Predicting Taxi tip prices

Group of 2: Anonymous1 and Anonymous2

ACM Reference Format:
Group of 2: Anonymous1 and Anonymous2. 2022. Predicting Taxi tip prices. 1, 1 (November 2022), 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

This report documents data analysis and models that were used to predict how much a taxi driver is tipped, using NYC-TLC Taxi Trip data set. The following sections describe exploratory data analysis, predictive tasks, model details, related literature, results and conclusions.

2 EXPLORATORY DATA ANALYSIS

The NYC-TLC taxi trip data set contains data about taxi trip records of different months and years. The yellow taxi trip records of February and June were used for this task ,and we randomly sampled 2 million data points from them. The data set provides information mentioned in Table 1.

Some extra features were extracted and added to the data set. These features are mentioned in Table 2.

Field name	Description
VendorID	ID of record provider
tpep_pickup_datetime	pickup date and time
tpep_dropoff_datetime	drop-off date and time
Passenger count	number of passengers in the vehicle
Trip Distance	elapsed trip distance in miles
PULocationID	Pickup location
DOLocationID	Drop-off location
RateCodeID	1= Standard rate 2=JFK 3=Newark 4=Nassau or Westchester 5=Negotiated fare 6=Group ride
Store_and_fwd_flag	Y = store and forward trip N = not a store and forward trip
Payment_type	1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown 6= Voided trip
Fare_amount	fare calculated by the meter
Extra	Miscellaneous extras and surcharges
MTA_tax	$0.50 MTA tax based on the metered rate
Improvement_surcharge	$0.30 improvement surcharge
Tip_amount	Amount of tip from passenger
Tolls_amount	Amount of all tolls
Total_amount	Total amount charged
Congestion_Surcharge	amount for congestion surcharge
Airport_fee	$1.25 for pick up

Table 1. Data Dictionary – Yellow Taxi Trip Records

pickup_hour	Hour of the day at pick up
pickup_day	Day of the week at pick up
dropoff_hour	Hour of the day at drop off
dropoff_day	Day of the week at drop off
trip_month	Month of the trip
duration	Duration of the trip in minutes (drop-off time - pickup time)

Table 2. New features

Pickup/dropoff_hour/day, trip_month were added since the actual timestamps can't be directly treated as features. We didn't capture a feature for time in seconds or minutes as we believe that would be very noisy. Finally, duration was added as a feature to add a relative time based feature b/w the pickup and drop.

We analyzed how various features in the dataset affect the average tip price, by modeling the variation in tip prices for different features. Plots in the following figures represent some of the interesting findings:-

