Analyzing Tipping Behavior in Ride-Sharing Services
Applied Statistics: Project 2

Georg Khella

March 2025

Abstract

This study examines the determinants of tipping behavior in ride-sharing services, focusing on key trip attributes such as fare, duration, shared ride status, and temporal factors. Using trip-level data from Chicago's Transportation Network Providers (TNPs), we employ logistic regression within the framework of Generalized Linear Models (GLMs) to predict the likelihood of tipping. Model selection is performed via likelihood ratio tests, and predictive performance is assessed through cross-validated Area Under the Curve (AUC) metrics. The results offer insights into consumer decision-making in digital service markets, with potential implications for ride-sharing platform policies and pricing strategies.

1 Introduction

Tipping behavior in ride-sharing services represents a complex interplay of economic, social, and contextual factors. Unlike traditional taxi services, modern ride-sharing platforms introduce additional elements such as automated fare calculation, digital payment systems, and the option to share rides with other passengers. These features influence consumer tipping patterns, making it essential to investigate the underlying determinants.

The dataset analyzed in this study includes detailed trip-level information from ride-sharing services operating in Chicago during October 2024. This dataset provides an opportunity to explore whether tipping is influenced by ride cost, trip characteristics, and temporal patterns. Specifically, we aim to answer the following research questions:

- How frequently do passengers tip in ride-sharing services?
- What role do fare amount, trip duration, and ride-sharing options play in tipping behavior?
- Are there temporal variations in tipping across different times of the day and days of the week?

To address these questions, we apply logistic regression models under the GLM framework, treating tipping as a binary response variable. The model is refined through likelihood ratio tests, and performance is validated using AUC metrics. By identifying key predictors of tipping, this research contributes to a better understanding of consumer decision-making in digital transportation services and offers valuable insights for platform optimization.

2 Data Overview and Analysis

This section provides an overview of the dataset and an exploratory analysis of it, including some relevant plots.

2.1 Dataset Description

The dataset analyzed in this study contains approximately 4.5 million ride-sharing trip records. Each record includes detailed information about a single trip, including start and end timestamps, trip duration, distance traveled, fare amount, tip amount, additional charges, and geographical details of pickup and drop-off locations. The dataset also includes categorical variables such as whether the trip was pooled and whether the passenger authorized a shared ride.

The dataset is structured with 25 variables, encompassing both numerical and categorical attributes. The key numerical variables examined in this analysis include:

- **Fare Amount (USD):** The total fare charged for the trip, excluding tips and additional fees.

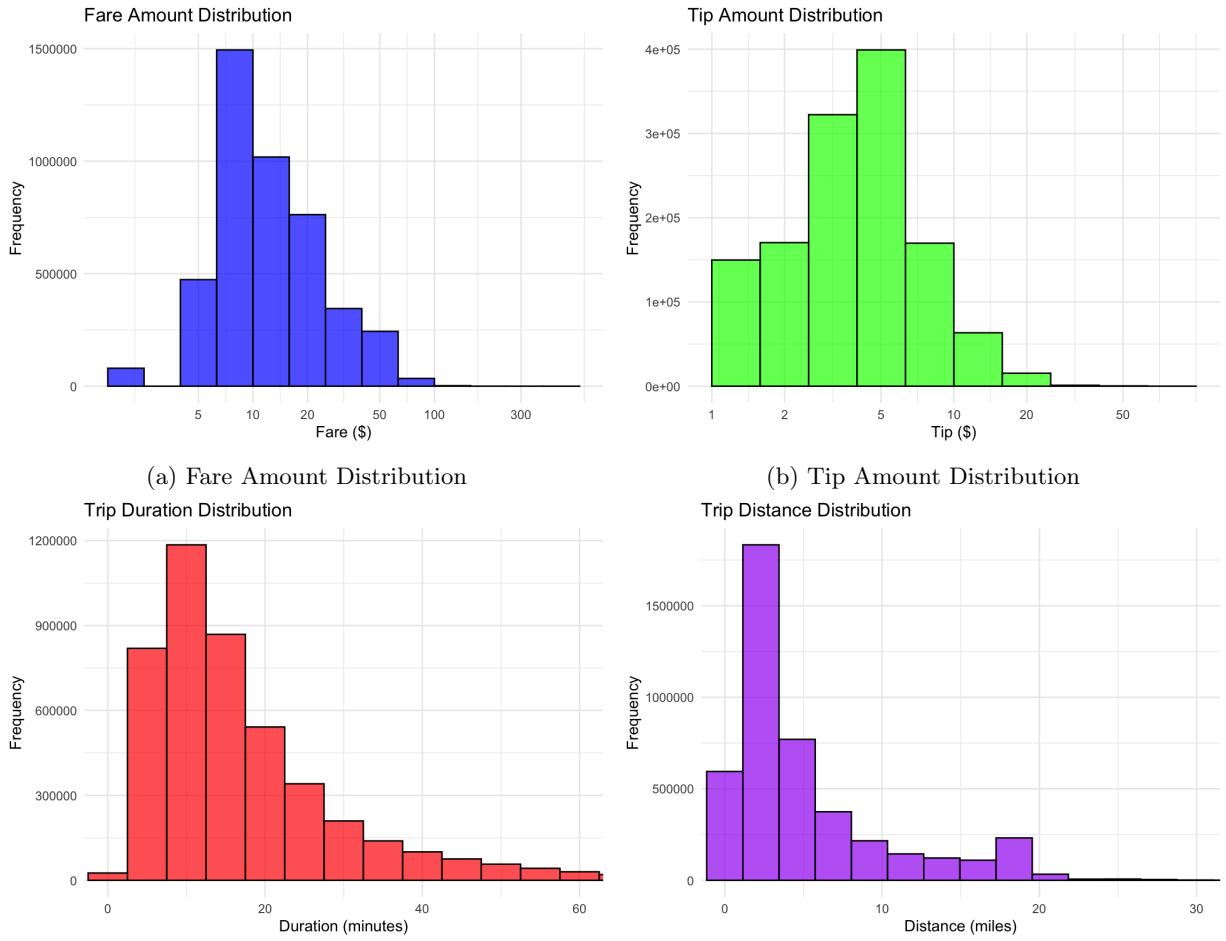
- **Tip Amount (USD):** The gratuity provided by the passenger, which may reflect service satisfaction.
- **Trip Duration (Seconds):** The length of the trip, computed from start and end timestamps.
- **Distance (Miles):** The total distance covered during the ride.

