Predicting NYC taxi tips using Machine Learning

1st Shaun Heffernan
Student
University of Florida
Gainesville, Florida
shaunheffernan@ufl.edu

***Abstract*—This document shows a analysis of the NYC 2023 Taxi Cab data set. It includes exploratory data analysis and model creation and performance evaluation.**

I. Introduction

Our data set is the NYC 2023 taxi cab data set, which is generated from the submission of the trip record by yellow taxi trip service providers. Each trip record includes fields capturing pick-up and drop-off dates/times, pick-up and drop-off taxi zone locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.

We want to figure out what the most important factors are in predicting tip prices, to help taxi drivers optimize the amount of tips they collect. To find the best predictors, we will train lasso and linear regression models to predict the tip amount feature of this data set.

To test our models we will consider the coefficient of determination as one of the metrics of success and report its 95% confidence interval on the validation set.

II. Preparing The Data

A. Data cleaning

To clean our data, we want to make sure that the data we provide our models are representative of the actual pattern of tipping for these yellow cab drivers. Upon inspection of the dataset, it shows that only credit card tips are recorded and cash tips are not logged in the total amount or the tip amount. This means that we do not want to use any cash transactions, because we can't train our model to predict a value that we do not know. So in order to make our model as strong as possible, we have to remove the cash transactions, which is roughly 20% of the nearly 10,000 trip dataset.

B. Feature engineering

For the attribute organizer I encode the pickup day of week and the hour. I also create the pre tip total amount category by combining all of the costs that make up the total amount before factoring in the tip. I also created the ride duration by subtracting the pickup time by the drop off time and converting it into seconds. I converted the pickup time into it's time of day and it's day of week, then I dropped the original pickup time feature. I dropped the drop off time feature because of multicollinearity with the pickup date and ride duration. I dropped the tip amount because that's the feature we are predicting. I dropped the total amount because it includes the tip amount so there would be data leakage. And I removed the payment type because we are only concerned with trips paid for by credit cards. I also turned the ratecodeid and vendorid into strings because those are categorical values that should be one hot encoded.

TABLE I
Pearson's correlation

	tip_amount	fare_amount
vendorid	0.065000	0.032000
passenger_count	0.023000	0.062000
trip_distance	0.573000	0.887000
ratecodeid	-0.046000	0.084000
pulocationid	-0.065000	-0.132000
dolocationid	-0.053000	-0.098000
payment_type	-0.334000	-0.083000
fare_amount	0.602000	1.000000
extra	0.174000	0.145000
mta_tax	-0.017000	0.016000
tip_amount	1.000000	0.602000
tolls_amount	0.474000	0.642000
improvement_surcharge	0.076000	0.174000
total_amount	0.719000	0.981000
congestion_surcharge	-0.055000	-0.220000
airport_fee	0.401000	0.615000

C. Pipelines

For my numerical pipeline I use a median imputer because it is resistant to outliers, and the standard scaler so all features are on the same scale which is necessary for lasso regularization to fairly penalize. I used the one hot encoder for the categorical pipeline because I don't want the model to predict based off of the numerical values of the attributes. I used most frequent imputer because the mode of the categories is a good substitution if there is not a value for it. For the numerical attributes I didn't include the features that comprised pre_tip_total_amount because of multicollinearity. I also didn't include the pulocation or dolocation, because the pulocationid and dolocationid are more specific versions of those features.

III. Exploratory Data Analysis

A. Pearson's correlation

As shown in Table I, for tip_amount: trip_distance, fare_amount, airport_fee, tolls_amount, and total amount all have high positive correlations which make sense because the higher the cost the higher the tip because tip is usually

a percentage of cost. For the fare_amount: the trip_distance, tip_amount, tolls_amount, total_amount, and airport_fee were all highly positively correlated. This also makes sense and has similar values to tip amount, because those features signal a higher total price, which is correlated with a higher tip and higher fare_amount. It is interesting how the total_amount has a 0.98 correlation with fare_amount but only a 0.71 with tip_amount, I would expect both of them to have similar correlations with the total_amount. The lower correlation with the tip_amounts suggests not everyone tips as a percentage of the total_amount or are not tipping at all.

