Predicting Taxi Tip-Rates in NYC

[A predictive model focused on location binning and averages]

Sahil Jain
srj004@ucsd.edu

Alvin See
alvinsee@ucsd.edu

Anish Shandilya
ashandil@ucsd.edu

1. INTRODUCTION

Almost half a million taxi trips are made daily in the city that never sleeps, producing a plethora of information that can prove useful for both the passengers and drivers. We choose to understand what features (besides the quality of the driver) actually factors into the tip received by a cab driver. After examining quite an extensive data set, we have come up with the following observations.

2. EXPLANATORY ANALYSIS

We decided to use a subset of the data provided by the following link:

`https://archive.org/details/nycTaxiTripData2013`

The data is mainly distributed into 2 different files: one having to do with the trip data, and one containing the trip fare.

trip_data:
medallion, hack_license, vendor_id, rate_code, store_and_fwd_flag, pickup_datetime, dropoff_datetime, passenger_count, trip_time_in_secs, trip_distance, pickup_longitude, pickup_latitude, dropoff_longitude, dropoff_latitude

Important Features
pickup_datetime, pickup_longitude, pickup_latitude, trip_time_in_secs, trip_distance

trip_fare:
medallion, hack_license, vendor_id, pickup_datetime, payment_type, fare_amount, surcharge, mta_tax, tip_amount, tolls_amount, total_amount

Important Features
payment_type, tip_amount, total_amount

Note: We used payment_type to filter transaction data into cash-payments and non-cash payments. Cash-payments always had a tip of 0 (most likely because a tip paid in cash could not be recorded), so all such transactions were ignored.

Basic Statistics:

- 200000 data points ($\frac{1}{2}$ for training, $\frac{1}{2}$ for testing)
  - Randomly selected into training/testing
- Data set from 12-01-13 to 12-09-13
- Average Tip Rate = 14.9 %
- Distribution of Tip Rate

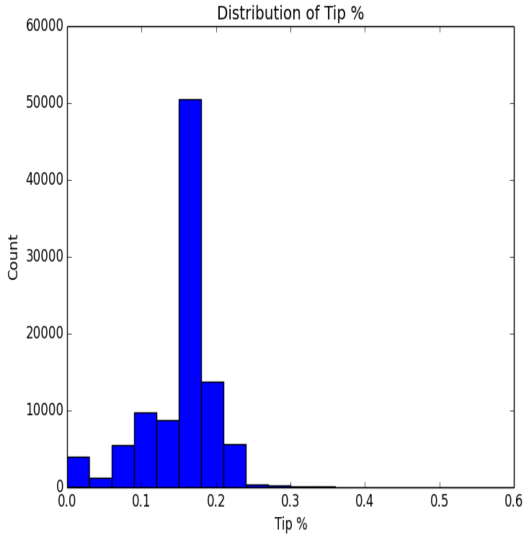

Figure 1: Tip

3. PREDICTIVE TASK

3.1 Data

The predictive task that is being analyzed is the percentage of tip in relation to the total amount paid for taxi trips in New York City (NYC). The predictive task was chosen due to our curiosity of what factors causes people to tip higher percentages. Ultimately, this analysis can be used to assist taxi drivers in considering these factors in order to better understand their business and how they can maximize the tip received.

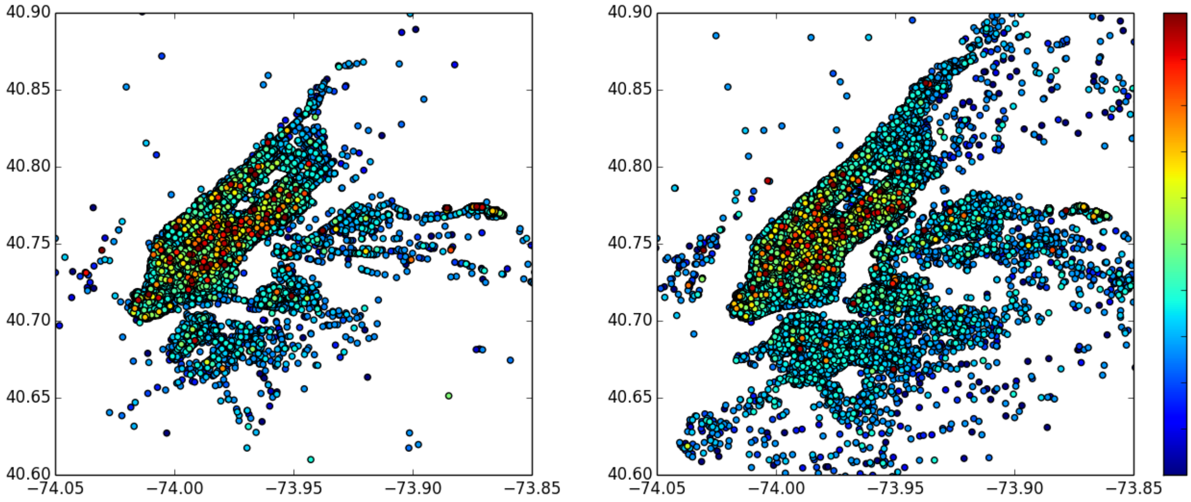

Figure 2: Pick-up and Drop-off Locations

3.2 Evaluating Success

Since we will try many different predictive models, we need some sort of measurement that tells us whether the model is actually considered an optimization. Specifically, we are looking to the Mean Squared Error (MSE) when the model is applied to the training for evaluation. The MSE is then compared with the variance of the set such that if the MSE is less than the variance, the predictive model is considered an optimization.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i - Y_i)^2$$

Specifically, we chose our baseline to be the average tip percentage of the training data after looking at *Figure 1*. Although the baseline is already a good model, we have incorporated models that creatively try to beat it.

