Technical University of Denmark

42186 Model-based machine learning

Spring 2023

Project Report

Authors:
Jort Buitenhuis, s226499
Robin Paul Augustin Gaborit, s231147
Hans Christian Munter Hansen, s103629
Dimitrios Salogiannis, s212471

May 30, 2023
Lyngby

Table of Contents

1 Abstract

1.1 Abstract

This report aims to predict the tip amount for taxi trips in New York using a dataset from January and August 2022, which used data from the New York City Taxi and Limousine Commission, it includes pick-up/drop-off details, passenger count, trip distance, payment type, and tip amounts. The data preparation involved removing irrelevant columns, redefining the total fare amount, handling missing and invalid entries, as well introducing two additional variables for trip duration and time representation. To represent the tip amount distribution the exponential distribution was chosen, where the parameters were estimated using linear regression. Gaussian priors were used for the model's coefficient and intercept parameters. The model's performance was evaluated using correlation coefficients, mean absolute error, root mean squared error, and coefficient of determination. The visualization of the model's accuracy to predict included histograms, scatter plots, residuals plots, and line plots. The result indicates a negative correlation with low errors but with an R2 with a value of '0' giving it a limited predictive capability. Further improvements or alternative modeling approaches should be investigated to improve the accuracy.

2 Introduction

In New York, tips are an important part of the wage of employees in the service industry [2], which includes taxi cabs and thus taxi drivers. Therefore, it may be useful to be able to predict the tip sizes based on the characteristics of a trip for multiple purposes. For example, this allows taxi drivers to operate in areas or at times where tips are likely to be higher than average. At the same time, some insight into the relationship between trip characteristics and tip size could be valuable information for the New York City Taxi and Limousine Commission when making decisions on fare amounts. Finally, an accurate estimate of the tips received by taxi drivers is useful to derive the actual earnings of taxi drivers for statistical purposes.
This report consists aside from the introduction of data and model description, the results, and the conclusion. The data description consists of a description of the dataset, the steps that were undertaken to prepare the data, and an overview of the data. Furthermore, the model is described and finally, the results are presented through various metrics and figures after which these are interpreted.

3 Data description

3.1 Dataset

The dataset consists of taxi trips in New York from January and August 2022 [1], which used data from the New York City Taxi and Limousine Commission [3]. Both yellow and green taxis are included in the dataset [1] with yellow taxis being allowed to pick up passengers anywhere, whereas green taxis may only pick up passengers outside the city center [4]. The dataset includes pick-up and drop-off date and time, pick-up

and drop-off location, passenger count, trip distance, payment type, price and price composition, and tip amount [1].

3.2 Data preparation

Some columns were removed from the dataset as the pick-up and drop-off location and the different price components. The variable for the total fare amount was redefined as the total fare amount without tips because the goal of this project is to predict the size of the tips. Additionally, a variable for the trip duration was introduced which is calculated by subtracting the pick-up time from the drop-off time. After this, the entries that miss values of the leftover attributes were removed as well as the entries with negative values or zero passengers. Furthermore, entries that are not paid for by card are removed as the data set contains no tip data for those entries.
The outliers in the data set were removed, with outliers being defined as values more than three standard deviations from the mean. Furthermore, new columns were introduced with the day of the week represented by an integer and the pick-up time represented by an integer that indicates the whole hour.

3.3 Overview - EDA

After the data preparation, the data set consists of approximately 3.95×10^6 entries. In Table 1 an overview of the data set is given, with the mean, standard deviation, minimum, maximum, and percentiles of all variables except day of the week and hour of the day.

Table 1
Overview of the dataset

	Passenger count	Trip distance (mi)	Tip amount ($)	Total amount ($)	Trip duration (s)
Mean	1.41	2.64	2.78	15.24	759
Standard deviation	0.96	2.58	1.77	7.84	483
Minimum	1.00	0.01	0.00	0.30	1
25th percentile	1.00	1.11	1.85	10.30	412
50th percentile	1.00	1.80	2.36	12.80	645
75th percentile	1.00	3.04	3.32	17.30	985
Maximum	7.00	214.10	12.34	50.30	8,735

Furthermore, some extra statistics were calculated after importing the library "Pandas-profiling", which can be seen in the notebook. As a result of this, the correlation heatmap matrix was also calculated. It can be seen that the target variable "tip_amount" is highly positively correlated to the attributes "total_amount", "trip_duration", as well as "trip_distance", which was an expected result.

