2023

# *Taxi Tip Predictions with Parametric and Non-Parametric Linear Models*

SARA CAPOZIO | S.CAPOZIO@CAMPUS.UNIMIB.IT

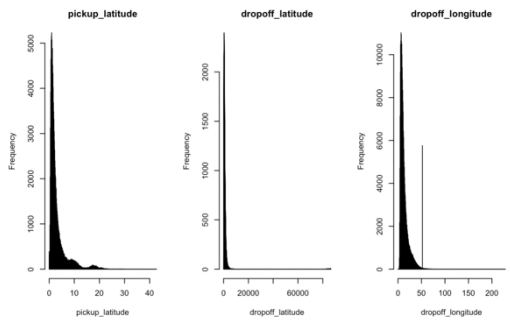
*GENERAL DATASET ANALYSIS*

The dataset under examination comprises a total of 243,179 observations and 21 predictors, including 6 qualitative variables and the remaining 15 quantitative variables. First, a careful pre-processing phase was carried out to identify the nature of the variables under analysis and detect any anomalies within the data. From the analysis, apart from the ID variable, the only variable affected by zero variance was excluded, namely the variable 'PICKUP_MONTH', as it had only one level corresponding to the month in which the observations were recorded in May.
After merging the training and test datasets to perform a unified pre-processing, an analysis was conducted on the characteristics of qualitative variables followed by quantitative variables.

*QUALITATIVE VARIABLES ANALYSIS:*

• **VARIABLES WDAY, WEEK, PICK_DOY**: After transforming the variables wday, week, and hours into factors, the consistency between the temporal variables was analyzed to identify any coding errors. Through careful analysis, it was observed that the level '1' of the variable wday, which should identify the first day of the week, actually represents the first day of the first week of the month, which happens to be Friday. Therefore, it was decided to recode the days of the week to assign them the correct coding for interpretative purposes: (1 = Friday, 2 = Saturday, 3 = Sunday, etc.). After coding the 'wday' variable, it was decided to remove the 'pick_doy' variable from the analysis because it is a combination of the 'wday' and 'week' variables. This choice was made because, for predictive purposes, it would be more interesting to evaluate whether a trip taken at a particular time or on specific days of the week significantly influences the tip received, rather than considering the individual days of the year.
• **PICKUP_HOUR**: The transformation of this variable was conceived after reading the documentation related to the taxi fare descriptions provided by the NYC government website. For all trips made during nighttime hours (8 p.m. to 6 a.m.) and peak hours (4 p.m. to 8 p.m.), there appear to be fare increases of $1 and $2.5, respectively. For this reason, it was decided to group the levels of this variable to summarize its information according to an objective criterion. By grouping, the levels of the variable were reduced from 24 to four levels: 6:00-12:00 --> morning, 12:00-16:00 --> afternoon, 16:00-20:00 --> peak hours, 20:00-6:00 --> nighttime.
• **DROPOFF_NTA_CODE AND PICKUP_NTA_CODE**: Using these variables during estimation without any modification would result in a highly complex model with a number of parameters equal to the number of levels excluding the baseline level. These variables are more informative than the Boro Code variables because they contain more detailed information about the trips made by each individual. According to the documentation, the NTA CODES are identification codes for neighborhoods, which in turn belong to larger districts. Since Manhattan is the most frequent area in our dataset, it was decided to aggregate its NTAs into macro-areas, namely: 'Manhattan_Facilities', 'Lower Manhattan', 'Upper Manhattan', and 'Upper Mid Manhattan', while for the other territorial areas, the Boro Code methodology was maintained.