3.3 Attempted and Considered Models

3.3.1 Predicting based on taxi driver

The first obvious thing to do was to simply keep a history of taxi driver data, and simply predict based on their previous tips. However we decided that would be too simple of a model. Also this faces a cold-start problem if there is an unseen taxi driver. Therefore we thought it would be more interesting to predict tips based on other potential factors.

Unique Medallions: 6833
Unique Hack Licenses: 13825

3.3.2 Predicting based on location

The next thing we realised is that a diverse city such as New York is bound to have more affluent neighborhoods and boroughs than other parts. Therefore we decided to plot and consider every data point in the training set, and we were presented with the following results:

The left and right scatter-plots (*Figure 2*) show the pick-up and drop-off location coordinates, respectfully, along with an interpolated color of the tip percentage. We can see that the more inner, rich parts of Manhattan (i.e Times Square) have more occurrences of a higher tip percentage. Therefore we thought it would be wise to build a linear regressor that would take into consideration a coordinate euclidean distance from these "hot-spots".

This ultimately failed however to produce a smaller MSE than our baseline, as we did not consider the vast amount of lower tip percentages that lie underneath the higher tips percentages in the scatter plot.

3.3.3 Predicting based on distance traveled, time spent, and speed of journey

We then looked into features that have to do with characteristic of the taxi's trip, and see if they correlated with the tip percentage.

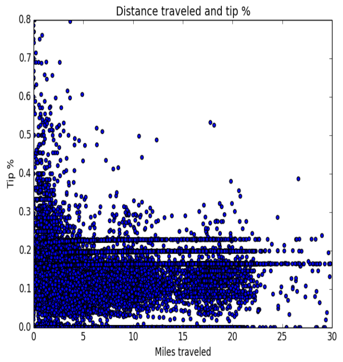

Figure 3: Distance

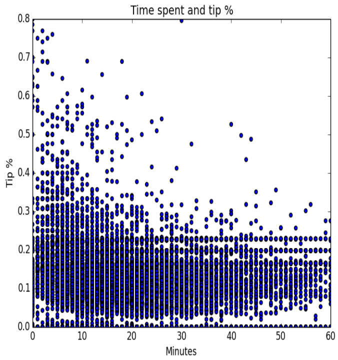

Figure 4: Time

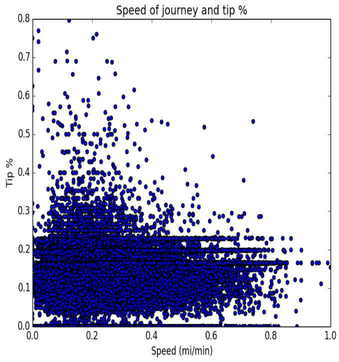

Figure 5: Speed

However, after looking at the graphs more closely, we noticed that the most dense parts of the graph lie close to the average, and that a linear regressor would not be too helpful.

3.3.4 *Predicting based on time of day*

We thought time would be a factor for how a person tips. For example, a person may tip more later in the nights or early in the morning, perhaps due to rush hour.

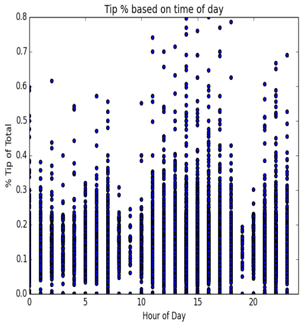

Figure 6: Time of Day

The data shows a slight increase in tip percentage, especially during peak hours. Therefore a linear regression model would make some sense here, especially if it is similar to the one we created in our past homework. In other words, our model would have to have a θ value for every hour of the day.

Surprisingly, this model worked pretty well, and actually reduced our MSE from the baseline, but we thought we could take it even a step further.

3.3.5 *Predicting based on location with bins*

We decided to revisit the location model from earlier, but this time create a heat-map that takes into consideration the points within the area. Our implementation is as follows:

1. Divide the map into NxN bins based on longitude and latitude
2. For every item in the training set:
   - (a) Assign it to its location-appropriate bin
   - (b) Update the average and count of items in the bin
3. For every item in the test set:
   - (a) Find the appropriate bin it would have been placed into
     - i. If the count < 50, predict the global average
     - ii. Else predict the average of the bin.

We found that an N of 20 works nicely, but we could probably have found a more optimal parameter. We also chose to predict the average for bins that do not have enough data, since a few points are not at all representative of that area.