In addition, some extra visualization graphs were plotted, in order to extract some

extra information regarding our target variable. In the upcoming figures, the line plot of the average tip amount by an hour of the day was visualized, which helps to see the trend of the average tip amounts across different hours.

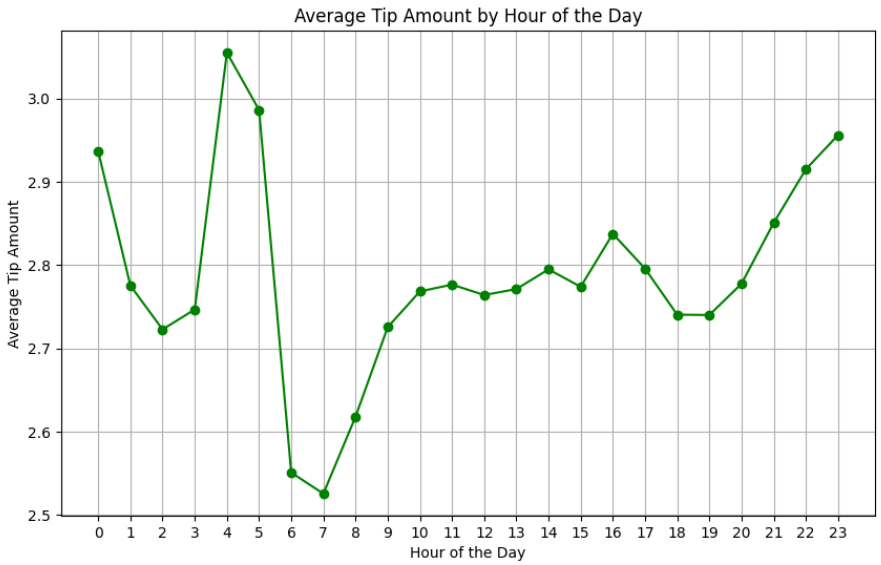

Figure 1: Average tip amount

By looking at the figure, it is visible that the customers tend to tip the drivers a bigger amount of money around 4 am, and the lowest mean of tipping amount is at 7 am. Another interesting fact that is extracted from the graphs, comes from the plot below. By taking the sample of just 1,000 instances of the dataset, the scatter plot of the trip distance and the tip amount of each trip, it can be seen that these two features are somewhat proportional to each other. Specifically, most values show that the longer the trip, the more the amount of the tip at the end.

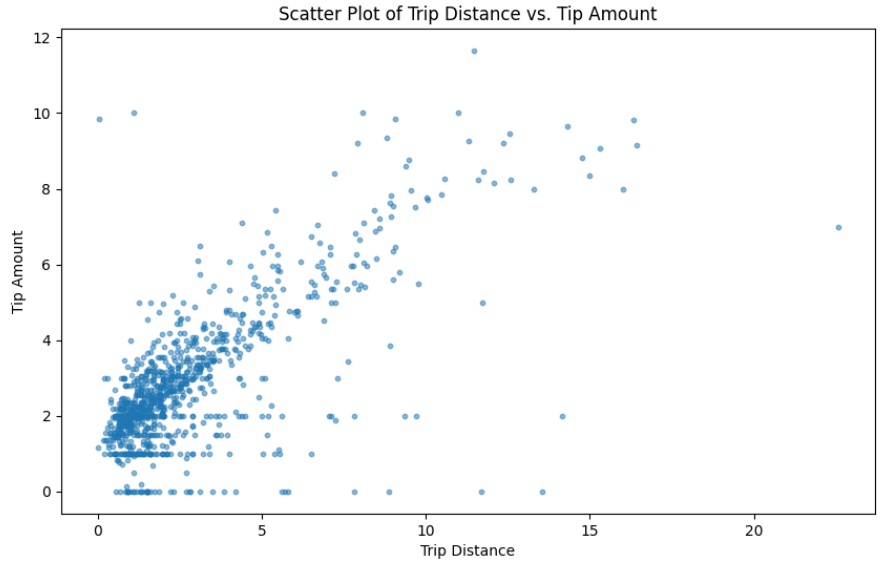

Figure 2: Trip distance vs. tip amount

4 Model description

4.1 Exponential model

Tips are non-negative numbers and theoretically without a maximum value. We considered that the value of a tip is a continuous variable (in reality, its precision can not exceed 1 cent). Consequently, the corresponding distribution should be a continuous distribution with support $[0;\infty[$.