The analysis primarily focuses on understanding the distributions and relationships of these numerical variables, aiming to identify patterns and potential anomalies within the dataset.

2.2 Exploratory Data Analysis

2.2.1 Univariate Analysis

To gain insights into the dataset, we begin with a univariate analysis of key numerical features. Figure 1 illustrates the distributions of fare amounts, tip amounts, trip duration, and trip distance.



(a) Fare Amount Distribution

(b) Tip Amount Distribution

(c) Trip Duration Distribution

(d) Trip Distance Distribution

Figure 1: Figures (a)–(d): Distributions of key numerical features in the dataset. Figures (a) and (b) show the distributions of fare and tip amounts, plotted on a logarithmic scale to enhance readability. Figure (c) represents trip duration converted from seconds to minutes, considering only trips shorter than 60 minutes. Figure (d) illustrates the distribution of trip distances, in miles, as provided in the dataset.

The fare distribution exhibits a right-skewed pattern, indicating that most rides are relatively inexpensive, but a small number of trips have significantly higher fares. The log-transformed x-axis enhances readability, revealing that fare values cluster around discrete price points, potentially due to predefined pricing structures.

Similarly, the tip distribution follows a skewed pattern, with most trips having tips around a few dollars. Notably, tipping behavior may depend on factors such as ride quality, passenger preferences, and trip length.

The trip duration distribution follows a unimodal pattern, with most rides lasting between 5 and 20 minutes. A small fraction of trips exceed 40 minutes, which may correspond to long-distance rides or traffic delays.

The trip distance distribution also exhibits a strong right-skew, with the majority of trips covering a short distance, typically below 5 miles. Longer trips are less frequent, but their presence is relevant for understanding fare variations and potential outliers in the dataset.

2.2.2 Multivariate Analysis

Correlation Analysis of Key Variables Understanding the relationships between key variables is fundamental in analyzing tipping behavior in ride-sharing services. Figure 2 presents the correlation matrix for four primary variables: distance, duration, fare, and tip amount.

The results indicate a strong positive correlation between *fare* and *distance* ($\rho = 0.81$) as well as between *fare* and *duration* ($\rho = 0.78$), which is expected given that longer trips tend to be more expensive. Similarly, a high correlation is observed between *distance* and *duration* ($\rho = 0.84$), reaffirming the intuitive relationship between these two travel attributes.

Regarding tipping behavior, the correlation between *tip amount* and *fare* is moderate ($\rho = 0.41$), suggesting that higher fares are generally associated with higher tip amounts. However, the correlation between *tip amount* and *distance* ($\rho = 0.33$) or *duration* ($\rho = 0.32$) is relatively weaker, implying that factors other than trip length may influence tipping decisions.

These findings suggest that fare is a more substantial determinant of tip amount than distance or duration, highlighting the potential role of pricing strategies and passenger generosity in shaping tipping behavior.

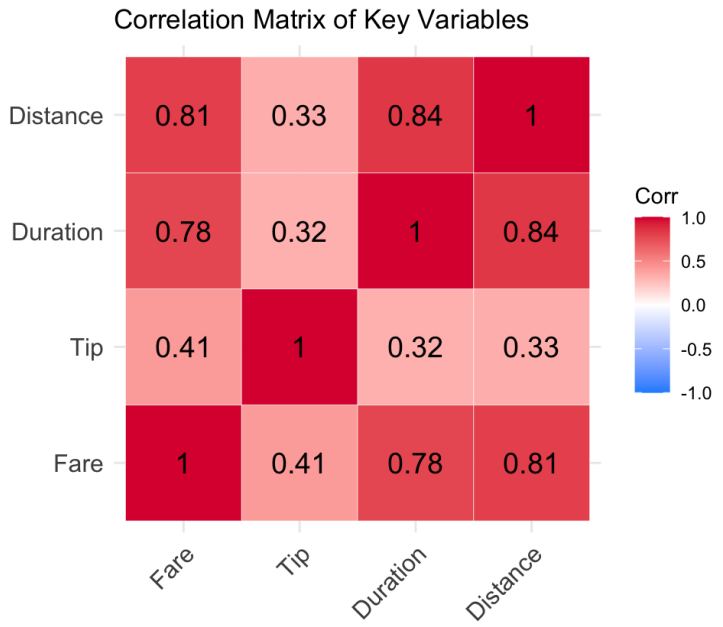


Figure 2: Correlation matrix of key variables: distance, duration, fare, and tip. Strong correlations (darker red) indicate higher positive associations.

Temporal Patterns in Tipping Behavior Temporal variations in tipping behavior provide crucial insights into passenger preferences and contextual influences. Figure 3 illustrates the average tip amount across different periods of the day and days of the week.

A notable trend emerges in tipping behavior during nighttime hours, particularly on Mondays, where the highest average tip amounts are observed. This pattern may be attributed to a lower frequency of rides at night, leading to a greater appreciation for driver availability. A general increase in tipping is also observed in evening and nighttime periods across most weekdays, suggesting that passengers might be more inclined to tip generously after late-hour trips.

On weekends, tipping behavior appears more uniform throughout the day, with slightly lower average tip amounts compared to weekdays. This could be indicative of a different ride composition, such as a higher proportion of leisure trips or shared rides, which may influence tipping decisions.