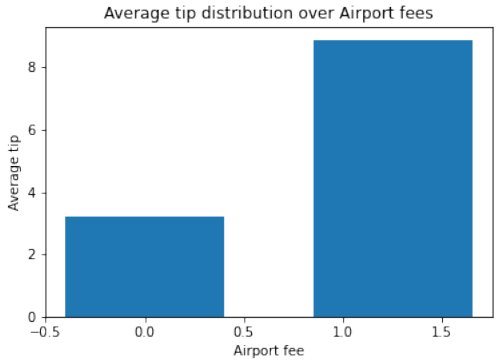

Fig. 1

XXXX-XXXX/2022/11-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

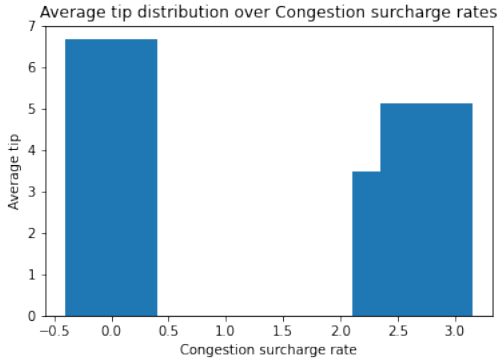

Fig. 2

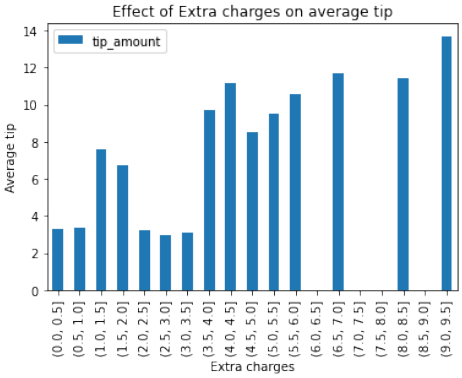

Fig. 5

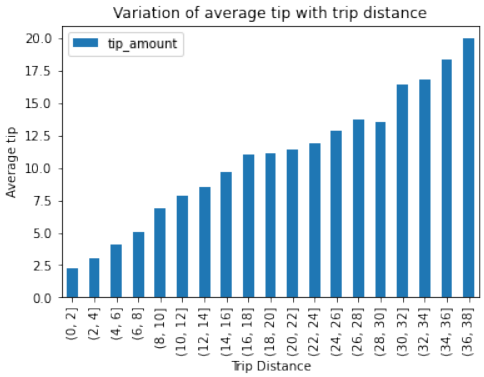

Fig. 3

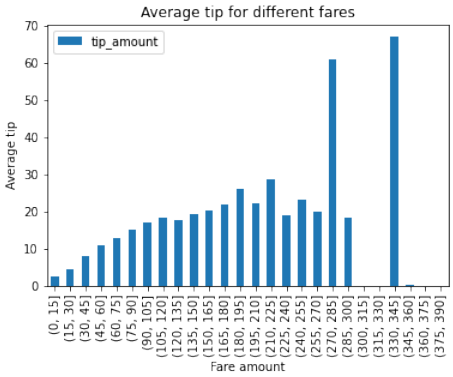

Fig. 6

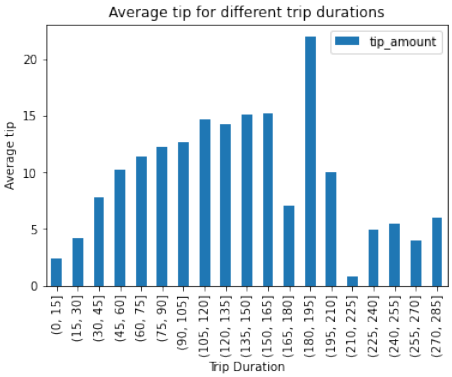

Fig. 4

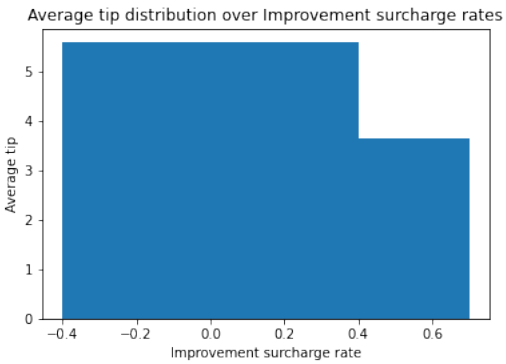

Fig. 7

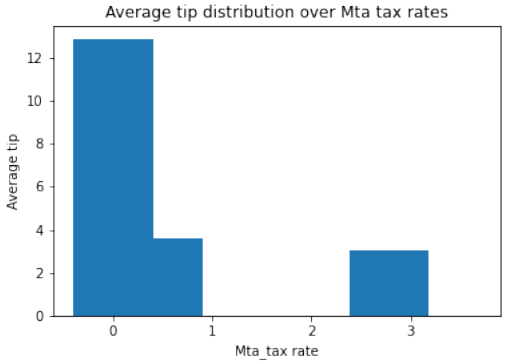

Fig. 8

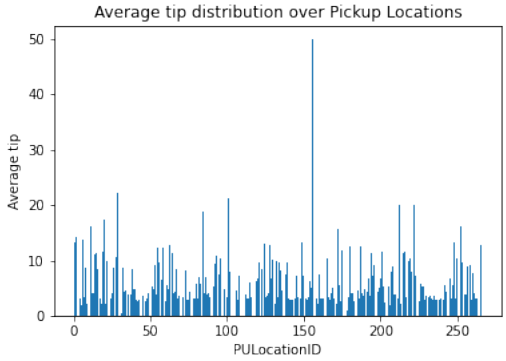

Fig. 11

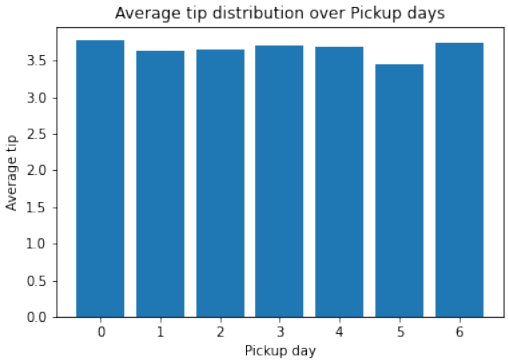

Fig. 9

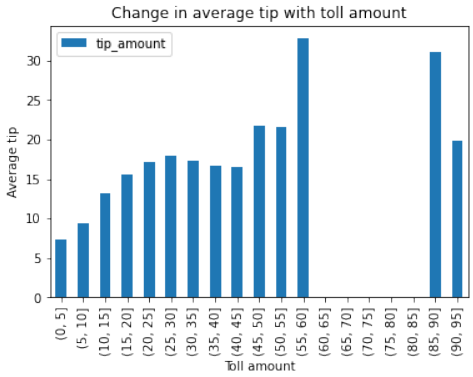

Fig. 12

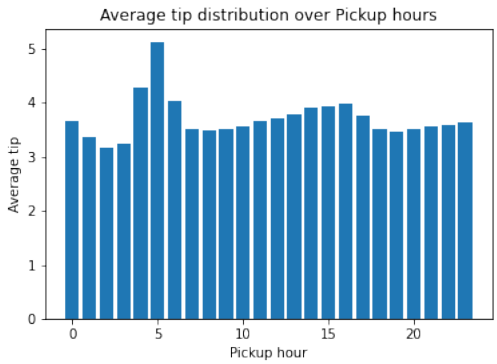

Fig. 10

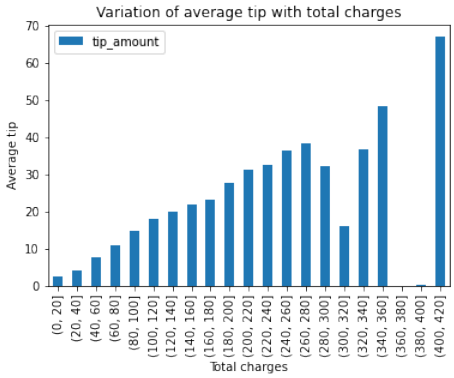

Fig. 13

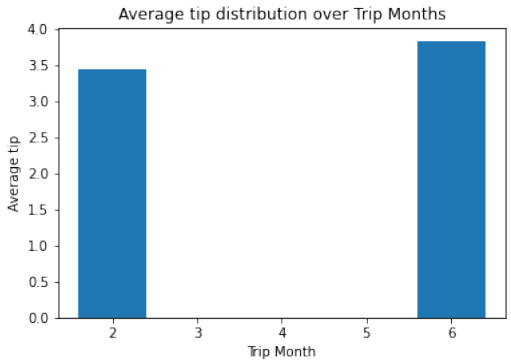

Fig. 14

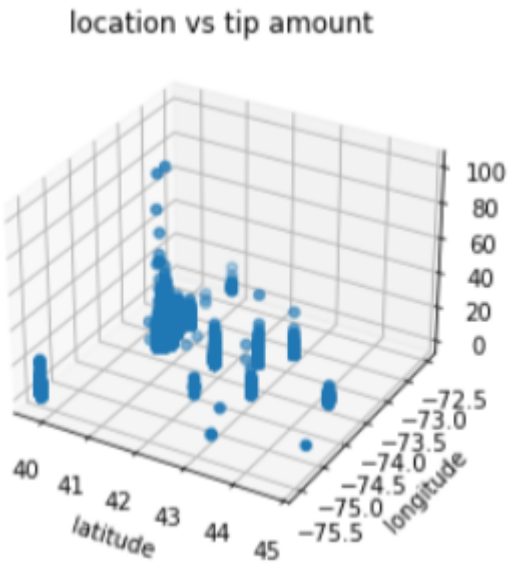

Fig. 15

Following information can be extracted from the plots :-

(1) Average tip was higher when airport fee was applied.
(2) Average tip was higher when there was no congestion surcharge.
(3) Average tip roughly varies linearly with trip distance.
(4) Average tip was not the same every hour of the day or every day of the week and varied with time.
(5) Average tip was higher when there was no improvement surcharge.
(6) Average tip was higher when MTA tax rate was 0.
(7) Average tip varied almost linearly with total charges.
(8) Average tip was higher in the month of June as compared to February (No data was extracted from NYC TLC dataset for other months, so the same may not hold for those months)

Note: While we performed analysis for various features, we have only shown the interesting graphs here. Also, we performed correlation analysis, however, we have not shown the exhaustive correlation matrix here due to its size, but, we have mentioned the results wherever appropriate.