We tried several distributions to make trip predictions: Gaussian, log-Normal, truncated Gaussian, exponential, Gamma, half-Normal, and half-Cauchy. We also tried to implement those distributions in a zero-inflated model. Such a model, similar to a mixture model, would first estimate if the tip is zero or non-zero (e.g., with a logit sub-model). Then, if the tip is non-zero, another distribution is applied to estimate the value of this tip (e.g., log-normal).
However, most of these attempts resulted in a crash of Pyro or absurd outcomes due to the particularities of the problem studied:

- Tips are non-negative values $\rightarrow$ impossible to use distributions that are not left-truncated (e.g., Gaussian distribution), otherwise some outputs may be absurd.
- A substantial number of tips are exactly equal to zero $\rightarrow$ impossible to use distributions with a support that does not include zero (e.g. log-normal) as numpyro crashes otherwise.
- Truncated distributions (other than half-distributions) do not work properly on numpyro, ancestral sampling results in (huge) unsigned integer values instead of floats.

- Tips are a continuous variable → impossible to use a zero-inflated distribution because it is limited to discrete variables (e.g. counts) on numpyro.

```
def model_truncNormal(X, obs=None):
    alpha = numpyro.sample("alpha", numpyro.distributions.Normal(0.0, 3.0))
    beta = numpyro.sample("beta", numpyro.distributions.Normal(jnp.zeros(X.shape[1]), 3.0).to_event())
    sigma = 3
    with numpyro.plate("data", X.shape[0]):
        y = numpyro.sample("y", numpyro.distributions.TruncatedNormal(loc=alpha + jnp.matmul(X, beta), scale=sigma, low=0.0), obs=obs)
    return y
```

```
[ ] #little ancestral sampling to show the problem
    rng_key, rng_key_ = random.split(rng_key)
    Xdebug = np.array([[1,1,1],[1,1,1]])
    model_truncNormal(Xdebug, rng_key)

    Array([1962718978, 4188231543], dtype=uint32)
```

Figure 3: Result of ancestral sampling with truncated Normal distribution

Consequently, we used an Exponential distribution which has the merits of being continuous and having support $[0;\infty]$. However, since it is a decreasing function, output values have necessarily a higher probability as they are closer to zero, which may not correspond to reality. Other distributions with the same support we tested (half-Normal, half-Cauchy, Gamma) had the same inconvenient and no better results.

The exponential distribution has one parameter, noted θ.

4.2 Parameters

θ is estimated as the exponential linear combination of the explanatory variables. A log-link function is applied because the parameter of an exponential distribution is positive. The coefficient parameters are noted β_i and the intercept parameter α.

$$\theta = e^{\alpha + \sum_i \beta_i x_i}$$

Thus, the likelihood distribution for the tips is the following:

$$tip \sim Exponential(tip | e^{\alpha + \sum_i \beta_i x_i})$$

4.3 Priors

Since a log-link function is in the likelihood distribution, we do not know a corresponding conjugate prior. We used Gaussian priors for the parameters β_i and α, which is a classical choice for linear regressions. The hyper-parameters mean and variance are set to 0 and 3, respectively.

$$\alpha \sim N(\alpha | 0, 3)$$
$$\beta_i \sim N(\beta_i | 0, 3)$$

4.4 Generative process and PGM

The generative process is the following:

- Draw $\alpha \sim N(\alpha | 0, 3)$

- For each explanatory variable x_i, draw $\beta_i \sim N(\beta_i|0, 3)$
- For each tip, draw $tip \sim Exponential(tip|e^{\alpha+\sum_i \beta_i x_i})$

The corresponding Probabilistic Graphical Model is the following:

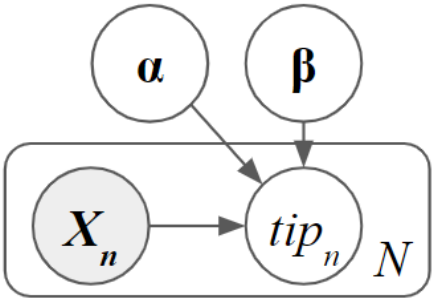

Figure 4: Probabilistic Graphical Model

5 Results

5.1 Presentation

This chapter focuses on the result of the model predicting the tip amount for taxis in New York. To evaluate the performance of the model, the following metrics are used: Correlation coefficient (corr), Mean absolute error (MAE), Root Mean Squared Error (RMSE), and Coefficent of determination (R2). Furthermore, to visualize the performance a histogram, residual plot, line plot, and scatter plot are used.