These findings underscore the role of temporal factors in shaping tipping patterns, emphasizing the potential impact of contextual variables such as ride purpose, passenger demographics, and service expectations.

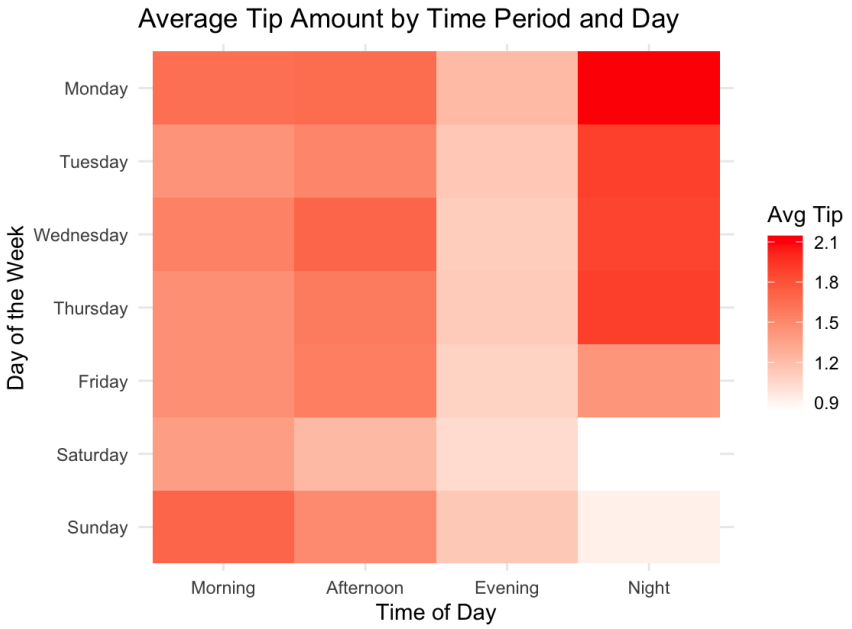


Figure 3: Average tip amount by time period and day of the week. Darker shades indicate higher average tips, with a noticeable peak in nighttime tipping on Mondays.

Effect of Distance and Fare on Tipping Behavior As indicated by the correlation matrix (Figure 2), fare and distance exhibit a strong positive correlation ($\rho = 0.81$), which is expected since longer trips typically result in higher fares. This relationship has direct implications for the analysis of tipping patterns.

Figure 4 presents two perspectives: the probability of tipping as a function of trip distance (left) and tipping as a percentage of fare across different fare ranges (right).

While there is a slight increase in tipping probability for longer trips—especially for distances exceeding 10 miles—the overall variation remains moderate. This suggests that distance alone is not a primary determinant of tipping behavior.

Regarding the percentage of fare left as a tip, a key observation is that the values remain relatively stable across fare categories, with no strong increasing or decreasing trend. However, the highest percentage is observed in the lowest fare category ($0–$5). This can be attributed to the fact that even a minimal tip (e.g., rounding up the total or leaving spare change) represents a relatively high percentage of the total fare. In contrast, for higher fares, the absolute tip amount may be larger, but its percentage relative to the fare remains fairly constant.

These findings suggest that tipping behavior may be less influenced by fare or distance than initially expected. Instead, other contextual factors—such as service quality, passenger demographics, or situational circumstances—may play a more significant role in determining tip amounts.

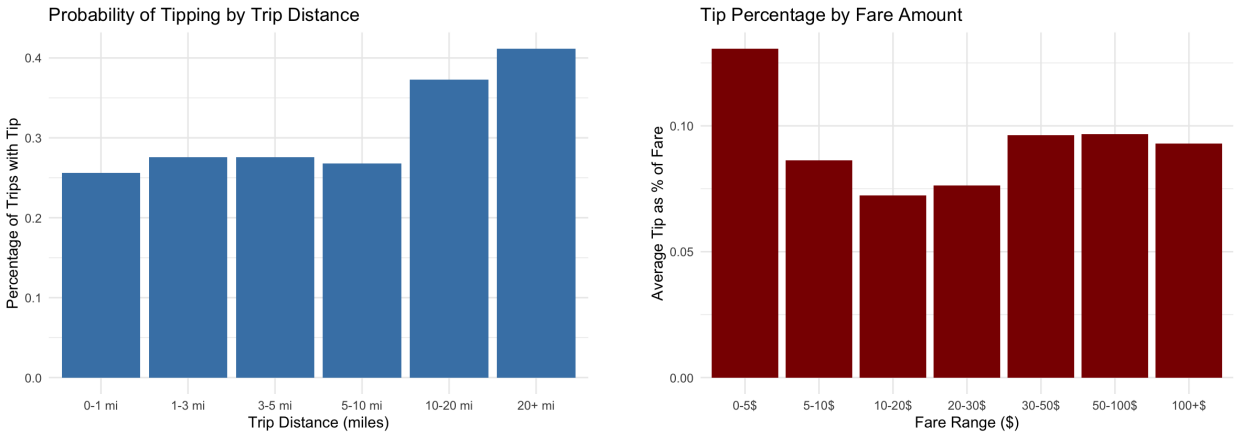


Figure 4: Comparison of tipping probability by distance (left) and tip percentage by fare (right). The trends indicate that tipping behavior remains relatively stable across fare and distance categories.

Geographical Patterns of Tipping Behavior To analyze the geographical distribution of tipping behavior, we examined tipping rates across the 77 community areas of Chicago. For clarity and visualization purposes, Figures 5 display only the top 20 areas with the highest percentage of trips that included a tip, considering both pickup and dropoff locations.

The results highlight that tipping behavior is not uniformly distributed across the city. Instead, it is concentrated in specific areas that may reflect differences in passenger demographics, trip purposes, or local economic factors. Notably, several areas appear in both rankings, such as Garfield Ridge, O'Hare, Lincoln Square, and Near South Side, indicating that these are high-tipping service zones regardless of whether they are the starting or ending point of a trip.

