ECO 520

Group: Liana Cramb, Rohil Kansagara and Viraj Tanti

**Final Project**

Table of Contents

[Variable List 2](#_Toc193301603)

[Business Idea 3](#_Toc193301604)

[Motivation: 3](#_Toc193301605)

[Our Question 4](#_Toc193301606)

[Data Analysis Method Plan: 4](#_Toc193301607)

[Data and Empirical Methodology 5](#_Toc193301608)

[Data Set: 5](#_Toc193301609)

[Summary Statistics: 5](#_Toc193301610)

[Scatter Plots 6](#_Toc193301611)

[Regression Methodology 8](#_Toc193301612)

[Full Data Analysis 9](#_Toc193301613)

[Correlation Analysis 9](#_Toc193301614)

[Proc Frequency Analysis 10](#_Toc193301615)

[ANOVA Analysis on Shared\_Trip\_Authorized 10](#_Toc193301616)

[Hierarchical Cluster Analysis Used to Determine Number of Clusters 11](#_Toc193301617)

[Pseudo T-Squared Line Plot 12](#_Toc193301618)

[Dendrogram 12](#_Toc193301619)

[Non-Hierarchical Cluster (K-Means) Analysis: 13](#_Toc193301620)

[K-Means Scatter Plots 14](#_Toc193301621)

[ANOVA Test on Tips based on Clusters 17](#_Toc193301622)

[Multiple Regression Model 18](#_Toc193301623)

[Stepwise Regression Model 19](#_Toc193301624)

[Performance of the Regression Models 19](#_Toc193301625)

[Logistic Model 20](#_Toc193301626)

[ROC Plot 20](#_Toc193301627)

[Performance of Model 21](#_Toc193301628)

[Random Forest Model 22](#_Toc193301629)

[Best Number of “mtry” 22](#_Toc193301630)

[Best Number of “ntrees” 22](#_Toc193301631)

[ROC Plot 23](#_Toc193301632)

[Performance of the Model 23](#_Toc193301633)

[Neural Network Analysis 24](#_Toc193301634)

[Model 1 (One hidden Layer and Two Neurons) 24](#_Toc193301635)

[Model 2 (Two hidden Layer and Three Neurons) 25](#_Toc193301636)

[Performance of Neural Network Model 1 26](#_Toc193301637)

[Summary of Project 27](#_Toc193301638)

[Bibliography 28](#_Toc193301639)

[Appendix 28](#_Toc193301640)

# Variable List

|  |  |
| --- | --- |
| **Variable Name** | **Variable Description** |
| Trip\_Id | Unique identifier for the trip |
| Trip\_Start\_Timestamp | When the trip started (rounded to the nearest 15 min) |
| Trip\_End\_Timestamp | When the trip ended (rounded to the nearest 15 min) |
| Trip\_Seconds | Time of the trip in seconds. |
| Trip\_Miles | Distance of the trip in miles. |
| Pickup\_Census\_Tract | The Census Tract where the trip began. Outside of Chicago are blank. |
| Dropoff\_Census\_Tract | The Census Tract where the trip ended. Outside of Chicago are blank. |
| Pickup\_Community\_Area | The Community Area where the trip began. Outside of Chicago are blank. |
| Dropoff\_Community\_Area | The Community Area where the trip ended. Outside of Chicago are blank. |
| Fare | The fare for the trip, rounded to the nearest $2.50. |
| Tip | The tip for the trip, rounded to the nearest $1.00. Cash tips will not be recorded. |
| Additional\_Charges | The taxes, fees, and any other charges for the trip. |
| Trip\_Total | Total cost of the trip. This is calculated as the total of the previous columns, including rounding. |
| Shared\_Trip\_Authorized | Whether the customer agreed to a shared trip with another customer, regardless of whether the customer was actually matched for a shared trip. |
| Trips\_Pooled | If customers were matched for a shared trip, how many trips, including this one, were pooled. All customer trips from the time the vehicle was empty until it was empty again contribute to this count, even if some customers were never present in the vehicle at the same time. Each trip making up the overall shared trip will have a separate record in this dataset, with the same value in this column. |
| Pickup\_Centroid\_Latitude | The latitude of the center of the pickup census tract or the community area if the census tract has been hidden for privacy. Outside of Chicago are blank. |
| Pickup\_Centroid\_Longitude | The longitude of the center of the pickup census tract or the community area if the census tract has been hidden for privacy. Outside of Chicago are blank. |
| Pickup\_Centroid\_Location | The location of the center of the pickup census tract or the community area if the census tract has been hidden for privacy. Outside of Chicago are blank. |
| Dropoff\_Centroid\_Latitude | The latitude of the center of the dropoff census tract or the community area if the census tract has been hidden for privacy.  Outside of Chicago are blank. |
| Dropoff\_Centroid\_Longitude | The longitude of the center of the dropoff census tract or the community area if the census tract has been hidden for privacy.  Outside of Chicago are blank. |
| Dropoff\_Centroid\_Location | The location of the center of the dropoff census tract or the community area if the census tract has been hidden for privacy.  Outside of Chicago are blank. |
| Date | Date of trip (in form of “Year-Month-Day” |