Also it is important to see how payment type is negatively correlated with tip_amount, because type 0 is cash and type 1 is credit card. This is because every tip on a ride paid in cash is not logged so it is effectively 0.

B. Tip by pickup location

To find the highest tipping pickup location I searched for the mean tip across the pulocation and the pulocation ID. For the pulocation I found that Queens had the highest mean of $8.18. And for the best pulocationid I found it was ID of 134 with a mean of $18.73.

C. Tip distribution by time and day

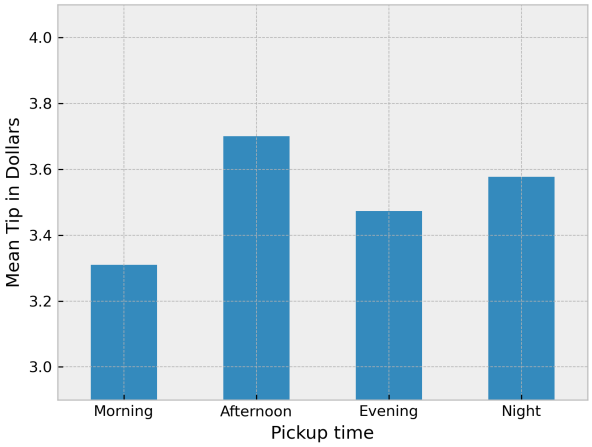

Fig. 1. Mean tip amount time of the day

Mean tip in dollars plotted over the pickup time categories. Afternoon has the highest mean tips with morning having the lowest mean tips. This shows that taxi drivers might want to prioritize working in the afternoon and night.

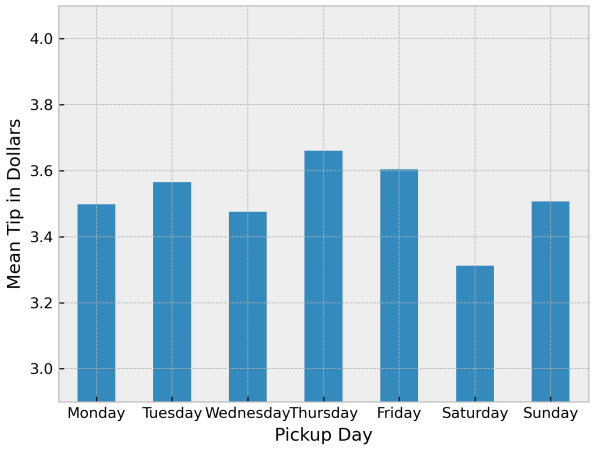

Fig. 2. Mean tip amount by day of the week

Mean tip in dollars plotted over each pickup day, from Monday through Sunday. There isn't a large variation between days, but Saturday has the lowest mean tip which is surprising because I would assume that is a day with a lot of tourists who I'd expect to tip well.

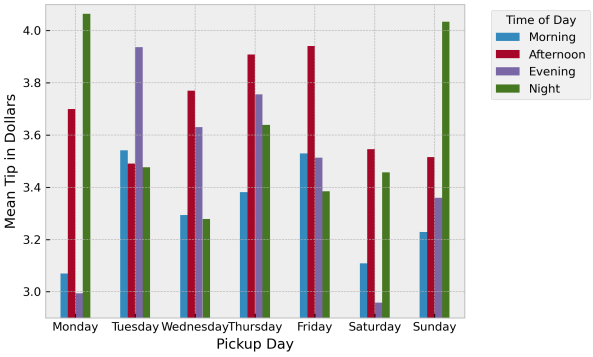

Fig. 3. Mean tip amount by day of the week

Mean tip in dollars plotted over each day of the week from Monday through Sunday. And for each day of week, showing the mean tip in dollars for the time of day. One trend is that people tip higher on the evening on Tuesday to Friday, which could be caused by people leaving the office on weekday evenings. But surprisingly Monday and Sunday night have much higher mean tips, possibly being cause by people taking trips home from across the city.

