acted as the key between the taxi and weather data allowing us to perform a join and append the weather features (PRCP, SNWD, TMAX, TMIN) for that day. Once the join was completed, the date field was removed.

9.7 Final List of Features

Table 4. Final Dataset

Field	Type	Description
trip_time	Categorical/ Numerical Code	Code to determine length of trip
day_of_week	Categorical/ Numerical Code	Code to determine the day of the week
month	Categorical/ Numerical Code	Code to determine the month
year	Categorical/ Numerical Code	Code to determine the year
season	Categorical/ Numerical Code	Code to determine the season
pickup_hour	Categorical/ Numerical Code	Code to determine the hour of pickup time
passenger_count	Numerical	Number of passengers
fare_amount	Numerical	Fare calculated by time and distance
tip_percentage	Categorical/ Numerical Code	Code to determine the tip percentage
trip_distance	Numerical	Trip distance in miles
pickup_area	Categorical/ Numerical	Pickup location on X/Y matrix
dropoff_area	Categorical/ Numerical	Drop-off location on X/Y matrix
PRCP	Boolean	Determines if there was rain on that day
SNWD	Boolean	Determines if there was snow on the road that day
SNOW	Boolean	Determines if there was snow that day
TMAX	Numerical	Maximum Temperature in Fahrenheit
TMIN	Numerical	Minimum Temperature in Fahrenheit

10. Algorithms Testing

After performing our data reduction and feature engineering, we were able to reduce 22 GB of taxi data down to 4.67 GB. Our initial target dataset reduced the tip percentage into nine categories of 0%-5%, 5%-10%, 10%-15%, 15%-20%, 20%-25%, 25%-30%, 30%-35%, 35%-100%, and greater than 100%. Initial classification tests with the target split into nine classifications returned an average accuracy of roughly 45% on even the best performing classification models. We continued to massage the dataset by classifying features and removing features, but we were still underperforming in our predictions. In an effort to improve the accuracy or our model, we eventually reduced the tip percentage into two categories of either under or over 20%. By reducing our tip percentage into two categories we drastically improved the performance of our classification models. In testing a subset of two million records, both the Random Forest and Decision Tree classifiers achieved 5-fold cross-validated accuracies around 75%. Additionally, in the classifier category, Naive Bayes and K-Nearest Neighbor classifiers underwent tests. However, the models did not show results as promising as the Random Forest and Decision Tree. Lastly, we attempted models for Linear Regression and K-Nearest Neighbor Regression. Unfortunately, the results for both regression models were not conducive enough to warrant further evaluation. Unsupervised learning models were skipped because we wanted to focus on experimenting with classification models.

11. Details of Analysis

We tested and tuned our algorithms extensively and we were able to get the Decision Tree and Random Forest as our top two performing algorithms. In our first round of feature engineering, we were struggling to get passed 50% accuracy for any of our algorithms. As we continued to classify our features we saw improvements in our accuracy.

Our algorithms needed to tuned to filter out all of the noise the data generated. For the Decision Tree, we saw a significant improvement in the metrics by reducing the max_leaf_nodes down to 1,000. The Random Forest metrics was able to compete by setting the n_estimators to 50. K-Nearest Neighbor Classifier and Naive Bayes Classifier were neck and neck once the K-Nearest Neighbor set 20 as K. Our regression algorithms performed so low as we moved away from continuous values in feature engineering and the other algorithms kept on improving.

The metrics for the classification models were valued by the accuracy of training + testing set, classification Report, 5-Fold Cross Validation, and Mean Score. The regression models were valued by the Variance Score, Residual Sum of Squares, Mean Absolute Error, Mean Squared Error, Root Mean Squared Error and 5-Fold Cross Validation.

11.1 Decision Tree Classifier

Accuracy on Training Set: 76.520%
Accuracy on Testing Set: 76.339%

Table 5. Classification Report

	Precision	Recall	F1-score	Support
<=20%	0.77	0.97	0.86	375005
>20%	0.60	0.16	0.25	124995
Avg/Total	0.73	0.76	0.71	500000

5-Fold Cross Validation
Scores: [0.7642725, 0.76297, 0.763835, 0.7631, 0.763465]
Mean Score: 0.764 (+/-0.000)

11.2 Radom Forest Classifier

Accuracy on Training Set: 99.966%
Accuracy on Testing Set: 75.523%

Table 6. Classification Report

	Precision	Recall	F1-score	Support
<=20%	0.77	0.97	0.86	375005
>20%	0.55	0.12	0.20	124995
Avg/Total	0.71	0.76	0.69	500000

5-Fold Cross Validation Scores: [0.755385, 0.754005, 0.75461, 0.75521, 0.754105]

Mean Score: 0.755 (+/-0.000)

11.3 K Nearest Neighbor Classifier

Accuracy on Training Set: 75.6902%

Accuracy on Testing Set: 74.7704%

Table 7. Classification Report

	Precision	Recall	F1-score	Support
0	0.76	0.98	0.86	375005
1	0.47	0.06	0.11	124995
Avg/Total	0.48	0.75	0.67	500000