Further information about feature construction has been discussed in following sections.

3 PREDICTIVE TASK

One predictive task for this data set is predicting the tip amount that a driver will receive from a passenger, given other trip details. This is the task that was worked upon in this project. Mean squared error measure (MSE) can be used to evaluate a model at this predictive task since a real value has to be predicted. Also, we remove the outliers in the pre-processing step from both features and tip-amounts, thus Mean Absolute Error(MAE) has not been considered. Metrics like accuracy, precision, recall and other ranking based metric cannot be used here since a real value is being predicted. To get the baseline value, a simple logistic regression model was used over the features that already exist in the data set (this does not include the new features mentioned in Table 2). To assess the validity, the data set was split into train,val and test sets. All models were trained on the training set and evaluated (using MSE) on the test set, while validation set was used to find optimal hyper parameters(λ for l2-regularization.)

3.1 Feature Generation Details

3.1.1 Cleaning. The data set had many outliers probably caused by faulty data collection methods which could skew the model predictions. So, some pre-processing was done to remove such data points. Following list summarizes the pre-processing steps :-

(1) Trips with distance less than 1 mile or greater than 40 miles were discarded, they seemed to have unreasonable trip amount and other feature values.
(2) Records with non-positive passenger count were removed.
(3) Trips with fare amount less than $2.5 were removed, as this is the base price.
(4) Cases where tip amount was greater than $200 were ignored.
(5) Records with total amount greater than $500 were discarded.
(6) The NYC TLC data set mentions that tip amount is automatically populated for credit card tips but cash tips may not be included in the data set. So, for this task, only trips with credit card payments were considered.
(7) Entries with out of domain values for RatecodeID were removed.
(8) Trips with duration > 300 minutes were discarded as for many of these trips(>80%) the distance travelled was very low, 3-4 miles only. So this data seems faulty even if we consider traffic jams.

3.1.2 Features. Some new features were created (to model temporal variations) namely - 'pickup_hour', 'pickup_day', 'dropoff_hour', 'dropoff_day', 'trip_month' and 'duration' using the existing columns in the data set 'tpep_pickup_datetime' and 'tpep_dropoff_datetime'.

The new features are also mentioned in Table 2.

However, out of all the features we removed following,

- Store and fwd: Since this feature only captures information regarding data collection process
- Payment type as we only consider card payments. For other payment types, the tip amount was either not recorded or was faulty.
- Dropoff day, Drop off hour as these had high correlation with corresponding pickup values, 0.98.
- trip_distance, which had very high correlation with total_amount, 0.9462.
- fare_amount, which had very high correlation with total_amount, 0.984247.

We categorized all the features that we used into categorical or continuous. Categorical features were converted to one-hot vector while continuous features were standardized to have 0 mean and unit variance.

- Categorical: Airport Fee, Congestion Surcharge,Improvement Surcharge, Vendor ID, Passenger Count, RateCodeID, PU Location ID, Du Location ID, pickup hour, pickup day, trip month, MTA tax
- Continuous: extra, total amount, tolls amount, duration.

3.1.3 Temporal Dynamics. In addition, the amount given as tip by a passenger may depend on how the passenger perceives a trip feature at different times. For instance, a passenger who has to reach office on time in the morning may give more tip to the driver if the trip duration is short. They may not give any tip if the duration is long and they reach the office late. So, some features (like duration) were split into different time bins. Taking duration as example(Table-3), different hour bins were created for each hour of the day. For any entry in the data set, the hour bin corresponding to the 'pickup_hour' value of the entry is assigned the value of trip duration of that entry and other hour bins are assigned a value of 0. Thus, by using this technique we capture temporal dynamics in each feature. Note, that this is in addition to providing temporal features like pickup hour to the model. As we show later, using temporal dynamics improves the performance.