While some variations exist between pickup and dropoff locations, the general pattern remains consistent: areas with a higher likelihood of tipping tend to be those with significant traveler traffic, including major transit hubs and residential neighborhoods with higher-income populations. This suggests that the nature of trips, such as airport rides or rides in well-served urban neighborhoods, might influence tipping behaviors more than spatial location alone.

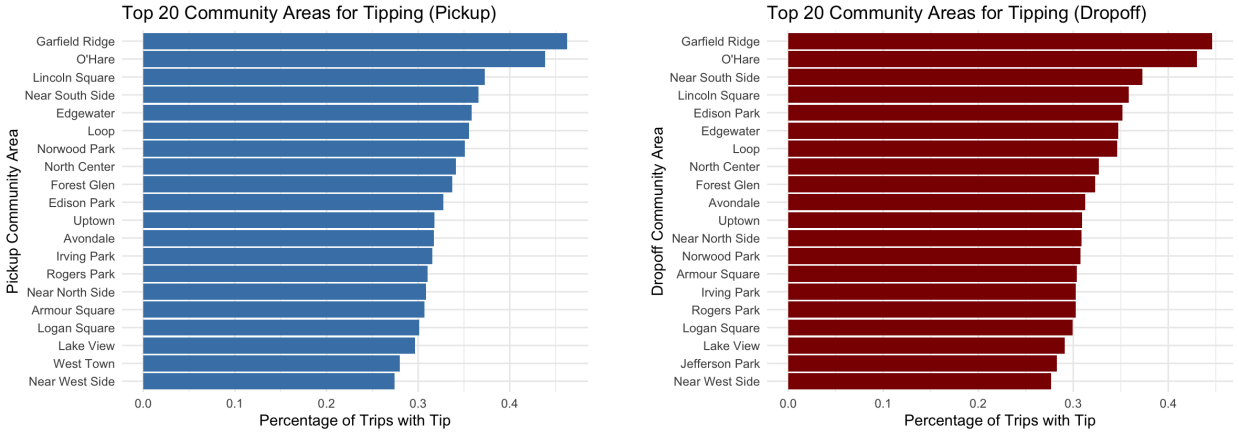


Figure 5: Comparison of the top 20 community areas for tipping, based on pickup (left) and dropoff (right) locations. The rankings highlight key service areas where tipping is more prevalent.

Impact of Shared Rides on Tipping Behavior An important factor influencing tipping behavior is whether a ride is shared or not. The results indicate a substantial difference: non-shared rides have a significantly higher tipping rate (29.3%) compared to shared rides (7.92%). This disparity suggests that passengers may perceive shared rides differently in terms of service quality or social norms related to tipping. In a shared ride, passengers might attribute the responsibility for tipping to others, or they might perceive the service as less personalized, leading to lower tipping rates.

These findings highlight the role of ride structure in shaping passenger tipping habits, reinforcing the idea that tipping is not solely determined by fare or distance but also by contextual and psychological factors.

3 Statistical Methods

The analysis of tipping behavior in ride-sharing services requires a suitable statistical framework to model the probability that a passenger leaves a tip. Given the binary nature of the response variable—whether a tip is given or not—the logistic regression model within the framework of Generalized Linear Models (GLMs) is an appropriate choice. This model allows for a probabilistic interpretation of tipping behavior while maintaining interpretability through the estimation of odds ratios.

3.1 Logistic Regression

Logistic regression is a statistical method used to model binary response variables. Given a response variable Y that takes values in $\{0, 1\}$, we assume that it follows a Bernoulli distribution with success probability π_n, such that

$$Y_n \sim \text{Bernoulli}(\pi_n),$$

where π_n represents the conditional probability that $Y_n = 1$ given a set of covariates X_n. The primary objective of logistic regression is to model this probability as a function of explanatory variables [1, 2].

3.1.1 Model Specification

Generalized Linear Models combine a random component, a systematic component, and a link function that relates the response variable to the explanatory variables. In the case of logistic regression, the systematic component is given by a linear predictor:

$$\eta_n = \beta_0 + \beta_1 X_{n1} + \cdots + \beta_p X_{np},$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are the regression coefficients and $X_{n1}, \ldots, X_{np}$ are the covariates.

Since the response variable is binary, it is necessary to use a link function that ensures the estimated probabilities remain within the interval $(0, 1)$. The standard choice for logistic regression is the logit function:

$$\text{logit}(\pi_n) = \log\left(\frac{\pi_n}{1-\pi_n}\right),$$

which is the inverse of the logistic function:

$$\pi_n = \frac{\exp(\eta_n)}{1+\exp(\eta_n)}.$$

This formulation ensures that the probability estimates are properly constrained and provides an interpretable relationship between covariates and outcome probabilities [3].

3.1.2 Interpretation of Coefficients

The coefficients β_j in logistic regression do not represent marginal changes in probability, as in linear regression. Instead, they are interpreted in terms of the odds ratio:

$$\frac{\pi(X_n)}{1-\pi(X_n)} = \exp(\beta_0 + \beta_1 X_{n1} + \cdots + \beta_p X_{np}).$$

For a one-unit increase in X_j, keeping all other predictors constant, the odds of $Y = 1$ change by a multiplicative factor of $\exp(\beta_j)$. If β_j is positive, the odds increase, whereas if β_j is negative, the odds decrease [1].

3.1.3 Estimation via Maximum Likelihood

The parameters β are estimated using Maximum Likelihood Estimation (MLE). The likelihood function for N independent observations is

$$L(\beta) = \prod_{n=1}^{N} \pi_n^{y_n} (1-\pi_n)^{1-y_n}.$$

Taking the log-likelihood,

$$\ell(\beta) = \sum_{n=1}^{N} \left[y_n \log \pi_n + (1-y_n) \log(1-\pi_n) \right].$$

Since this function is concave, optimization is performed using Iteratively Reweighted Least Squares (IRLS), which employs the Fisher scoring algorithm [3]:

$$U_n(\beta) = \sum_{n=1}^{N} \frac{(Y_n - \pi_n) X_n}{V(\pi_n)},$$

where $V(\pi_n) = \pi_n(1-\pi_n)$ is the variance function for the Bernoulli distribution.