5-Fold Cross Validation Scores: [0.7479925, 0.7466725, 0.7470375, 0.7469425, 0.746235]

Mean Score: 0.747 (+/-0.000)

11.4 Naïve Bayes Classifier

Accuracy on Training Set: 74.7794%

Accuracy on Testing Set: 74.8254%

Table 8. Classification Report

	Precision	Recall	F1-score	Support
0	0.75	1.00	0.86	375005
1	0.26	0.00	0.01	124995
Avg/Total	0.63	0.75	0.64	500000

5-Fold Cross Validation Scores: [0.7485275, 0.747995, 0.7478725, 0.7486525, 0.7471675]

Mean Score: 0.748 (+/-0.000)

11.5 Linear Regression

Variance Score: 0.01

Residual sum of square: 0.19

Mean Absolute Error: 0.37

Mean Squared Error: 0.19

Root Mean Squared Error: 0.43

5-Fold Cross Validation Scores: [0.0051262, 0.00521547, 0.0052033, 0.00493357, 0.00525286]

Mean Score: 0.005 (+/-0.000)

11.6 K-Nearest Neighbor Regression

Variance Score: 0.00

Residual sum of square: 0.19

Mean Absolute Error: 0.36

Mean Squared Error: 0.19

Root Mean Squared Error: 0.43

5-Fold Cross Validation Scores: [0.00340432, 0.0023052, 0.00173741, 0.00164286, 0.00162464]

Mean Score: 0.002 (+/-0.000)

12. Insights Gained

Due to the limited memory capacity of our environments, the insights in this section are based upon analysis of 10 million records in our cleaned datasets.

12.1 What day of week and/or time of day is likely to generate the highest tip?

Interestingly enough, we found that days during the workweek tended to have a higher likelihood of an individual tipping greater than 20%. Some potential explanations for the decrease in the likelihood over the weekend could be due to the increase in tourism and party goers who may not be as generous. Whereas, the core work days are more likely to be business individuals.

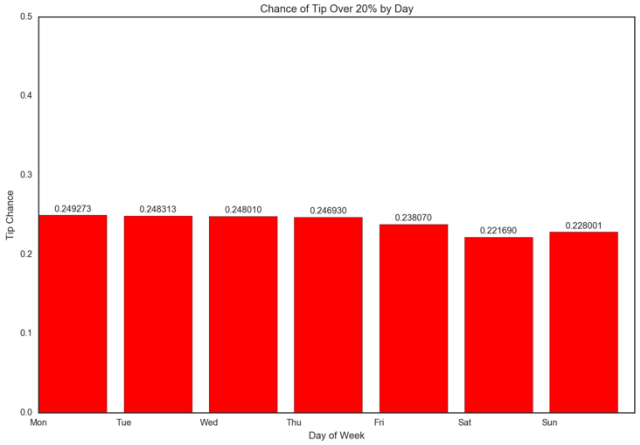

Figure 1. Day of Week

On a per hour basis, we saw some interesting trends in the likelihood that a driver would receive a tip over 20 percent. Early morning hours tended to show a significant decrease in likelihood, while afternoon and evening hours regularly show a higher percentage of likelihood of a greater tip. Please note the range of y-axis below is zoomed to provide a better view of the variation.

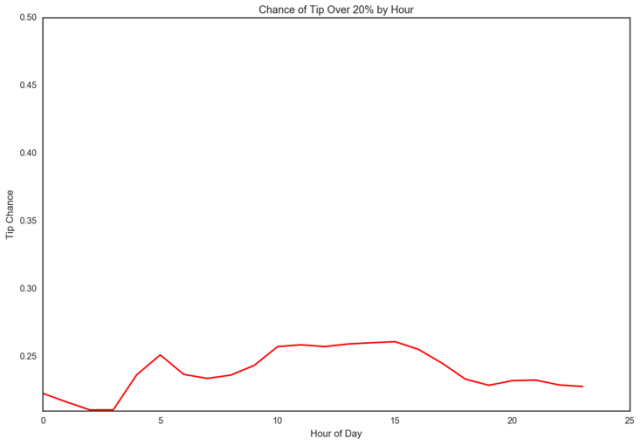

Figure 2. Hour of Day

12.2 Are certain pickup or drop-off locations more apt to produce consistently higher tips?

To answer this question, we created a visualization showing a heat map of the city based on the mean tip chance at a given location. In an effort to not skew the visualization, a minimum count of five tips was required for a location to be populated on the map. The

image on the left is representative of pickups, and the graph on the right is representative of drop-offs.