We considered temporal dynamics only based on pickup hour of the day, and not on other features like trip month or pickup days, since, as can be seen in the figures-14, 9, we only observe slight variation in the tip amount based on these features. Also, to reduce the model size we created various bins out of pickup_hour, which are as follows, [0-3, 4-5, 6-7, 8, 9, 10-12, 13-15, 16, 17, 18, 19, 20, 21-23].

Since, considering this type of temporal dynamics for all the features would be infeasible in terms of the model size and the dataset size, we considered temporal dynamics for only the following features which seemed to be most effected by temporal dynamics.

- extra
- total_amount
- tolls_amount
- duration

Even for each of these, we performed independent analysis to find out the temporally dynamics based features. However, only for the duration we got improvement by considering temporal dynamics. Thus, we only considered duration for temporal dynamics for building our model.

Entry index	pickup hour	Trip duration	hour bin (00:00-03:00)	hour bin (05:00-06:00)
1	1:27	10	10	0
2	5:30	12	0	12

Table 3. Example of hour bin temporal features

3.1.4 Complex features. We also considered various complex features by which we mean that we consider the linear relation of the tip amount with a $f(feature)$ rather than just the feature values. Please note complex features of this sort only make sense for continuous features. To save ablation study time, we looked at the corresponding dynamics of how the tip amount varies with these features, and decided to experiment with following definitions of f for these features

- extra: $extra^4$, $extra^5$.
- total_amount: Identity function, since we think tip amount various linearly with total amount, except for few data points with higher total amount, fitting to which will probably cause overfitting.
- tolls_amount: $Log(1 + tolls)$, $sqrt(tolls)$, 1 was added to the tolls for numerical stability, when tolls equal zero.
- duration: piece-wise linear functions for each of these ranges, [0-155, 155-300]

We have summarized the results based on these complex features in the next sections.

3.1.5 Geographical features. We considered two representations for geolocation features, PULocationID, DOLocationID, as categorical features where each zone represents a category, and as continuous variables by converting IDs to latitude and longitude. Results with these have been discussed later.

4 MODEL DESCRIPTION

The baseline model feeds the features that already existed in the raw data set into a linear regression model to predict tip amount. This model was chosen as the baseline because it is a simple model that captures the existing features and any optimal model should at least perform better than this simple baseline model.

We tried converting location IDs to longitudes and latitudes. This model has fewer parameters because there are only 2 features for location - latitude and longitude. But this model did not perform better than the baseline model. This model is referred to as baseline + continuous location features.

Next, we tried using some complex features by replacing the basic features in baseline model by $f(feature)$ for different functions f. For instance, in this model, the feature 'extra' (which represents extra charges) is replaced by $('extra')^5$. We tried this because tip amount seemed to vary linearly with a higher power of feature 'extra'. Similarly, the feature 'tolls_amount' was replaced by log(1+tolls_amount). Here also, the plot of tip amount vs tolls amount looked like a logarithmic graph, so we tried using log(1+ tolls_amount) in place of tolls_amount. We tried multiple ablation studies to understand which representations worked best, and then the best representations were used going ahead.

Moreover, it was observed that tip_amount varies with time. This meant that there must be some parameters to explain the temporal variations. So, additional features - 'pickup_hour', 'pickup_day' etc (mentioned in Table 2) - were extracted to model direct variation of tip amount with time. This model is subsequently referred to as Baseline + direct temporal features which captures temporal features in the data.

Next, we develop on top of the previous models by adding indirect temporal logic by considering temporal dynamics of other features. It introduces the concept of hour bins for feature 'duration' which was explained in the previous section. The motivation behind this model was the idea that a passenger's perception of a feature might change with time and it might affect how much they tip the driver. For every trip record, the hour bin corresponding to the pickup hour of the trip gets the value of duration while other hour bins become 0. Every hour bin is treated as a feature and each one gets a parameter in the model. This model then imitates coefficients of the form $\beta(t)$ for a particular feature, where we have considered t to be discrete variable with value as hour bin. We tried the hour bin logic for other features apart from 'duration' but they did not perform well. So this logic was restricted to 'duration' only. This model is subsequently referred as Baseline + direct temporal + indirect temporal features.