3.2 Decision Trees

Decision trees are a non-parametric supervised learning method used for classification and regression tasks. They recursively partition the feature space into homogeneous regions, where each terminal node corresponds to a class label in the classification case. The goal of a decision tree is to construct a model that predicts the value of a target variable by applying a sequence of decision rules learned from the data [4].

3.2.1 Model Specification

A decision tree consists of a hierarchical set of splitting rules, where each internal node represents a decision based on one of the predictor variables. Given a dataset with N observations and p explanatory variables, the tree partitions the feature space into disjoint regions R_m such that each region is assigned a predicted class. The prediction for a new observation is given by:

$$\hat{Y}(X) = \arg\max_k P(Y = k \mid X \in R_m),$$

where $P(Y = k \mid X \in R_m)$ is the proportion of training observations in R_m belonging to class k.

At each step, the algorithm selects the variable X_j and threshold s that define the split:

$$X_j < s \quad \text{or} \quad X_j \geq s.$$

The splitting criterion is chosen to maximize homogeneity within the resulting subsets. The quality of a split is typically measured using an impurity function, where lower impurity indicates more homogeneous groups.

3.2.2 Impurity Measures

To determine the optimal split, the algorithm minimizes an impurity function at each node. The most commonly used impurity functions for classification trees are:

- Gini Index:

$$G(R) = \sum_{k=1}^{K} p_k(1 - p_k),$$

where p_k is the proportion of observations in region R belonging to class k. The Gini index measures the probability of misclassification if a random observation is assigned according to the class proportions in the node.

- Entropy:

$$H(R) = -\sum_{k=1}^{K} p_k \log p_k.$$

Entropy measures the uncertainty in the class distribution, with higher values indicating greater disorder.

For a given split into two subsets, R_1 and R_2, the reduction in impurity is computed as:

$$\Delta I = I(R) - \left(\frac{N_1}{N} I(R_1) + \frac{N_2}{N} I(R_2) \right),$$

where $I(R)$ is the impurity measure (Gini or entropy), and N_1 and N_2 are the number of observations in R_1 and R_2, respectively. The split that maximizes ΔI is chosen.

3.2.3 Estimation via Recursive Partitioning

The tree is built using a top-down, recursive algorithm known as recursive binary splitting. The process follows these steps:

1. At each node, for each variable X_j, find the threshold s that minimizes an impurity function.
2. Split the data into two subsets based on $X_j < s$ and $X_j \geq s$.
3. Repeat the process recursively for each subset until a stopping criterion is met, such as reaching a minimum node size or a maximum depth.
4. Assign class labels to terminal nodes based on the majority class in each region.

Since decision trees tend to overfit the training data, pruning techniques such as cost-complexity pruning are often applied. The complexity of a tree is controlled by introducing a penalty term that balances model fit and generalization. The cost-complexity criterion is given by:

$$C(T) = \sum_{m=1}^{|T|} N_m I(R_m) + \alpha |T|,$$

where $|T|$ is the number of terminal nodes, N_m is the number of observations in node m, and α is a regularization parameter. Larger values of α result in smaller, more general trees.

4 Classification and Model Evaluation

Evaluating the performance of classification models is crucial to ensure their predictive reliability and generalization capability. In the context of tipping behavior analysis, model evaluation allows us to assess the effectiveness of different classifiers in distinguishing between passengers who tip and those who do not. This section discusses the key performance metrics and methods used to validate the classification models.

4.1 Confusion Matrix and Accuracy

The confusion matrix provides a comprehensive summary of classification performance by displaying the counts of correctly and incorrectly classified observations. Given a threshold c, predictions can be defined as:

$$\hat{Y}_n = \begin{cases} 1, & \text{if } \pi_n > c, \\ 0, & \text{otherwise.} \end{cases}$$

The confusion matrix is structured as follows:

	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	True Negatives	False Positives
$Y = 1$	False Negatives	True Positives

The overall accuracy of the classifier is given by:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

While accuracy is a straightforward measure, it can be misleading in the presence of class imbalance, where one class is significantly more frequent than the other [5].

4.2 Receiver Operating Characteristic Curve and Area Under the Curve

To account for different classification thresholds, the Receiver Operating Characteristic curve is employed. It plots the true positive rate against the false positive rate for varying decision thresholds:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}.$$

A classifier with better discrimination ability will have a curve that is closer to the top-left corner of the plot. The area under the curve provides a single numeric summary of classification performance and is defined as:

$$\text{AUC} = P(\pi(Y = 1) > \pi(Y = 0)),$$

where $\pi(Y = 1)$ and $\pi(Y = 0)$ represent the predicted probabilities for a randomly selected positive and negative instance, respectively. A value close to 1 indicates strong classification performance, while a value near 0.5 suggests random guessing [3].

4.3 Cross-Validation for Model Selection

To ensure that model performance generalizes well to unseen data, cross-validation techniques are applied. The k-fold procedure involves partitioning the dataset into k equal-sized subsets. The model is trained on $k-1$ subsets and tested on the remaining subset, with the process repeated k times to compute an average performance metric. Formally, the cross-validated area under the curve is given by:

$$\text{CV-AUC} = \frac{1}{k} \sum_{i=1}^{k} \text{AUC}_i,$$

where AUC_i is the value computed on the i-th validation fold.

This methodology mitigates overfitting by providing an unbiased estimate of out-of-sample performance. For logistic regression, cross-validation helps assess the impact of different predictor subsets, while for decision trees, it guides hyperparameter tuning such as tree depth and pruning strategies [5].