*ANALYSIS OF QUANTITATIVE VARIABLES*

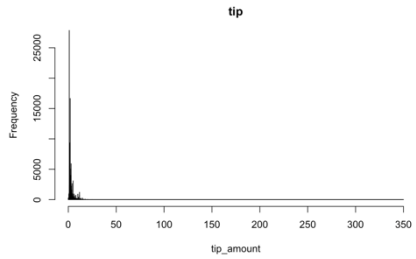



Distance in miles, travel time, and fare amount exhibit a strongly right-skewed distribution with significant outliers present in the right tails, increasing the range of variation for the examined variables. These anomalous values will be analyzed in detail later. Notably, there is a strong systematic pattern in the distribution of 'fare_amount' around the $52 mark.
First, the types of trips taken by the subjects were analyzed, revealing that 20% of them traveled on the following routes: QN98-MN17 and MN17-QN98. By grouping the levels of the NTA_code variable, it was found that the QN98 code identifies trips originating from or arriving at John F. Kennedy or LaGuardia airports, both located in Queens. According to the documentation provided by the NYC government website, taxis serving airport routes have fixed fares that do not depend on time or distance.

For this reason, it was decided to include a dummy variable called 'airport' in the model, constructed based on NTA_code levels corresponding to 'QN98'.

$$airport = \begin{cases} 0 \ \textit{The person did not take an airport route} \\ 1 \ \text{The person has taken an airport route} \end{cases}$$

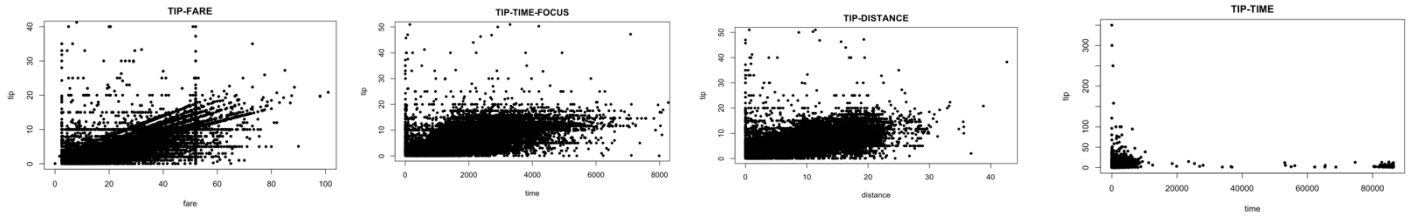
***TARGET ANALYSIS AND GRAPHICAL ANALYSIS***




The distribution of the dependent variable is strongly skewed. It is important to note the presence of significant peaks corresponding to certain tip values. These peaks are observed at the three most common tip amounts: \$1, \$2, and \$1.5. Generally, according to the documentation, if the fare is less than \$10, a tip of at least one dollar is left. Therefore, it seems that fixed tips exist. Furthermore, it's no coincidence that the tip_amount variable follows a distribution similar to that of the fare_amount. Typically, a person leaves a tip equal to 15%, 20%, or 25% of the total fare.

There is a noticeable correspondence between the two variables, especially in terms of their distribution and the peaks they exhibit.

Now, the relationship between the dependent variable and the explanatory variables is analyzed using graphical methods to preliminarily identify the variables with the greatest explanatory power.




The following plots were obtained by restricting the domain and range of the reference variables. This was necessary because the strong presence of outliers tends to obscure the true relationship between the tip and the three variables under analysis. Among the three covariates and tip, there appears to be a growing linear relationship, which is masked by systematic effects due to the presence of fixed fares and tips.

The length_time variable, in particular, exhibits a significant group of outliers that strongly impact the distribution trend of length_time. There are subjects with length_time greater than 60,000 seconds (16 hours). This is an anomaly, as the documentation states that the maximum number of hours a taxi driver can work is 12 hours.

When analyzing the relationship between the qualitative variables and the target variable, it seems that none of the qualitative variables are particularly discriminative.