# Business Idea

## Motivation:

Rideshare services are a popular mode of transportation in Chicago, offering the convenience of booking a ride through a smartphone app and generally lower fares compared to taxis. Uber and Lyft dominate the rideshare market in the city, but both companies face challenges. Many drivers struggle with low earnings, largely due to passengers not tipping.

A **2019 University of Chicago study**found that, “Uber customers tip on roughly 16% of rides. Those who do tip add an average of $3.11, about 26% of their fare. However, the paper also found that nearly 60% of ride-share customers never tip, while only about 1% always tip.” This lack of tipping significantly impacts drivers, many of whom rely on tips to earn a livable income. A 2024 Huffington Post article highlighted this issue, stating: “When you get a reminder to tip, during the ride and after, you might choose to ignore it. It may seem like a small thing to you, but according to drivers, not tipping makes it harder than ever for them to make a decent living behind the wheel.”

If rideshare companies continue to lose customers due to rising costs, the pressure to tip and the decline in drivers caused by low wages, it poses a threat to their business. This analysis will explore issues related to tipping behavior, possibly uncover potential patterns, and assess whether the findings can drive meaningful changes.

## Our Question

Through our data analysis, we seek to answer two questions: Do any variables in our dataset significantly influence tipping, and can we accurately predict whether a passenger will tip or not? We will examine numerical variables, such as Total\_Trip, Trip\_Miles, and Trip\_Seconds, and categorical variables, such as trip date, start and end times, and the coordinates of the pick-up and drop-off locations. By exploring these variables, we aim to identify a passenger’s decision to tip or not. Understanding these influences could help rideshare companies adjust their business model and create better opportunities for drivers to earn extra income. In addition, identifying reasons why passengers choose not to tip could highlight areas of the rideshare experience that need improvement, as tipping can reflect a customer’s overall experience with the trip.

## Data Analysis Method Plan:

To answer the research question, we will first use summary statistics to provide an overview of the dataset, followed by correlation analysis to explore relationships between variables and identify the most relevant ones for further investigation. Next, cluster analysis will be performed to identify distinct groups of rideshare trips, while Analysis of Variance (ANOVA) will assess the impact of categorical variables on tipping behavior. Multiple regression analysis will then be used to examine the effects of numerical variables. Finally, logistic regression will predict whether a passenger will tip, and machine learning techniques, including neural network and random forest analysis, will be employed to uncover complex patterns in tipping behavior.

# Data and Empirical Methodology

## Data Set:

The dataset used in this analysis comes from Transportation Network Providers reporting to the City of Chicago as required by city ordinance. We are specifically using the **TransportationRideshare\_10percent\_2018.csv** from the Big Blue server as our starting point. Since this dataset contains over one million observations, we have selected a random sample of 30,000 observations to make the analysis more manageable.