5 Results

5.1 Logistic Regression

In this study, logistic regression was employed to model the probability of tipping behavior in ride-sharing services. The binary response variable indicates whether a passenger left a tip ($Y = 1$) or not ($Y = 0$). The explanatory variables include fare, trip distance, additional charges, whether the trip was pooled, drop-off locations, and the time of day when the trip started. The model aims to identify key factors influencing tipping decisions while ensuring robustness through appropriate feature selection and preprocessing.

5.1.1 Preliminary Analysis and Feature Selection

An initial logistic regression model was estimated using all available features in the dataset. However, this preliminary analysis revealed that all predictors exhibited high p-values, consistently tending to 1. This result indicates that none of the features had a statistically significant impact on the response variable, suggesting the presence of collinearity and irrelevant predictors that could reduce the model's efficiency.

To address these issues, several preprocessing steps were performed:

- **Logarithmic transformations**: The variables `Fare` and `Trip.Miles` were transformed as follows to reduce the influence of extreme values:

$$\text{log\_Fare} = \log(\text{Fare} + 1), \quad \text{log\_TripMiles} = \log(\text{Trip.Miles} + 1).$$

- **Temporal normalization**: The time of trip initiation was converted into a continuous variable:

$$\text{TimeContinuous} = \frac{\text{Hour}}{24}.$$

- **Spatial filtering**: Only trips with drop-off locations in significant areas $(50, 51, 53, 69, 71)$ were retained to improve spatial interpretability.
- **Multicollinearity assessment**: The variance inflation factor (VIF) was computed for all predictors. Features with VIF > 5 were removed to prevent multicollinearity issues.
- **Feature significance testing**: Only predictors with statistically significant contributions (i.e., with p-values below conventional thresholds) were retained in the final model.

5.1.2 Final Model and Results

After applying these preprocessing steps, the logistic regression model was re-estimated, yielding the following results:

Variable	Estimate	Std. Error	z-value	Pr(¿—z—)
(Intercept)	-3.06305	0.28809	-10.632	$< 2e-16^{***}$
log_Fare	0.46419	0.13564	3.422	0.000621^{***}
log_TripMiles	0.49326	0.11585	4.258	$2.06e-05^{***}$
Additional.Charges	0.03902	0.01651	2.364	0.018098^{*}
Trip.Pooled	-1.42321	0.39922	-3.565	0.000364^{***}
Dropoff.Community.Area50	-2.25151	0.51604	-4.363	$1.28e-05^{***}$
Dropoff.Community.Area51	-2.02418	0.46588	-4.345	$1.39e-05^{***}$
Dropoff.Community.Area53	-2.03524	0.51969	-3.916	$8.99e-05^{***}$
Dropoff.Community.Area69	-1.53861	0.24409	-6.304	$2.91e-10^{***}$
Dropoff.Community.Area71	-1.81329	0.28043	-6.466	$1.01e-10^{***}$
TimeContinuous	-0.93777	0.21496	-4.363	$1.29e-05^{***}$

Table 1: Logistic regression model estimates. Significance codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

5.1.3 Interpretation of Results

The final model demonstrates significant relationships between tipping behavior and multiple explanatory variables. The key findings include:

- **Fare and Distance**: Both `log_Fare` ($\beta = 0.464$) and `log_TripMiles` ($\beta = 0.493$) exhibit positive coefficients, suggesting that higher fares and longer trips increase the likelihood of tipping.

- **Additional Charges**: A marginal but significant effect is observed for `Additional.Charges` ($\beta = 0.039$), indicating that extra costs slightly impact tipping behavior.
- **Shared Rides**: The coefficient for `Trip.Pooled` is negative ($\beta = -1.423$), implying that passengers in shared rides are significantly less likely to tip.
- **Dropoff Location**: The negative coefficients for different drop-off areas indicate that tipping probability varies across spatial regions.
- **Time of Day**: The `TimeContinuous` variable ($\beta = -0.938$) suggests that trips occurring later in the day are associated with a lower probability of tipping.

These results highlight the primary factors influencing tipping behavior in ride-sharing services and provide a basis for further model refinement and interpretation.

5.1.4 Model Validation

To evaluate the predictive performance of the logistic regression model, we assessed key classification metrics, including the confusion matrix and the Area Under the Curve (AUC). The classification performance, using a probability threshold of 0.5, is summarized in Table 2.

	Non-Tippers (0)	**Tippers (1)**
True Positive Rate (Sensitivity)	–	4.5%
False Negative Rate	–	95.5%
False Positive Rate	0.5%	–
True Negative Rate (Specificity)	99.5%	–
Precision (PPV)	–	54.5%
Negative Predictive Value (NPV)	88.1%	–
Overall Accuracy	87.7%	

Table 2: Classification performance metrics for the logistic regression model.

The model exhibits a high specificity of 99.5%, meaning it is highly effective at identifying non-tippers. However, the sensitivity is notably low at 4.5%, indicating that the model struggles to correctly identify individuals who leave a tip. This imbalance is also reflected in the high false negative rate of 95.5%, which suggests that a large proportion of actual tippers are misclassified as non-tippers. While the positive predictive value of 54.5% implies that when the model predicts a tip, it is correct in just over half the cases, the negative predictive value is much stronger at 88.1%, confirming that most predicted non-tippers are indeed non-tippers. The overall accuracy of 87.7% may seem satisfactory at first glance, but given the class imbalance, it largely reflects the model's ability to correctly classify non-tippers rather than providing balanced performance.

To further quantify the model's discriminative ability, we computed the AUC, obtaining a value of 0.7866 when applied to the full dataset. To ensure robustness in the estimation of model performance, we also performed a five-fold cross-validation, yielding a slightly lower but consistent AUC of 0.7811. These values indicate a moderate predictive capability, significantly better than random classification (AUC = 0.5), yet still leaving room for improvement.

