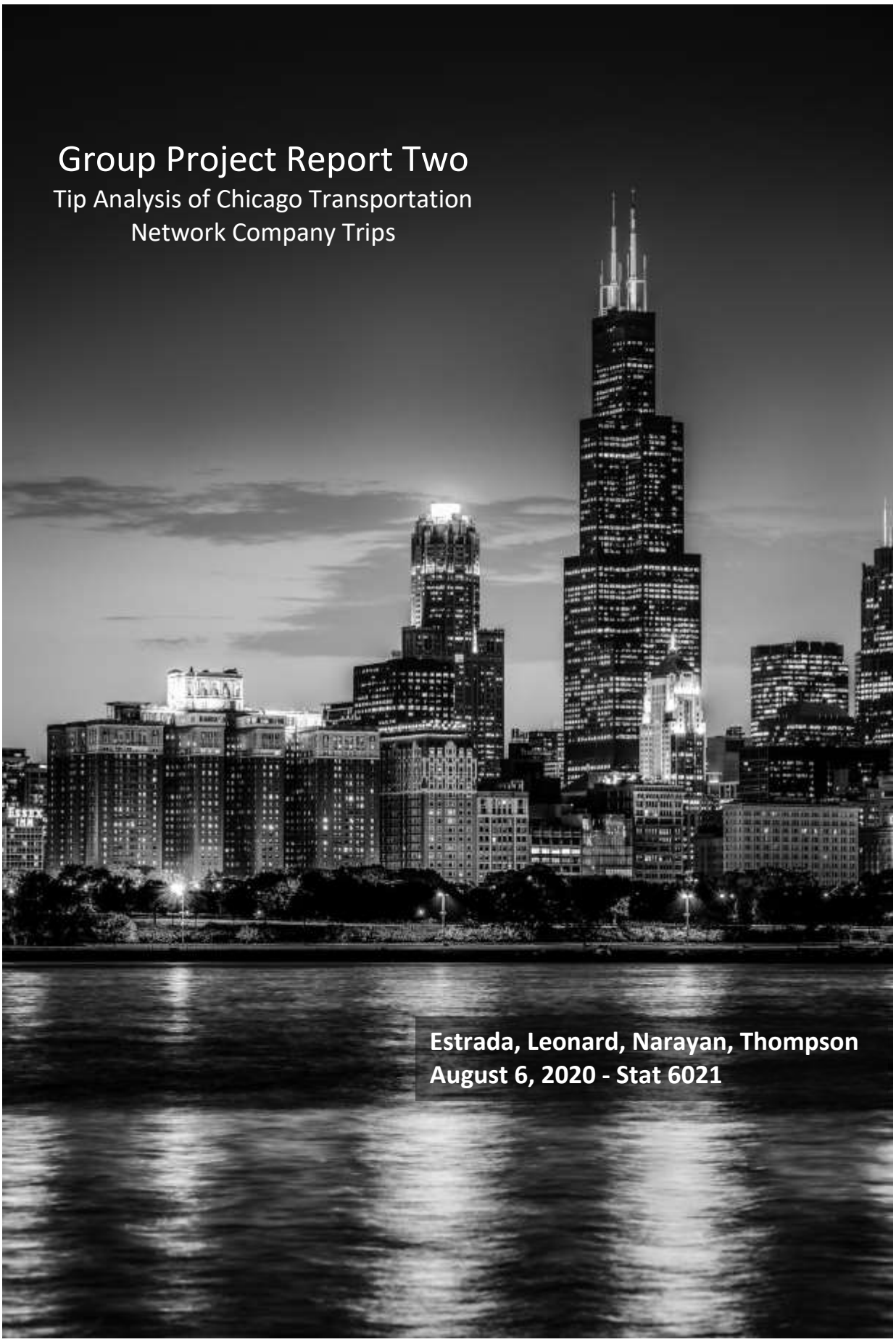
# Group Project Report Two

## Tip Analysis of Chicago Transportation Network Company Trips



**Estrada, Leonard, Narayan, Thompson**

**August 6, 2020 - Stat 6021**

## Introduction

We conducted statistical analysis using the techniques we covered in STAT 6021 course for the City of Chicago's Transportation Networking Companies (TNC) "Trips" dataset[i]. We believe that this dataset sufficiently challenged our statistical abilities while falling within the scope of this class. The dataset is comprised of anonymized data from all trips by Uber, Lyft, and other TNCs in Chicago beginning November 1, 2018. The dataset is updated monthly and has data on more than 152 million trips. Due to the large size of the dataset, a subset of the data was chosen for the analysis. The criteria chosen to narrow the scope of our subject dataset is described in detail in the Exploratory Data Analysis section. Our goal in this analysis was to determine what predictors contribute to a user's decision of whether to tip or not. The response variable was classified as a binary response.