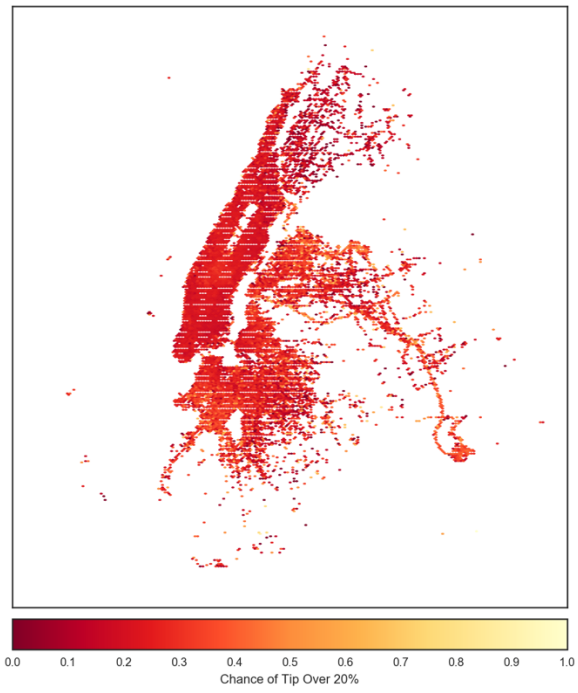

Figure 3. Pickups

Figure 4. Drop-offs

Interestingly enough, we see that riders from the Queens and Brooklyn boroughs tend to tip greater than the riders from the Manhattan and Bronx boroughs. In future studies, it would be interesting to have more data about the passengers such as age demographic, the average cost of housing in locations, and so forth to draw some additional conclusions about the riders.

12.3 What impact does the passenger count, if any, have on the tip?

The data shows the tip chance remains fairly steady for groups of up to 4 passengers. Once a passenger count of greater than four passengers is achieved, there is a greater chance of a tip greater than 20% A group of seven passengers produces the greatest tip percentage, while eight passengers produce the lowest chance. Further analysis may show that eight passengers value is an outlier. Due to the fact that passenger count is a driver entered the value, some trips show a passenger count of zero. These trips were not removed if the rest of the data points were accurate and in scope.

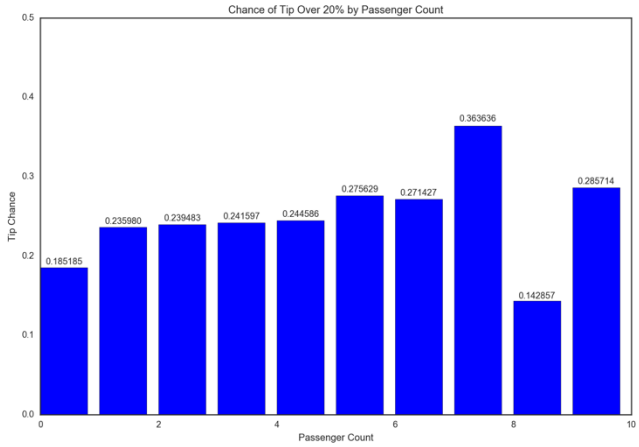

Figure 5. Passenger Count

12.4 Does the weather have any effect on the amount a driver is tipped?

After performing some visual analysis based on rain and snowfall there was a minimal increase in the chance that a driver will be tipped greater than 20%. However, when analyzing the tip chance based on the temperature the data showed that when temperatures were very cold or very hot passengers showed a great chance to tip over 20%. This could possibly be explained by the gratitude to get out of the extreme conditions.

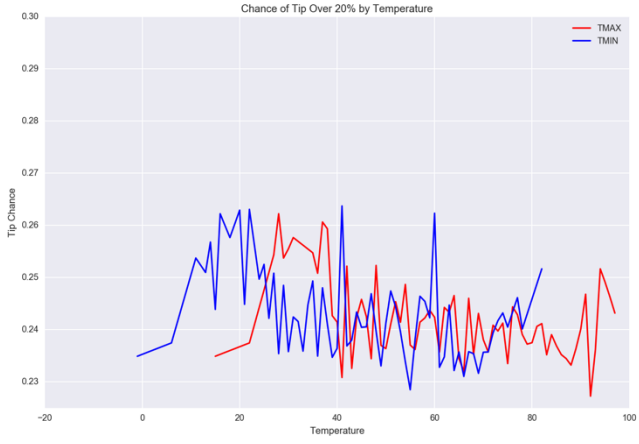

Figure 6. Temperature

13. Summary of Insights

After our analysis, we were able to determine that a few actionable insights. Based on our findings, were able to determine that weekends, winter/summer, and evening hours are likely to generate the highest tip. After we grouped location data into broader location values we find that the outer borough pickup and

drop-off locations were able to result in a higher likelihood of a tip over 20%. Our data was able to gather insights on the impact of the passenger count, and we found that groups of greater than four produced a higher likelihood of a greater tip percentage. Lastly, we studied the effect of the weather data and found that while rain/snow have little effect on the tip chance, the temperature does show a small visible effect on the tip chance during more extreme conditions.

14. Feature Impact on Observed Solution

The features extracted from the data help predict if the customer is likely to tip greater than 20 percent. The most notable features in our analysis showed that location, time of day, and trip distance had the greatest impact on predicting the tip. While taxi drivers have little control over the trip distance of a customer, knowing what locations and timeframes are likely to present the highest chance of increased revenue is a serious advantage.