With this model we got an improved MSE of 0.260274 compared to the baseline MSE of 0.261128.

4. RELATED LITERATURE

For our dataset, we decided to use existing data that can be found here:

`https://archive.org/details/nycTaxiTripData2013`

We also found the following study which relates to our predictive analysis on taxi tips in NYC:

`http://www.intetics.com/new-taxi-tip-prediction-map-can-change-the-way-we-take-cabs/`

Like our study, it primarily uses location to help predict the tip percentage. However it's difference lies in the fact that it looks at "the popularity of a particular location by looking at number of trips vs. tip generosity". Still, it confirms location to be a crucial feature in optimizing any model that has to do with tip prediction.

A lot of documentation from 3rd party libraries and forum posts also served as our literature for understanding what data visualization models could and could not work. For example, we naively implemented a scatter plot initially, thinking that that would be the best way to represent our day. We then looked at various other models like 3d bar graphs, surface plots, and 3-D histograms. However we finally chose a 2-D histogram as a means of recording a proper heat map, only after referring to the problems that other people had when trying to plot similar data.

Bloomberg also published an article that reports some statistics on the same data set we ended up using. It was mainly useful in showing a more comprehensive explanatory analysis of all the data, as well as proving that tips, in general, and higher in the evening. It can be found at `http://www.bloomberg.com/bw/articles/2014-07-31/heres-how-much-you-should-be-tipping-your-cab-driver`

Finally, Jose Camacho in his analysis at `https://github.com/josemazo/nyc-taxi-tip-predictor`, was able to successfully predict tips with an accuracy of 71.74%. He used the same data set as we did, but instead ran his model on the larger subset of the data. However instead of considering linear regression models like we did, he used a random forest model.

Because he used a completely different model as we did, he ended up using a lot more features, including those that we may have never considered. Some of these include:

pickup_day, vendor_id, and even passenger_count.

However he did used similar features as us, like location and time.

Camacho, also through his explanatory analysis, found the cash payment type to be unnecessary noise, and filtered it from his data set as well.

5. RESULTS

5.1 Optimizations and Scaling

The main reason we decided to create the heat-map from the ground up ourselves, is because we felt that we had more control over certain features when compared to third party libraries. Unfortunately the trade-off was that our solution is invariably slower. Therefore, it was hard to optimize the parameters such as number of bins divisions, or the sparse data cutoff value. Furthermore, we could not increase the number of bins beyond a certain value without having to wait a while for the data to process. For example a division parameter of 100, resulted in nearly a half-hour waiting time.

λ_1=number of divisions
λ_2=sparse data cutoff value

	λ_2=10	λ_2=50
λ_1=20	0.260367	0.260274
λ_1=30	0.260505	0.260361
λ_1=40	0.260404	0.260363

Table 1: MSE of model with various parameters

Baseline MSE: 0.261128

Based on table (a) the two parameters seem to be inversely correlated, which makes sense. The more you increase λ_1, the smaller the area each bin encompasses. This means λ_2 should be reduced to compensate for the more detailed area since there will be less points generally categorized in that area. λ_1 however cannot be too big, because eventually we will just be predicting the average, since there will be an increase in bins that do not have any points assigned to it.

It is also important to point out that the data that was initially given had quite a lot of noise in it having to do with cash-type payments. Half of our data had a tip of 0 because cash payments were collected without keeping track of the tip, which caused a ridiculously high MSE. After we came to this realization through our exploratory analysis, we filtered a larger data set to retain a similar amount number of data points, and proceeded forward.

5.2 Models and Features

After looking back at all the features and models that we evaluated, location and time of day seemed like the two most promising ones. We ultimately chose to go with location, because of its performance on the test set, the intuition of certain part of New York City being more affluent over others, as well as the conclusions reached by relevant literature. We feel strongly that if we fitted our model to include time of day, or even the day of the week, we could have even reached an even better model.

On the other hand, features like distance, time and speed, seemed like a good idea to train models on. However after we looked at the visualizations of the data, we realised that building a regressor would hardly help, as much of the data remains so close to the average.

5.3 Final Thoughts

Although it seems like our MSE has not made too much of a numerical decrease, it is important to realise the tough baseline we are already comparing our model to. It is no surprise that people will tip the suggested rates between 15% and 25%, and that is in fact what our exploratory analysis already told us. Even if our analysis helps earn a few percent better tip for a driver, it could amount to quite a bit in the long run.

Also, when looking back at our models, we feel like we could have taking a better sample that was spread across a wider time-line. For example, it would have been interesting to see whether tips were given generously in context to the holiday season, as well as other temporal features.

6. CONCLUSION

When we think of how a tip is evaluated, we generally assess the person for the quality of their service. This for the most part is intuitively true, but when we look deeper we find that it is not the only factor in play. In the case of New York City's taxi drivers, location and time both play a role in determining how much of a tip they receive. It can be said that in more affluent parts of the city and during peak hours, a higher percentage tip will be paid to any taxi driver. Hopefully with this information, taxi drivers can better plan their routines for the day in order to increase their own earnings.