Usage frequency of Chicago's Transportation Networking Companies for short trips, equal to or under 20 miles, was high and provided enough data to conduct a non-biased or skewed investigation into what factors could contribute to the rider's decision of whether to tip or not.

The data obtained from Chicago's transportation database was already anonymized and organized to a great extent. The available predictors selected for our analysis were empirical in nature and include trip length in seconds, trip length in miles, fare, tip amount, and additional charges.

## Summary *(Lay Audience)*

We created a success / failure model that could predict whether a driver should expect a tip from their rider or not. Trip length in miles, trip length in seconds, fare, additional charges,

were considered as variables that could predict if a rider would tip. Each entry in the dataset corresponded to an individual trip.

In order to conduct our analysis, we employed a variety of techniques to assist in determining which variables would contribute to predicting if a rider would tip. The models displayed visual or numerical indicators which were then interpreted using the discriminating criteria learned throughout the semester. Graphical and tabular outputs produced by a general linear regression model were used to explore the relationship between the predictor variables(s) and the response variable.

We created a model that is able to predict the likelihood a driver will receive a tip. Our model is 8% more accurate than random guessing. Our model as shown below

**Log(Tip_Odds) = -2.181 + 0.120*Additional.Charges + 0.023*Fare**

is based on the fare and additional charges (toll, airport fee, etc.) paid by the rider. Tip_Odds is the odds ratio for a rider tipping or not tipping.

In order to generate a better model, we need other additional variables which were not available to us through this dataset. Additional variables that may contribute to building a better model include demographic and socioeconomic data about the riders, which is not in the data provided to the City of Chicago by the various TNCs. Ridesharing services like Uber and Lyft collect information about their drivers such as ratings, car information, and amenities provided. It would be interesting to explore how the likelihood of a rider tipping is related to driver data.

Based on a 2019 study[ii] conducted by the Nation Bureau of Economic Research of 2017 TNC data, it was found that 60%[iii] of riders don't tip for Uber and Lyft services. Our dataset showed 84.3% of riders did not tip for TNC services in the City of Chicago in January of 2019.

## Detailed Description

***Data Scope and Cleaning***

In order to select a manageable and relevant dataset, we chose limiting criteria for our predictors. We selected the data for all rides with a trip length of less than 20 miles that occurred in January 2019.

The location data was anonymized such that we could not associate trip start and end points in a meaningful way. Therefore, we removed all predictors associated with location from our data subset.

The trip total was removed from the dataset because it is the sum of fare, tip, and additional charges. Considering trip total as a predictor would have produced a poorly performing model due to multicollinearity.

Along with the location data, we also removed the date and time related fields. Length of trip was already captured in another predictor and we weren't interested in determining the effect of time of day on the likelihood of tipping. Therefore, we excluded the data columns associated with trip start and end time stamps.

Pooled rides are when different riders share a single trip together and have different origin and destination points. It's a cheaper fare than an individual ride. Trips.Pooled was a variable representing the count of how many pooled rides were strung together before the car was empty again. We did not consider this variable to be appropriate for our analysis goals and removed it from the dataset.

After cleaning the dataset the predictors left in our data subset were: trip length in seconds (Trip.Seconds), trip length in miles (Trip.Miles), fare (Fare), tip (Tip), additional charges (Additional.Charges). Additional.Charges is the sum of any taxes, fees (such as airport pickup), and other charges (such as waiting time) that are included in the trip total.

To construct a binary logistic regression, we transformed our tip predictor from a numeric variable to a binary "No Tip" or "Tip" categorical response variable (Figure 1). We were interested in predicting whether a rider would tip at all and not how much they tipped. Due to societal tipping norms, we'd expect the tip amount fall within 10-20% of the total fare. Therefore, the amount of tip would have been dependent on the amount of the fare and necessitated consideration of interaction effects.

```
> str(udata)
'data.frame':   8510906 obs. of  5 variables:
 $ Trip.Seconds      : num  1652 1661 488 1848 700 ...
 $ Trip.Miles        : num  3.5 11.9 2.1 7.3 2.7 2 6.7 6.8 11.9 1.8 ...
 $ Fare              : num  7.5 15 5 7.5 7.5 5 7.5 12.5 47.5 5 ...
 $ Additional.Charges: num  2.89 2.55 2.55 2.55 2.55 2.55 2.55 2.55 7.8 2.55 ...
 $ Tipped.or.Not     : Factor w/ 2 levels "No Tip","Tip": 1 1 2 1 1 1 1 1 1 1 ...
```

*Figure 1: Data Structure.*

The data set was searched for missing data. Fourteen rows were found to be missing values. Due to the size of the dataset, over 8.5 million observations, the 14 rows were considered an insignificant loss that would not affect the results and therefore removed from the data set.