5.2 Performance Metrics

For performance evaluation, the following metrics were calculated:

CorrCoef	-0.708
MAE	1.702
RMSE	2.520
R2	0.000

5.3 Visualization

To better understand the distribution of errors between the predicted and actual tip amount a histogram is used for this visualization.
To visualize the predicted vs. the actual values the following plots are used. The scatter plot visualizes the correlation, the residuals plot shows the distribution of errors, and line plots visualize the predicted vs. actual over time.

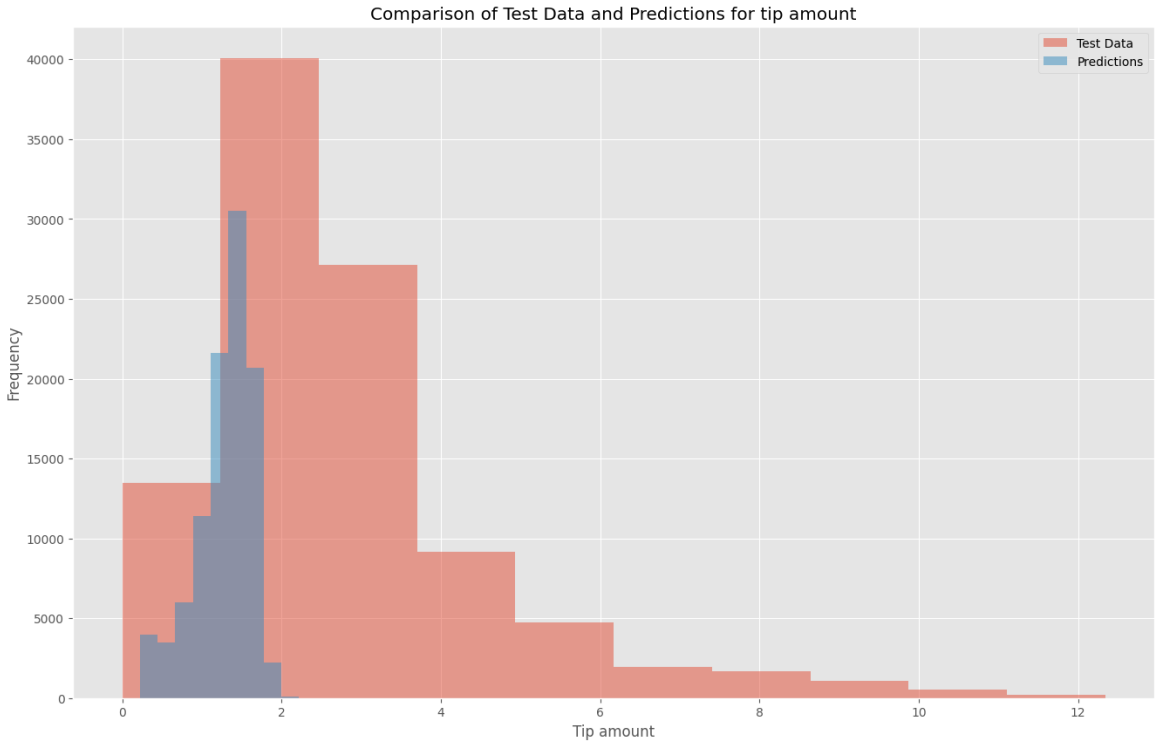

Figure 5: Comparison of Test Data and Predictions for tip amount

The histogram shown in Figure 5 gives a visual representation of the frequency distribution of the difference of errors between the predicted and actual values. Here can be seen that there's some overlapping between the test and predicted data.