The model's performance is further illustrated through the Receiver Operating Characteristic (ROC) curve in Figure 6. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1 - specificity), providing a visual representation of the model's ability to distinguish between tipping and non-tipping behavior. The curve deviates significantly from the diagonal reference line, confirming that the model performs better than chance. However, the presence of a relatively shallow slope in certain regions suggests that further optimization could enhance its predictive accuracy.

One of the key challenges identified in this analysis is the effect of the classification threshold on model performance. The default threshold of 0.5 minimizes overall misclassification error, but it prioritizes specificity over sensitivity, leading to a poor detection rate for tippers. By lowering the classification threshold (e.g., to 0.3 or 0.4), the sensitivity could be improved at the cost of increased false positives. This trade-off is crucial for applications where failing to detect a tipping event carries more consequences than falsely predicting one.

Furthermore, the relatively moderate AUC values indicate that tipping behavior is likely influenced by factors beyond those captured in the current dataset. Psychological, situational, and personal preferences may play a role that logistic regression, based solely on numerical and categorical trip attributes, struggles to model. More advanced techniques, such as non-linear models or ensemble learning methods, could be explored to better capture complex relationships.

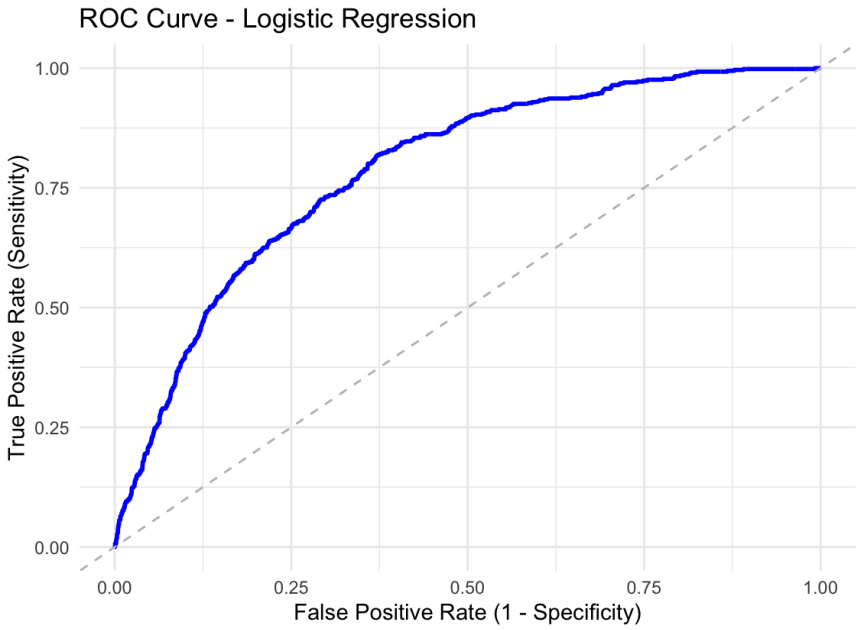


Figure 6: Receiver Operating Characteristic (ROC) curve for the logistic regression model.

5.2 Decision Trees

5.2.1 Final Model and Interpretation

Building on the features identified as significant in the logistic regression model, we employed a decision tree classifier to predict tipping behavior in Chicago ride-sharing trips. The tree was constructed using the *Gini impurity* criterion to determine the optimal splits. Gini impurity is defined as:

$$G = 1 - \sum_{i=1}^{C} p_i^2, \tag{1}$$

where p_i is the proportion of instances belonging to class i. A lower Gini impurity value indicates a more homogeneous node, guiding the tree towards maximally informative splits.

The resulting decision tree is visualized in Figure 7. The most influential splitting variable was *log_TripMiles*, reinforcing its significance from the logistic regression analysis. Shorter trips exhibited a lower likelihood of tipping, while longer trips, particularly those exceeding a threshold of 2 miles, were more likely to result in tips. Additionally, *TimeContinuous* played a role in finer decision boundaries, indicating temporal variations in tipping behavior.

Feature importance was also assessed, as depicted in Figure 8, highlighting *log_TripMiles* and *log_Fare* as the dominant predictors.

5.2.2 Model Validation

To evaluate the predictive performance of the decision tree, we computed the confusion matrix (Table 3) and assessed the model's AUC.

	Non-Tippers (0)	**Tippers (1)**
True Positive Rate (Sensitivity)	–	4.5%
False Negative Rate	–	95.5%
False Positive Rate	0.5%	–
True Negative Rate (Specificity)	99.5%	–
Precision	–	54.5%
Negative Predictive Value	88.1%	–
Overall Accuracy	87.7%	

Table 3: Classification performance metrics for the logistic regression model.

The decision tree achieved an AUC of 0.743, as illustrated in the ROC curve in Figure 9.

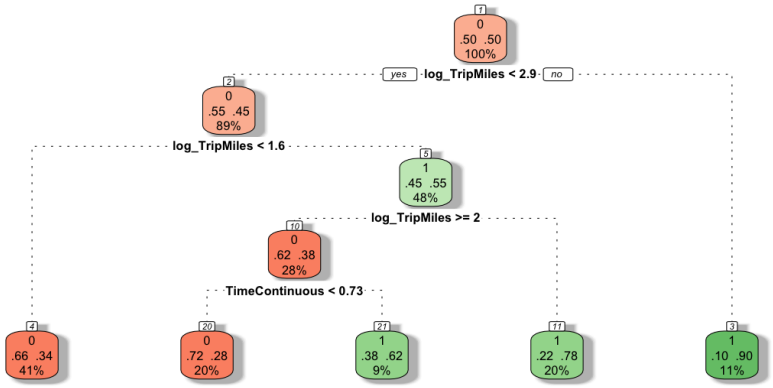


Figure 7: Decision tree structure for tipping prediction. The first split occurs at *log_TripMiles* ¡ 2, followed by conditions on *TimeContinuous* and additional features.