The sample period for our data ranges from **November 1, 2018, to December 31, 2018.** As detailed in the variable name and description table at the beginning of this report, the dataset includes valuable information such as the trip's start and end times, trip duration (in seconds) and distance (in miles), the general locations of pickups and drop-offs, a breakdown of trip fares, tip amounts (for cases where a passenger left a tip), and whether the trip was solo or shared.

## Summary Statistics:

A screenshot of a computer screen

AI-generated content may be incorrect.

Here is a summary of the statistics for the numerical variables used in our analysis. Our dependent variable, Tip, has a low mean of $0.54, with the median tip being $0, indicating that most customers do not tip. This aligns with the University of Chicago study, which found that about 60% of customers do not tip, contributing to the low average tip amount. The average fare is $10.98, which indicates that the rideshare trips initially reserved were on average low in cost, where we also see the average total cost of a trip being $14.30. Additional costs average $2.79, and the average number of passengers in shared trips is essentially one, with a maximum of ten passengers. The average trip length is 5.9 miles, and the average duration is 1,069 seconds (about 18 minutes). The average pickup and drop-off locations are in Chicago's Loop (latitude 41.8929227, longitude -87.6657833 for pickup and latitude 41.8939952, longitude -87.6682528 for drop-off), which is not surprising given the area's popularity.

Trip\_Seconds has 6 missing values, which is a very small portion of the dataset. However, the latitude and longitude for pick-up and drop-off locations have substantial missing values. This is not an issue with the data, as locations outside of Chicago are intentionally left blank, as stated in the variable descriptions. These missing values are unlikely to significantly affect the results.

## Scatter Plots

A graph of blue dots

AI-generated content may be incorrect.

This scatter plot illustrates the relationship between the total distance of each rideshare trip and the tips given to the driver at the end of the trip. While a general trend is visible, showing that tips tend to increase as trip distance grows, most tips are relatively low, with most being under $10. Additionally, there is a notable group of passengers who traveled between 0 and 50 miles and did not leave any tip at all. Even an extreme case where a trip being 125 miles resulted in no tip. This initial analysis highlights a significant proportion of passengers who do not tip, prompting us to further investigate the factors that may explain why so many passengers choose not to tip.

A graph with blue dots

AI-generated content may be incorrect.

This scatter plot illustrates the relationship between the fare price agreed upon by the passenger and the tip left at the end of the rideshare trip. As seen in the previous plot a clear trend is shown, higher fares tend to be associated with higher tips. There are two extreme cases where the tips were significantly higher, one around $50 and the other just under $80.

A graph with numbers and dots

AI-generated content may be incorrect.

This scatter plot shows the relationship between tips and the date of the trip. Overall trips between November and December 2018 show a consistent distribution of tips. However, given the popularity of these months in Chicago due to Thanksgiving and Christmas holidays, we observe that trips during the week of Thanksgiving (around November 25) and Christmas (around December 23) often show tips above the typical range. This suggests that the date of the trip, particularly near holidays, may influence passengers to leave higher tips.

## Regression Methodology

With the multiple regression model, we are estimating the relationship between the **tip amount** (dependent variable) and several independent variables such as **fare, trip distance,** **trip duration, date of trip, and pick-up and drop off locations.** The goal is to understand how these variables influence the tip amount left by passengers. Our regression equation:

Tip=β0​+β1​(Fare)+β2​(Trip\_Miles)+β3​(Trip\_Seconds)+β4​(Additional\_Charges)+ β5(Drop\_Centroid\_Latitude)+ϵ

# Full Data Analysis

## Correlation Analysis

A table of numbers with text

AI-generated content may be incorrect.