IV. Predictive Modeling

A. Linear regression

For the linear regression training with cross validation of 10 folds, the R^2 was 0.6033 with a confidence interval of (0.5301, 0.6764). The strongest coefficients were ratecodeid99 (-3.75), pre tip total (3.07), and the rate codes 1-5 had strong coefficients.

TABLE II
LINEAR REGRESSION COEFFICIENTS

	Feature	Coef
7	cat__ratecodeid_99.0	-3.750363
23	num__pre_tip_total_amount	3.068447
3	cat__ratecodeid_2.0	2.557320
2	cat__ratecodeid_1.0	1.826362
5	cat__ratecodeid_4.0	-1.247570
6	cat__ratecodeid_5.0	1.149519
4	cat__ratecodeid_3.0	-0.535268
1	cat__vendorid_2	0.283569
0	cat__vendorid_1	-0.283569
17	cat__pickup_time_of_day_Afternoon	0.184350
13	cat__pickup_day_of_week_Sunday	-0.154629
8	cat__store_and_fwd_flag_N	-0.147097
9	cat__store_and_fwd_flag_Y	0.147097
22	num__trip_distance	0.139153
19	cat__pickup_time_of_day_Morning	-0.137034
14	cat__pickup_day_of_week_Thursday	0.089194
11	cat__pickup_day_of_week_Monday	-0.080509
16	cat__pickup_day_of_week_Wednesday	0.058769
12	cat__pickup_day_of_week_Saturday	0.050275
25	num__pulocationid	0.048148
15	cat__pickup_day_of_week_Tuesday	0.041023
20	cat__pickup_time_of_day_Night	-0.038939
26	num__dolocationid	-0.021322
21	num__passenger_count	-0.011738
18	cat__pickup_time_of_day_Evening	-0.008377
24	num__ride_duration	-0.006414
10	cat__pickup_day_of_week_Friday	-0.004123

TABLE III
LASSO REGRESSION COEFFICIENTS

	Feature	Coef
7	cat__ratecodeid_99.0	-3.755795
23	num__pre_tip_total_amount	2.950557
3	cat__ratecodeid_2.0	0.998764
0	cat__vendorid_1	-0.527897
2	cat__ratecodeid_1.0	0.384616
22	num__trip_distance	0.193769
17	cat__pickup_time_of_day_Afternoon	0.171052
19	cat__pickup_time_of_day_Morning	-0.103543
13	cat__pickup_day_of_week_Sunday	-0.083691
25	num__pulocationid	0.042156
11	cat__pickup_day_of_week_Monday	-0.015917
14	cat__pickup_day_of_week_Thursday	0.008176
9	cat__store_and_fwd_flag_Y	0.000000
6	cat__ratecodeid_5.0	0.000000
24	num__ride_duration	0.000000
4	cat__ratecodeid_3.0	-0.000000
5	cat__ratecodeid_4.0	-0.000000
21	num__passenger_count	-0.000000
20	cat__pickup_time_of_day_Night	-0.000000
18	cat__pickup_time_of_day_Evening	0.000000
10	cat__pickup_day_of_week_Friday	-0.000000
8	cat__store_and_fwd_flag_N	-0.000000
16	cat__pickup_day_of_week_Wednesday	0.000000
15	cat__pickup_day_of_week_Tuesday	0.000000
1	cat__vendorid_2	0.000000
12	cat__pickup_day_of_week_Saturday	0.000000
26	num__dolocationid	-0.000000

B. Lasso regression

For the lasso regression training grid search and picking λ from 0.0001 to 1, the chosen λ was 0.01 with a R^2 of 0.6075 and a confidence interval of (0.5382, 0.6769). The strongest coefficients were ratecodeid99 (-3.76), pre tip total (2.95), and the rate codes of 2 (0.999).