While experimenting with the models, we ran into some scalability issues. As more features were created, a new column corresponding to the feature was added to the data set. Since, there were around 2 million entries, every new feature created 2 million more values which had to be stored in the data set. After adding a few features (especially in the last model where 24 new features come into picture), the kernel would die due to memory issues. We had to switch to a more powerful machine with more memory to handle the large augmented data set. Even then also, we had to use sparse matrix provided by scipy to generate all our results.

5 LITERATURE

The dataset we used is provided by the NYC taxi & Limousin Commission, on their website, https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data. Different versions of this dataset corresponding to different years and months have been used in the literature. Many of them have also augmented this dataset with auxiliary information such as weather information by https://towardsdatascience.com/newyork-taxi-demand-forecasting-with-sarimax-using-weather-data-d46c041f3f9c, for various prediction tasks.

There are a lot of unpublished tip prediction results available on GitHub. One such study by user Jenniferz28 achieved 0.02 MSE at test data. However, they focused on green taxi data rather than the yellow taxi data we used. josemazo performed a closely related task, a binary classification task of whether the tip would be lower or greater than 20% of the charge. A report by Sahil et al., https://cseweb.ucsd.edu/classes/sp15/cse190-c/reports/sp15/050.pdf also studied the task of predicting tips. We also found several key insights similar to them for data cleaning like tips were not registered for cash payments, and thus, are noisy and should be removed. Also, similar to them, we found that time features are quite useful. However, we also studied temporal dynamics by considering $\beta(t)$ kind of parameters in the model, which were also very helpful in reducing the MSE values. The following report by Adam: https://medium.com/@adam.hajjej.ah/nyc-taxi-tip-amount-prediction-1bacf9eac920, also derives the same conclusion as ours for the data cleaning part. However, instead of inspecting different complex feature transformations, they have focused on trying different types of models like Linear regression, random forests, etc. They also seemed to have neglected various useful features such as Miscellaneous extras and surcharges, which we have found useful for lowering mse values. In summary, these approaches have focused on feature engineering and have used sophisticated models such as random forest, gradient boosting. However, these approaches have not analysed temporal dynamics of features, and neither they have considered complex feature transformation(like log, sqrt, polynomials) as we have to improve mse values.

6 RESULTS AND CONCLUSIONS

Model	**MSE 60k data**	**MSE 2 million data**
Baseline	2.061092	2.283352
Baseline + continuous location	2.104399	–
Baseline + direct temporal	2.053177	2.268397
Baseline + direct temporal + $extra^4$	2.051852	–
Baseline + direct temporal + $extra^5$	2.051595	–
Baseline + direct temporal + Log(1 + toll_amount)	2.041603	–
Baseline + direct temporal + sqrt(toll_amount)	2.042512	–
Baseline + direct temporal + $extra^5$ + Log(1 + toll_amount)	2.039990	–
Baseline + direct temporal + complex features + indirect temporal	2.034301	2.244066

Table 4. MSE achieved by different models

We first performed ablation studies on a smaller set of data (which had 60k data points) to select optimal representation for continuous features. This study was not performed on the full data set because of

scalability issues. Only the final model and the baseline model were run on the full dataset. From the ablation study it was found that $extra^5$ and log(1+ toll_amount) are better representation for these features as compared to just 'extra' and 'toll_amount, which is also indicated by the variations we saw in the tip amount based on these features in fig-5, 13. Similarly, we found continuous location features did not perform well as compared to treating them as categorical features, which is consistent with the representation that is used throughout the literature for these features. We believe that perhaps latitude and longitude can't explain the tip amount linearly and more complex representation for these would have to be explored for getting better performance.

The model with complex features + temporal features + temporal dynamics performed better than all other models including just the baseline model on both small dataset (60k) and full dataset (2 million). These results show that temporal features are useful for explaining variation in data with respect to time. They can also explain how other non-temporal features might be perceived at different times and how this time-dependent perception can affect the quantity to be predicted. At the same time, considering better representations for features can boost the performance. Here, we could achieve better performance by considering $extra^5$ and log(1+ toll_amount) features.