Based on the correlation analysis, we find that several variables influence a customer’s tip amount:

* **Fare (0.306)**: There is a moderate positive correlation, suggesting that as fares increase, tips tend to increase as well.
* **Trip Miles (0.253)**: A weak positive correlation indicates that longer trips are somewhat associated with higher tips.
* **Trip Duration (0.204)**: A weak positive correlation with tips, suggesting that the amount of time spent on the trip has minimal impact on the tip amount.
* **Additional Charges (0.247)**: Extra charges (such as taxes) show a weak positive correlation with tips, indicating that higher additional charges may slightly influence higher tips.
* **Trips Pooled (-0.096)**: A very weak negative correlation, suggesting that shared trips may slightly decrease the tip amount, but the relationship is not significant.
* **Trip Total (0.449)**: A moderate positive correlation, indicating that higher total trip costs (fare + additional charges) are more strongly associated with higher tips.

Running the correlation analysis at the start of this project helps us identify key factors influencing tipping behavior and provides a foundation for selecting relevant variables for further analysis. This helps set expectations for what we might observe throughout our analysis.

## Proc Frequency Analysis

**The FREQ Procedure**

| **Shared\_Trip\_Authorized** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **false** | 21970 | 73.23 | 21970 | 73.23 |
| **true** | 8030 | 26.77 | 30000 | 100.00 |

Before conducting the Analysis of Variance (ANOVA) for the variables of Tip and Shared\_Trip\_Authorized, it's important to first examine the distribution of customers who authorized their trip information to be shared versus those who did not, so we have a general understanding. The table shows that about 73% of the trips in the dataset did not authorize sharing their trip while 27% did.

## ANOVA Analysis on Shared\_Trip\_Authorized

A screenshot of a graph

AI-generated content may be incorrect.



* **The null hypothesis (H0):** There is **no significant difference** in the mean tip between the two groups (those who authorized sharing their trip and those who did not).
* **The alternative hypothesis (HA):** There **is a significant difference** in the mean tip between the two groups (those who authorized sharing their trip and those who did not).

Since the p-value of <0.0001 is smaller than 0.05, we reject the null hypothesis and conclude that there is a statistically significant difference in the mean tip between the two groups: those who authorized sharing and those who did not.

The ANOVA sum of squares for the variable Shared\_Trip\_Authorized (1380.53569) represents the portion of the variance in tips explained by this factor. While this value is small compared to the total sum of squares (80156.46797), it indicates that trip-sharing authorization has some explanatory power regarding tipping behavior.

The R-squared value of 0.017223 (or 1.72%) shows that Shared\_Trip\_Authorized accounts for only a small portion of the variance in tips. While this percentage is low, the statistical significance suggests that trip-sharing authorization does have some influence on tipping behavior.

## Hierarchical Cluster Analysis Used to Determine Number of Clusters

A screenshot of a data table

AI-generated content may be incorrect.



Based on the cluster analysis, we found that the R-Square value improved significantly the most when it hit around 4 cluster groups. Since we are working with three variables, we chose to use the 'Ward's procedure' because this method provided better results based on the R-Square values versus the centroid method. This is because the ward procedure focuses on minimizing variance within clusters, leading to compact and well-separated groups.

### Pseudo T-Squared Line Plot

A graph of a number of clusters

AI-generated content may be incorrect.



As shown in the Pseudo T-Squared plot, there is an 'elbow' around 4 clusters. This suggests that 4 clusters might be an optimal choice, as the Pseudo T-Squared value drops sharply at this point.

### Dendrogram

A black and white drawing of a building

AI-generated content may be incorrect.

The dendrogram also supports the choice of using four clusters as it shows four distinct groups. This visualization effectively highlights the natural divisions within the data based on the three variables.

## Non-Hierarchical Cluster (K-Means) Analysis:

A table with numbers and text

AI-generated content may be incorrect.

The means of the clusters provide insights into the central tendencies of the three variables for each cluster:

* **Cluster 1:** Has an average fare of $68.00, a trip duration of 7,346.1 seconds, and an average trip distance of 50.24 miles.
* **Cluster 2:** Has an average fare of $19.66, a trip duration of 1,971.73 seconds, and an average trip distance of 12.73 miles.
* **Cluster 3:** Has an average fare of $37.61, a trip duration of 3,974.63 seconds, and an average trip distance of 23.76 miles.
* **Cluster 4:** Has an average fare of $7.55, a trip duration of 710.35 seconds, and an average trip distance of 3.34 miles.