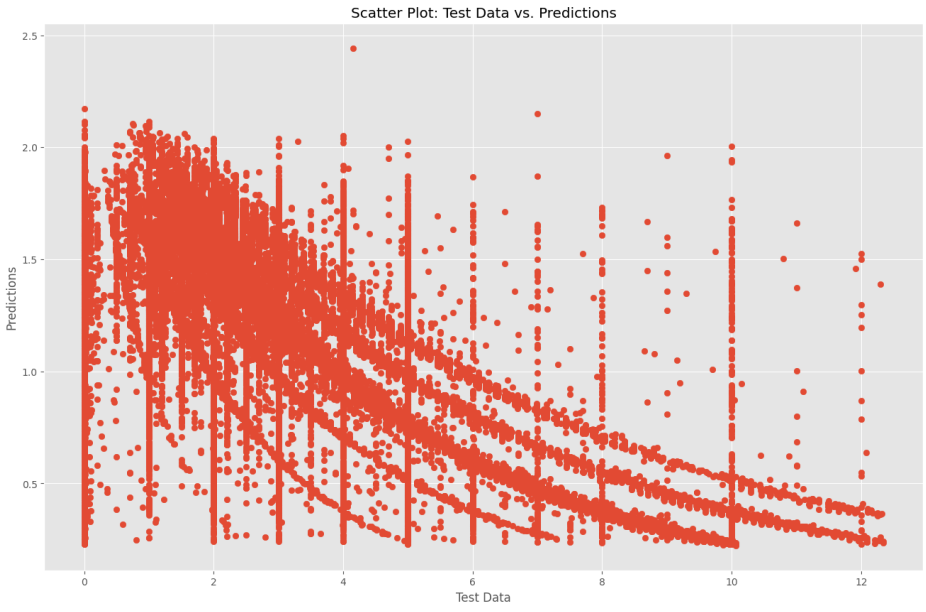

Figure 6: Scatter Plot: Test Data vs. Predictions

In Figure 6 where the CorrCoef is -0.708 indicates that there's a strong negative correlation. This means that when the predicted values increase, the actual values decrease.

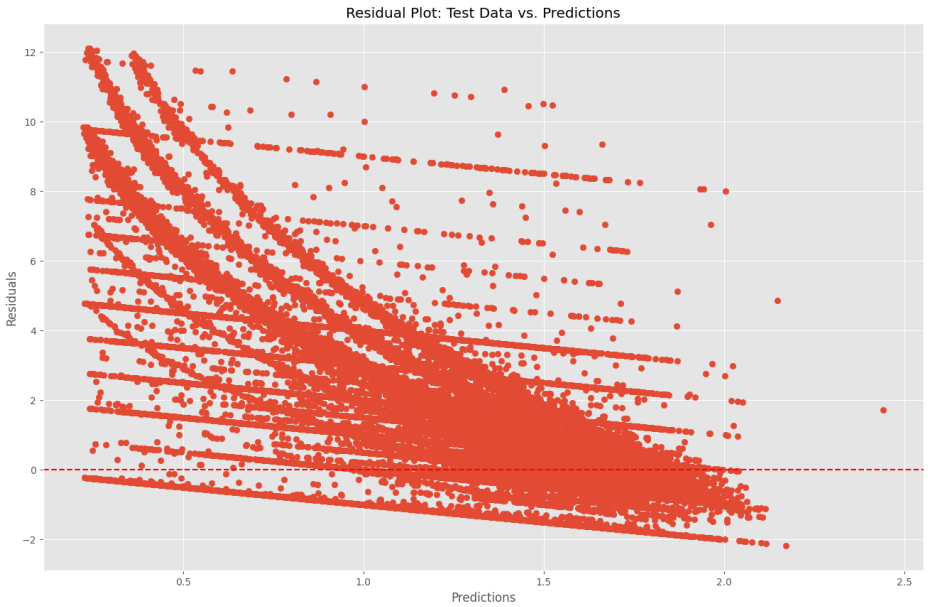

Figure 7: Residual Plot: Test Data vs. Predictions

In Figure 7 where MAE is 1.702 indicates that the model is off by 1.702 compared to the actual tip amount. This value suggests that the model is reasonable to predict the values.
A RMSE at 2.520 indicate a low relative overall error, meaning that the model predictions have a relatively low overall error. A low value for these two metrics indicates the model is reasonably accurate to predict the value.