***OUTLIER DETECTION AND POSSIBLE SOLUTIONS***

The following outliers were identified:

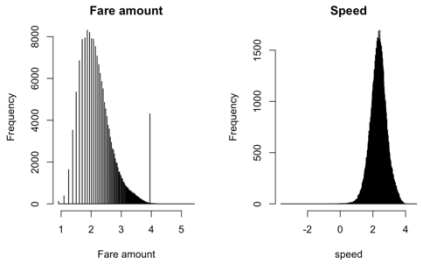
- Length_time values >= 50,000 (13 hours): These are unusual and likely errors.
- Null length_time values with positive distance and different latitude and longitude coordinates for the outbound and return trips: These observations seem to contain errors since if the subject has different coordinates but a null length_time, the distance should also be null.
- Non-zero distance values with identical outbound and return latitude and longitude coordinates and a fare of \$2.50: These observations indicate computational errors, as the activation fare is \$2.50, and if the coordinates are identical, the distance should be zero.
- Fare_amount < \$2.50: According to the documentation, the activation fee for a trip cannot be lower than \$2.50.

As the analysis continued, other outliers and computational errors were identified that are difficult to address on a case-by-case basis. To resolve this issue, it was decided to introduce a control variable:

$$speed = distance/(time/3600)$$

In this case, by constructing a new summary measure, it was possible to more easily identify additional outliers, such as individuals who traveled at speeds exceeding 120 miles per hour. In the analysis, this new variable will be kept instead of trip_distance and length_time, as it summarizes their informational contribution. The cases just listed were robustly imputed using regression trees. The outliers were managed in two ways:

1. Introducing appropriate transformations to the variables fare and speed.



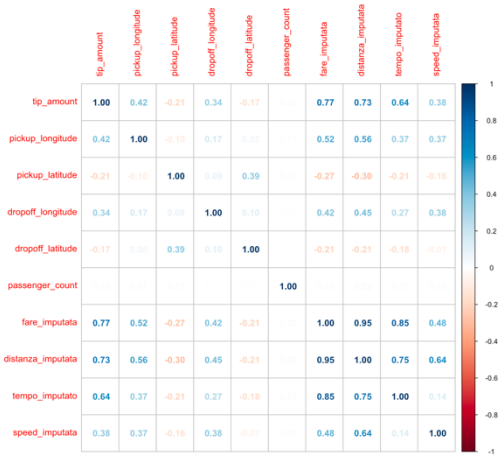

A logarithmic transformation was chosen because, in addition to being easily interpretable in terms of percentage increases, it is optimal in reducing the range of variability of the variables under analysis. By compressing the distribution and making it more symmetrical, the effect of the points that were present in the positive tails is less impactful. This transformation helps to mitigate the influence of extreme values while retaining the essential information in the data.

2. Application of Robust Regression: This is a family of regression tools designed to handle outliers in a robust manner. This methodology involves minimizing a new loss function that depends on the residuals, assigning each of them a weight function that aims to minimize the influence of values with greater magnitude through the use of an IRLS (Iteratively Reweighted Least Squares) estimator. In summary, the IRLS algorithm proceeds as follows:
   - It is initialized by penalizing a certain loss function.
   - The residuals are estimated.
   - A weight function is assigned to each residual.
   - The regression coefficients are estimated using the WLS (Weighted Least Squares) method: $b = (X'WX)^{-1}X'Wy$
   - Repeat the algorithm until convergence

   During the estimation phase, different loss functions with their respective weight functions can be used, such as the Huber, Tukey, and LAR methods. Among the three methods, the **LAR model** was chosen with the loss function $\rho(e) = |e|$ and weighted function $w(e) = \frac{1}{|e|}$.

This is because the Huber model has a weight function that, by design, is not capable of heavily penalizing strong outliers, while the Tukey model, on the other hand, is too penalizing, completely nullifying the effect of extreme outliers by assigning them a weight of zero. The goal is not to eliminate the informational contribution of outliers but rather to moderate their impact.