A table with numbers and letters

AI-generated content may be incorrect.



The R-squared values in this table represent the proportion of variance in each variable explained by the clustering. Fare has an R-squared value of 0.4578 (45.78%), indicating that the clustering explains a moderate portion of the variance by fare. Trip\_Miles has an R-squared value of 0.5202 (52.02%), indicating that trip distance also plays a moderate role in defining the clusters. Trip\_Seconds has the highest R-squared value of 0.7621 (76.21%), indicating that trip duration is the most influential factor in distinguishing the clusters. Finally, the overall R-squared value of 0.7620 (76.20%) suggests that the clustering model explains a substantial portion of the variance in the dataset, capturing meaningful patterns in trip characteristics.

### K-Means Scatter Plots

A graph of different colored dots

AI-generated content may be incorrect.

This scatter plot reveals four distinct clusters of trips based on fare and trip duration:

* **First Cluster:** Trips with fares ranging from $30 to $150 and durations between approximately 6,000 and over 8,000 seconds.
* **Second Cluster:** Trips with fares between $0 and $100, lasting around 1,500 to 3,500 seconds.
* **Third Cluster:** Trips with fares from approximately $5 to over $300 and durations spanning 3,500 to 6,000 seconds.
* **Fourth Cluster:** Trips with fares between $0 and $50, with durations from 0 to just over 1,500 seconds.

The scatter plot suggests a linear relationship between trip fare and trip duration, which makes sense—longer trips generally result in higher fares.

A graph of different colored dots

AI-generated content may be incorrect.

This scatter plot reveals four distinct clusters of trips based on distance and trip duration:

* **First Cluster:** Trips with distances ranging from approximately 25 to just over 125 miles and durations from a little over 6,000 to more than 8,000 seconds.
* **Second Cluster:** Trips covering about 5 to 50 miles, with durations between just over 1,500 and 3,000 seconds.
* **Third Cluster:** Trips spanning about 5 to just under 100 miles, with durations ranging from around 3,500 to just over 6,000 seconds.
* **Fourth Cluster:** Trips with distances from 0 to just under 25 miles and durations from just over 0 to a little over 1,500 seconds.

The scatter plot shows a clear linear relationship between trip distance and duration, which is obvious—longer distances result in longer trip durations.

A graph of a trip distance

AI-generated content may be incorrect.

This scatter plot reveals four distinct clusters of trips based on fare and distance:

1. **First Cluster:** A small group of just two observations with the highest values, where trip fares range from $125 to $150 and trip distances span approximately 100 to over 125 miles.
2. **Second Cluster:** Overlapping with clusters three and four, this group includes trips with fares ranging from $0 to about $100 and distances between roughly 5 and 50 miles.
3. **Third Cluster:** Trips with fares from about $10 to over $200 and distances ranging from approximately 15 to just under 100 miles.
4. **Fourth Cluster:** Trips with fares between $0 and just under $50, covering distances from 0 to just under 25 miles.

The scatter plot highlights a clear linear relationship between trip fare and trip distance, which also makes sense—longer trips generally lead to higher fares.

## ANOVA Test on Tips based on Clusters

A screenshot of a table

AI-generated content may be incorrect.



* **The null hypothesis (H0):** The clusters **are** **not significantly** associated with customer tips, meaning the mean of tips are the same across all clusters.
* **The alternative hypothesis (HA):** The clusters **are significantly** associated with customer tips, meaning at least one cluster has a significantly different mean of tip compared to the others.

Since the p-value (<.0001) is much smaller than 0.05, we reject the null hypothesis and conclude that the clusters are significantly associated with customer tips.

The ANOVA sum of squares for Tip (2945.74822) represents the portion of variance in tips explained by the cluster groups. While this is relatively small compared to the total sum of squares (80156.46797), it indicates that the cluster groups explain some variability in tipping behavior.