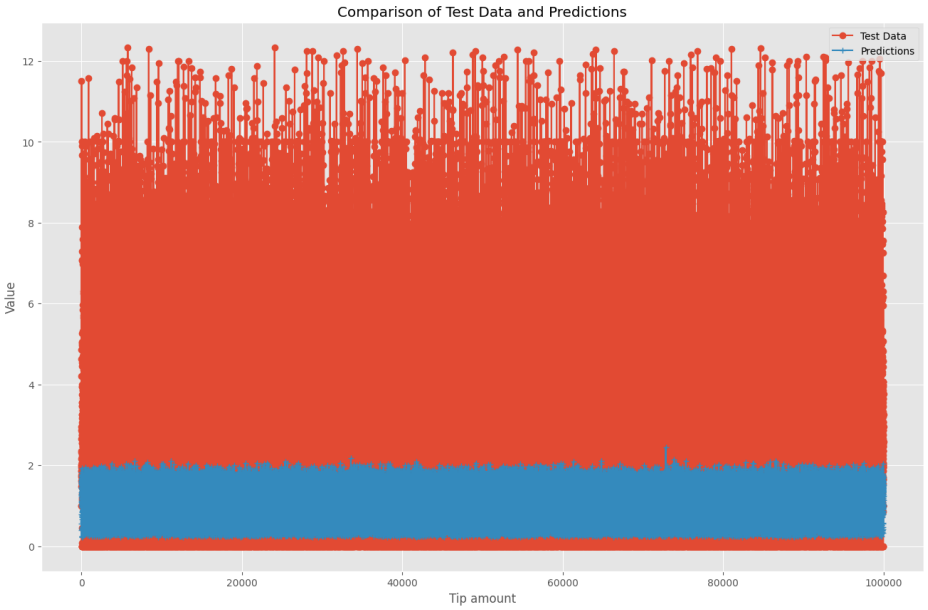

Figure 8: Comparison of Test Data and Predictions

In Figure 8 where R2 equals to 0, means that the model does not explain the variance in the actual values, this means that the model can't explain the underlying pattern nor relationships in the data and can't make meaningful predictions.

5.4 Interpretation

While the model has a negative moderate correlation between the predicted and actual values, and relatively low values for MAE and RMSE, which indicates a relatively reasonable accuracy. The value of 0 for the coefficient of determination means that the model can't explain the variability in the tip amount. Further improvements or alternative modeling approaches may be necessary to enhance the predictive capability of the model.

6 Conclusions

The purpose for this report was to predict the tip amount for taxi trips in New York. First, the data was chosen, and then it was prepared for further analysis. A model was described and implemented and then inference was performed. The result for the model was then calculated and visualized.
The result shows that even though the model have a negative correlation and relatively low MAE and RMSE, the R2 value of '0' indicate a poor predictive capability for the model.

The poor predictive capability of the model may be explained by the shape of the exponential distribution. For any values of explanatory variables, the density of probability decreases when the tip value increases. This is not in line with reality.
Further research may include the following directions

1. Manually design on Pyro a zero-inflated model accepting continuous support, if that is possible. Such a model, similar to a mixture model, would first estimate if the tip is zero or non-zero (e.g., with a logit sub-model). Then, if the tip is non-zero, another distribution is applied to estimate the value of this tip (e.g., log-normal).

2. Implement this same idea but with two different models ran sequentially. First, we infer the parameters of the first model to predict which trips will be subject to a (non-zero) tip. Then, we infer the parameters of the second to predict the values of the non-zero tips.

3. Implement a mixture model with one Gamma distribution of shape parameter < 1 and one Gamma distribution of shape parameter > 1. The former includes zero in its support. It would make predictions of tips close to zero. The latter has a mod that is not zero. It would predict the tips further from zero.

Also, tips may be substantially influenced by factors that we did not include in our model (e.g. exact origin-destination) or are not measured (e.g. customer profile).
In conclusion, this report highlights the limitation of the chosen model, and emphasizes the need to remodel to get a more accurate model.

7 Appendix

References

[1] V. Aguado. NYC Taxi Jan-Aug 2022 [Data set], 2023.

[2] freetoursbyfoot.com. Tipping in New York City - How Much Should I Tip?, n.d.

[3] New York City Taxi and Limousine Commission. TLC Trip Record Data [Data set], 2022.

[4] New York City Taxi and Limousine Commission. Your Ride, n.d.