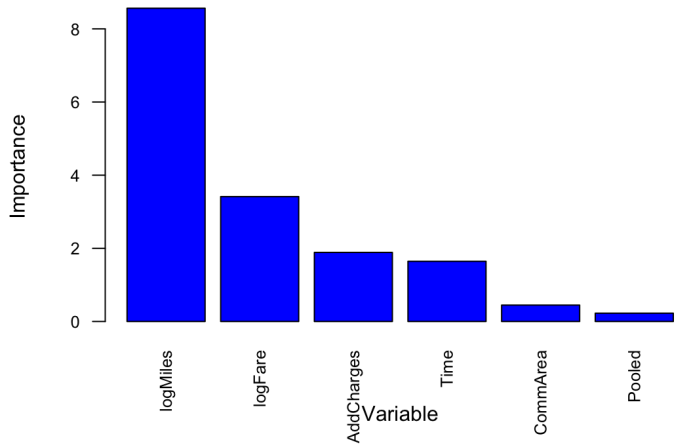


Figure 8: Variable importance in the decision tree model, with *log_TripMiles* being the most significant.

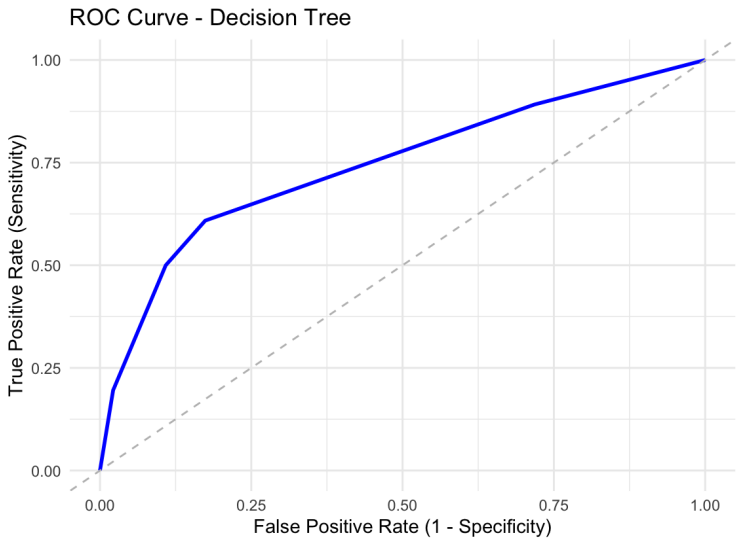


Figure 9: ROC curve for the decision tree model, showing an AUC of 0.743.

While the decision tree model provides an interpretable framework for predicting tipping behavior, its performance is slightly lower than that of logistic regression. The trade-off between interpretability and predictive power remains a consideration in selecting the optimal modeling approach for ride-sharing tipping predictions.

6 Conclusion

This study aimed to analyze the determinants of tipping behavior in ride-sharing services, leveraging logistic regression as the primary predictive model. The decision to select logistic regression over alternative methods, such as decision trees, was driven by its superior predictive performance and interpretability. With an AUC of 0.7866, logistic regression demonstrated a more robust capacity to distinguish between tippers and non-tippers compared to the decision tree classifier, which had a lower AUC of 0.743. Furthermore, the estimated coefficients provided meaningful insights into the relative impact of key variables on tipping likelihood.

The analysis revealed that tipping occurs in approximately 25.8% of ride-sharing trips, emphasizing that the majority of passengers do not tip. This behavior varies considerably based on trip attributes, including fare amount, trip distance, and time of travel. Higher fares and longer trips were found to be strong predictors of tipping behavior, as indicated by the positive and significant coefficients of these variables. This suggests that passengers might perceive longer rides as requiring more effort from drivers, leading to an increased likelihood of tipping.

Additionally, temporal patterns played a significant role in tipping behavior. The negative coefficient associated with time of day indicated that tipping probability declines as the day progresses. A notable observation was the peak in tipping during late-night hours on Mondays, which may be attributed to a different composition of riders during those times. These findings suggest that situational factors, such as passenger mood and ride purpose, influence the decision to tip.

Spatial characteristics also proved to be influential in tipping behavior. Certain drop-off locations exhibited significantly lower tipping rates, highlighting the potential effect of socio-economic factors or regional norms on tipping tendencies. This could be driven by variations in passenger demographics or different expectations regarding service compensation across neighborhoods.

One of the most striking findings of this study was the substantial negative impact of shared rides on tipping probability. The logistic regression model confirmed that passengers in shared rides were significantly less likely to tip compared to those in private rides. This supports the idea that the presence of multiple passengers dilutes the sense of individual responsibility for tipping, or that shared rides are perceived as a lower-tier service where tipping is less customary.

Despite the model's moderate predictive success, it is important to acknowledge potential areas for improvement. The relatively low sensitivity of the classifier suggests that additional factors not included in the dataset may influence tipping behavior. Psychological factors, such as passenger-driver interactions and service satisfaction, could provide further explanatory power. Future research could benefit from incorporating driver ratings, passenger feedback scores, or contextual trip details to enhance the model's predictive

capabilities.

Overall, logistic regression proved to be a valuable tool for understanding the factors influencing tipping behavior in ride-sharing services. The findings offer practical implications for ride-sharing platforms, which could use these insights to develop strategies for encouraging tipping, such as personalized incentives or optimized fare structures. Furthermore, this study underscores the broader significance of tipping as a socio-economic behavior shaped by economic rationality, social norms, and contextual variables.

References

[1] Dobson, A. J., & Barnett, A. G. (2018). *An Introduction to Generalized Linear Models*. CRC Press.

[2] Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. CRC Press.

[3] Mhalla, L. (2025). *Generalized Linear Models - MATH-516 Applied Statistics*. Course Notes.

[4] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth & Brooks/Cole.

[5] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.