The R-squared value (0.036750 or 3.68%) shows that clusters account for only a small fraction of the variance in tips. Although this percentage is low, the statistically significant F-value suggests that clustering still has some influence on tipping behavior. However, it also implies that other factors likely play a much larger role in explaining tip variability.

## Multiple Regression Model

A screenshot of a computer

AI-generated content may be incorrect.

The multiple regression model is used to predict the rideshare tipping behavior by analyzing variables such as trip distance, fare, trip time, and additional charges. The regression equation is:

Tip=β0​+β1​(Fare)+β2​(Trip\_Miles)+β3​(Trip\_Seconds)+β4​(Additional\_Charges)+ β5(Drop\_Centroid\_Latitude)+ϵ

Key independent variables from the regression results:

* **Fare (β1​)**: The coefficient for the fare is statistically significant and positive which suggest that as the fares are higher, the tip will also be higher. This is consistent with the correlation analysis, which indicated the positive correlation between tipping and fare price.
* **Trip Distance (β2​)**: There is also a positive impact with the coefficient for trip distance. It supports the assumption that the riders who are travelling for longer trips are more likely to pay more tips. However, the impact scale is less compared to the fare.
* **Trip Duration (β3​)**: There is a weaker coefficient for duration of trip than for fare and trip distance. It shows that the amount of time spent during the trip has very little impact on tipping.
* **Additional Charges (β4​)**: This variable is showing a positive effect. It suggests that trips which are having additional charges could impact higher tipping but not as strong as fare.
* **Drop\_Centroid\_Latitude (β5)**: This variable is also showing a positive effect with tips. It’s interesting that this variable has a significant impact on tipping, but it does make sense. The location of the drop off could affect the customer’s willingness to tip, such as cases where the drop off is at Chicago landmarks or busy areas.

The specified factors in the multiple regression model explain a moderate amount of tipping variability according to the R-squared measurement of .08168 (8.17%). The measured factors explain only a small percentage of tipping behavior according to the low R-squared value although other hidden factors, such as client demographics and driver contact and payment method, might significantly influence tipping.

## Stepwise Regression Model

A screenshot of a computer program

AI-generated content may be incorrect.

The stepwise regression model improves the multiple regression model by only picking variables that are most significant by their p-value.

* As per the result, the most important variables are the trip distance and the fare price.
* Trip duration was eliminated in certain iterations since it had very less impact on tips.
* Overall, the model fit was slightly less favorable in compared to multiple regression model. It suggests that eliminating less important variables is not resulting into better prediction.

The principal determinants of tipping behavior were fare and trip distance according to both analytical models. The stepwise regression method showed that removing certain variables would not decrease the predictive power of the model although those variables contributed less to tipping behavior.

### Performance of the Regression Models

A number and numbers on a white background

AI-generated content may be incorrect.

Based on the performance results from the test data, Model 1 (multiple regression model) performed better than the stepwise model – even though the difference in performance is very small. Model 1 did have a slightly better R-Square value meaning it is slightly a better fit of the data. In addition, Model 1 included all variables that could potentially influence tipping, which allows for a better understanding of how each factor affects tips, even though some of the predictors may not be statistically significant. This makes the multiple regression model more useful for interpreting how different variables numerically influence tipping, as it considers the full set of variables, rather than eliminating some through the stepwise process.

## Logistic Model

The logistic regression model was implemented to figure out binary outcome: if an individual will tip or not. The ROC curve was used for evaluating the performance of the model.

### ROC Plot

A graph of a logistic

AI-generated content may be incorrect.

As the result, the area under the curve score was moderate. It suggests that the model predictive power but is not 100% precise in separating tippers from non-trippers. The logistic model has an AUC of 0.65493, which indicates that it has a good predictive ability. As seen from the graph, the AUC is higher than that of Random Forest model’s AUC, the Logistic model performs better. This also indicates that the Logistic model is more effective at distinguishing tippers from non-tippers. Being a simpler model, it also provides a better classification of the performance than Random Forest. Logistic model is more interpretable which makes easier to understand key tipping factors. Random Forest captures complex interpretations but does not improve the accuracy. Overall, the Logistic regression is preferable model due to its better AUC and interpretability.