***MODELING***




Through the analysis of the correlogram, it is possible to evaluate whether the quantitative variables under examination are correlated with the dependent variable and to check for any potential multicollinearity. In general, the variables are not highly correlated with each other, except for 'distance,' 'time,' and 'fare,' which exhibit a strong linear relationship, as the latter is constructed as a function of the former.
From the correlation analysis, it is evident that applying models like Principal Component Analysis (PCA) or Ridge Regression might not be optimal, as these models perform particularly well when there are highly correlated variables, which was confirmed during the estimation phase. Therefore, it was decided to model the logarithm of tip_amount to ensure that its predictions never take on negative values.

The variables that were retained for analysis are as follows:
log(fare_imuted),pickup_hour,pickup_week,pickup_wday,pickup_longitude,pickup_latitude , dropoff_longitude, dropoff_latitude, vendor_id , airport , NTA_pickup_zone , NTA_droppof_zone, log(speed_imputed).

Firstly, a complete linear model was estimated, and an analysis of influential points was conducted. This step was necessary because robust regression is not efficient in the presence of leverage points or influential points, but in our case, this issue was not observed.
Subsequently, the optimal functional forms expressing the relationship between explanatory variables and the response variable were identified using Generalized Additive Models (GAM). The estimated univariate splines did not exhibit any particular functional patterns, so no transformation of the explanatory variables was carried out.
The fitted models that generated the best predictive performance are:

1. **Forward Stepwise Regression with LAD weights (cross-validated)**:
   This model was applied to the complete linear model, weighted with LAD weights, and implemented to identify the optimal model that minimizes cross-validated MAE. The resulting model was more parsimonious than the complete model, with a total of nine covariates: log(fare_imputed), airport, pickup_hour, NTA_pickup_zone, pickup_latitude, pickup_wday, vendor_id, NTA_dropoff_zone, and pickup_longitude. Once the optimal complexity of the model was identified, interactions between variables 'longitude' and 'latitude' and between the 'hour' and 'weekday' variables were introduced. This procedure significantly reduced the prediction error on the validation dataset. During the stepwise process, the first selected variables were: 'fare_amount' and 'airport.'
2. **Lasso with LAD weights**:
   The choice of the initial lambda grid values was made through several simulations. The optimal lambda value obtained via cross-validation was 0.0356. This value indicates that the penalization and, consequently, the degree of distortion inserted is very low. The Lasso model does not provide optimal predictive performance because, with a lambda value close to zero, it models a complete weighted linear model without reducing the magnitude of the regression coefficients. Analyzing the Lasso path for each variable, it was observed that 'fare_amount' contributes the most to predicting **tip_amount**, as its coefficient converges more slowly compared to the other explanatory variables.
3. **Elastic Net with LAD weights**:
   For the Elastic Net, the optimal cross-validated degree of distortion inserted into the model was higher than in the Lasso, with a lambda value of 0.05. Again, the penalization parameter was close to zero.
4. **GAM (Generalized Additive Model)**:
   GAMs, semi-parametric methods, were used not only in the initial phase to identify the optimal transformations of the explanatory variables but also as a predictive model, again using LAD weights. The covariates used in this model were those identified by the forward regression. As expected, the predictive performance of this model was very similar to that of the weighted linear model because, from the analysis of functional trends, they appeared to be linear.

| *Fitted Models* | MAE Validation |
|---|---|
| *Forward Stepwise regression cross validata con pesi LAD* | 0.590 |
| *Lasso con pesi LAD* | 0.633 |
| *Elastic Net pesato LAD* | 0.640 |
| *Modelli GAM* | 0.591 |

In conclusion, the best model obtained for predicting taxi tips in NYC based on specific characteristics of the trips is the **Forward Stepwise Regression** applied to a linear model with robust weights. The variables that contributed the most to improving its predictive performance are: log(fare_imputed), airport, pickup_hour*pickup_wday, NTA_pickup_zone, vendor_id, NTA_droppof_zone, pickup_longitude*pickup_latitude.

The link to the documentation used as a reference is as follows:
https://www.nyc.gov/site/tlc/passengers/taxi-fare.page