This also shows how ratecodeID of (3,4,5), ride_duration, passenger_count, store and fwd flag of (Y, N), pickup time (night, evening), and pickup day (Friday, Saturday, Tuesday, Wednesday) were all dropped with a 0 coefficient. A lot of these make sense such as some ratecodes being unimportant, the duration having multicolinearity with pre tip total amount, and the flags not being important. But it is interesting how a select amount of dates were dropped completely and how night and evening was dropped. Because when looking at the Mean Tip vs Date and Time it seems there are a few patterns between certain days and time of days influencing the tip amount. For the taxi driver perspective, they shouldn't focus on the time of day and the day itself despite the EDA suggesting otherwise. They also don't need to focus on the passenger count either. The main thing to focus on would be everything that drives up the pre tip total amount, so longer rides and rides with more fees (airport and tolls).

C. Model comparison

The R^2 for the best model with lasso was 0.607546903474456 with a 95% confidence interval of (0.5382025649456625, 0.6768912420032493). For the linear regression without lasso, the R^2 across the cross validation was 0.6032818605608286 with a 95% confidence interval of (0.5301225648994846, 0.6764411562221725). They had very similar R^2, which shows that they performed very similarly with Lasso being slightly better. They also had similar confidence intervals, with the lasso model having a slightly higher lower bound but also a slightly higher upper bound. The interval size for the lasso model is still smaller which means it has higher precision. Overall their statistics were very similar with lasso having slightly better scores for R^2 and the confidence interval. This shows that the lasso regularizer wasn't that important but still added value. Meaning that feature selection wasn't that necessary because the important features overshadowed the noisy features that had small coefficients. Since the R^2 are similar and the confidence intervals overlap a lot, it's hard to choose the better model based on those statistics alone. But lasso should be chosen because it had the slightly better CI and R^2, and the feature selection will help overfitting when there is new data to be tested on.

D. Taxi driver perspective

From the perspective of the taxi driver, we can see in both models that the tip is highly influenced by the rate code of 99, the pre tip total, rate code of 2, and rate code of 1. As shown in the NYC data [1], the rate code of 99 is unknown, but 2 is JFK and 1 is the standard rate. So this means finding travelers looking to go to the airport can increase tips, and

also regular rides will also help gain more tips. To increase the pre tip total, that is mostly comprised of fees and long ride durations. Which can be increased by airport trips (airport fee) and going through tolls (toll fee). For a taxi driver looking to maximize tips, they should look for long rides, airport rides, and routes that pass through tolls because these are the best predictors for the tip amount.

V. Test Results

A. Linear regression

The linear regression model has a R^2 of approximately 0.620 which is higher than the R^2 of the linear regression model during its cross validation. Its root mean squared error is approximately 2.39 which is about 54.2% of the mean tip for this test set.

B. Lasso regression

The lasso model has an R^2 of 0.624 which is higher than the R^2 of the model on the lasso grid search. Its root mean squared error is approximately 2.38 which is about 53.9% of the mean tip for this test set.

C. Comparison to training

Both models have a root mean squared error which is about half of the value it is predicting which is pretty high, but tipping is a very unpredictable variable. It depends on many more things such as the customer's mood and tipping habits just to name a few. Despite the high error, the model still provides a valuable predictor for tips and provides which predictors are most valuable for predicting taxi tips. Both models still had slightly better R^2 than their training model, which shows that the models were not overfit or underfit to the training data.

References

[1] NYC Taxi and Limousine Commission, "Data Dictionary – Yellow Taxi Trip Records," 2023. [Online]. Available: https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

[2] NYC Open Data, "2023 Yellow Taxi Trip Data," 2023. [Online]. Available: https://data.cityofnewyork.us/Transportation/2023-Yellow-Taxi-Trip-Data/4b4i-vvec/about_data