### Performance of Model

\*Threshold set at 30% for the predicted probabilities

A close-up of numbers

AI-generated content may be incorrect.

**Confusion Matrix Results:**

* **True Negative (TN):** 6815
* **False Positive (FP):** 535
* **False Negative (FN):** 1367
* **True Positive (TP):** 283

**Performance Rates:**

* **Accuracy Rate:**
* **Error Rate:**
* **True Positive Rate:**
* **Fale Positive Rate:**

The threshold for estimated probability had been set to 30%. It means that if the ride had a higher than 30% chance of tipping than it could be classified as a tipper. Analysis showed a reliable accuracy measure however there were many incorrect positives indicating consumers who wished to tip failed to do so.

* In 78.87% of cases the model shows accurate prediction abilities for choosing to tip or not. The categorization performance demonstrates high strength in this analysis.
* The prediction errors demonstrate that the model achieves a 21.13% failure rate in its output assessments. The existing accuracy level indicates space for better discrimination between people who tip and those who do not.
* The detection rate for correct tippers reaches only 17.15% among all detected traffic violations. Numerous tippers go undetected because the model exhibits low sensitivity rates.
* During incorrect analysis the model classifies 7.28% of users who did not tip as tipping individuals. The model shows moderate inaccuracies in its ability to identify the correct tipping habits.

## Random Forest Model

### Best Number of “mtry”

A table of numbers and symbols

AI-generated content may be incorrect.



Based on these results, the optimal value for 'mtry' is 2 for our random forest model. When we ran the code several times, it went between 1 or 2 being the optimal number. After running the random forest for both, we chose to use 2 mtrys.

### Best Number of “ntrees”

A graph with red lines

AI-generated content may be incorrect.

Based on the results, the optimal number of “ntrees” was around 250.

### ROC Plot

A graph of a function

AI-generated content may be incorrect.

The random forest model has an AUC of 0.63318, which indicates that it has moderate predictive ability. Although the AUC of this model is not significantly different from the logistic model, it is still lower. In result, the logistic model outperformed the random forest model.

### Performance of the Model

\*Threshold set at 30% for the predicted probabilities

A number with numbers and letters

AI-generated content may be incorrect.

**Confusion Matrix Results:**

* **True Negative (TN):** 6950
* **False Positive (FP):** 400
* **False Negative (FN):** 1447
* **True Positive (TP):** 203

**Performance Rates:**

* **Accuracy Rate:**
* **Error Rate:**
* **True Positive Rate:**
* **Fale Positive Rate:**

The model Accuracy rate is 79.50% which means it correctly predicts tipping behavior in most of the cases but however the error rate is 20.52%, which shows some misclassifications. The true positive rate is only 12.30% showing that the model struggles to correctly identify actual tippers. Whereas the false positive rate is 5.44% which means a small portion of non-tippers were incorrectly categorized as tippers. The overall model performs well, its low recall showcases it is biased towards predicting non-tipping behavior, underestimating the likelihood of tippers.

## Neural Network Analysis

### Model 1 (One hidden Layer and Two Neurons)

A diagram of a network

AI-generated content may be incorrect.

A graph of a function

AI-generated content may be incorrect.

After running the neural network analysis with one hidden layer and two neurons, our AUC in the ROC plot came out to 0.63274.

### Model 2 (Two hidden Layer and Three Neurons)

A diagram of a network

AI-generated content may be incorrect.

A graph of a function

AI-generated content may be incorrect.

After running the neural network analysis with two hidden layers and three neurons, our AUC in the ROC plot came out to 0.62779. When comparing both neural network analyses, the first neural network was the better performing model. However, when comparing all three models (logistic, random forest and neural network), the logistic model remains the better performing model.

### Performance of Neural Network Model 1

\*Threshold set at 30% for the predicted probabilities

A number and numbers on a white background

AI-generated content may be incorrect.

**Confusion Matrix Results:**

* **True Negative (TN):** 6800
* **False Positive (FP):** 550
* **False Negative (FN):** 1345
* **True Positive (TP):** 305

**Performance Rates:**

* **Accuracy Rate:**
* **Error Rate:**
* **True Positive Rate:**
* **Fale Positive Rate:**

The neural network model overall is accurate, with an accuracy rate of 78.94%. However, it struggled to identify those who tip, as shown by the low True Positive Rate of 18.48%, meaning many tippers are misclassified as non-tippers. On the other hand, the model performs very well at identifying non-tippers, with a low False Positive Rate of 7.48%, indicating that it rarely misclassifies non-tippers as tippers.

# 

# Summary of Project

After running the full data analysis of our random sample data, we were successful in completing our data analysis method plan and mostly successful in answering our research questions. We were able to use SAS to have an overview of our data and see what variables are likely to be important factors in our analysis. We were able to identify groups of rideshare trips within our data based on the trips fare, total distance and total duration through cluster analysis. We also were able to see how our cluster groups were statistically associated with our dependent variable, Tip.

We were able to identify five variables in our data set that had a significant impact on tipping numerically. We also identified how much of the variance in tipping could be explained by the variables but found that it was very little with about 8%. In addition, we conducted logistic, random forest and neural network analyses to identify binary outcomes: 0 passenger left no tip, and 1 passenger left a tip. We ran the ROC plots for all three to compare which model performed the best based on its AUC as well as ran confusion tables for each model to see how well they predicted in those who tipped or not with the threshold set at 30%. While all models were not significantly different from each other we concluded that the logistic model was the better performing model due to it having the highest AUC. We also concluded that our models were overall accurate in making predictions and while they were very successful in identifying non-tippers, they were unsuccessful in identifying those who would tip.

The shortcoming of our project was that we were not able to successfully predict those who would tip in our models and when trying to run the R codes, we had to make a lot of adjustments by removing missing values in some variables and replacing some of the missing values with the mode and mean so that the code can run. Some points of improvement would be having a data set with fewer missing values – so rideshare trips that were conducted only in Chicago so that there would be no missing values in the pickup/drop off coordinates and the Chicago community areas. In addition, if we had more time, using a much bigger sample data to see if there is a significant difference in results when we have more observations as the full data set from the City of Chicago has around ten million observations.

# Bibliography

Data Set Link: <https://bigblue.depaul.edu/jlee141/econdata/Chicago_RideShare/TransportatonRideshare_10percent_2018.csv>

News Articles Quoted in Motivation Section:

* <https://news.uchicago.edu/story/nearly-two-thirds-uber-riders-never-tip-study-finds>
* <https://www.huffpost.com/entry/tipping-rideshare-drivers_l_6622b856e4b0868a1b90fed8>

Data Source Link: <https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips-2018-2022-/m6dm-c72p/about_data>

* Information for Variable list came from this link.

Disclaimer on Data Set from Source Link:

“All trips, from November 2018 to December 2022, reported by Transportation Network Providers (sometimes called rideshare companies) to the City of Chicago as part of routine reporting required by ordinance.    
Census Tracts are suppressed in some cases, and times are rounded to the nearest 15 minutes. Fares are rounded to the nearest $2.50 and tips are rounded to the nearest $1.00.”

<https://www.latlong.net>

* Used to locate the pickup and drop off locations based on their latitude and longitude coordinates.

# Appendix

Please note that there are three separate code files for our project. There is a SAS txt file of our code done for the summary statistics, cluster analysis and ANOVA tests. There is a R txt file labeled as, Final Project Part 1, includes the code for the regression, logistic and random forest analyses. The other R txt file labeled as, Final Project Part 2, is the code for the neural network analysis. They were separated because part 1 of the R code was done on Liana’s R Studio desktop app and part 2 was done in the R Studio Pro server accessed via Big Blue.