***Data Exploration***

Correlation maps and box plots were generated to provide visual representations of the relationship between the variables. The correlation, ranging from negative to positive one,

indicates the strength of the relationship between two predictors. Values trending towards positive one indicate an increasingly strong positive linear relationship between variables. Values trending toward negative one indicate an increasingly strong negative linear relationship between variables. A zero value indicates the variables are uncorrelated and there is no linear relationship. We defined strong correlation as values that fall between -1.0 to -0.5 and 0.5 to 1.0.

The correlation matrix indicated that none of the predictors are significantly correlated with tip or no tip (Figure 2). There was a significant correlation between fare and seconds, miles. This stands to reason as seconds and miles are typically factors in calculating the fare amount. Seconds and miles were highly correlated as higher mileage trips are longer in duration due to average velocities on city streets. Additional charges were also mildly correlated to Trip Miles. This could be due to a subset of high mileage trip riders travelling to a location that includes an additional charge such as the airport. We did not further investigate the correlation between additional charges and miles.

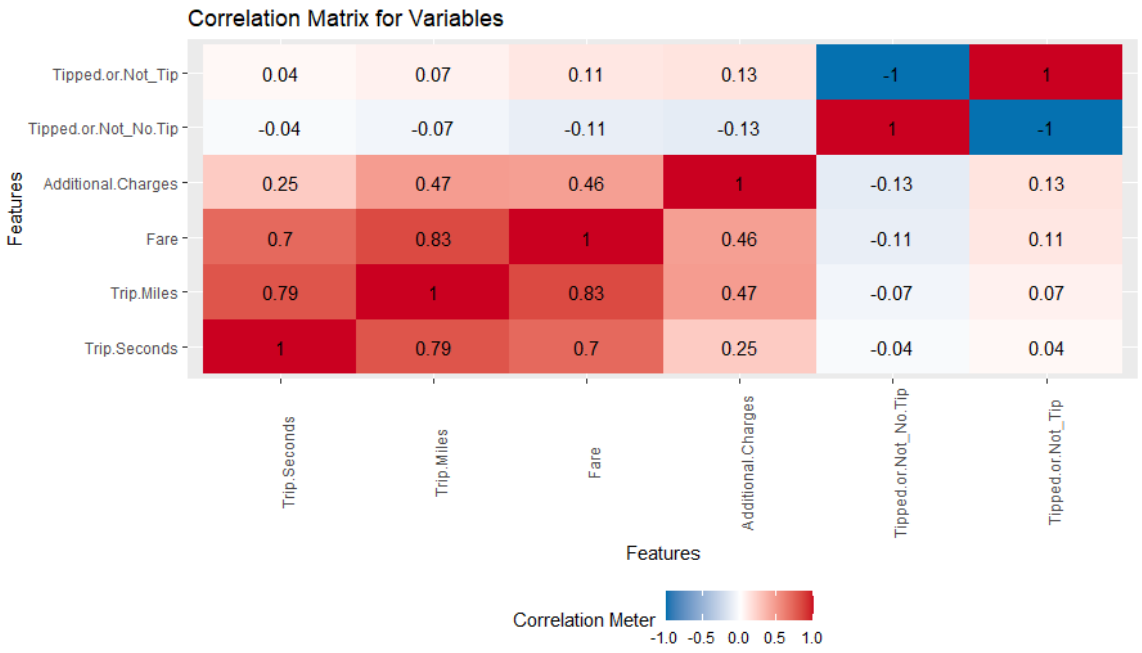



*Figure 2: Correlation Matrix of Variables in "Trips" Dataset*

Five percent of the selected dataset was separated into a training dataset consisting of more than 425,000 observations and the remaining 95% of the data set aside in a testing dataset. The training dataset was used for model development. All exploration, model development, and graphics reported after this point were produced from the training dataset.

Utilizing the training dataset, box plots were generated for each predictor against our tip or no tip categorical response variable to examine the variability of the response variable to the remaining predictors. The box plots did not provide any specific insights about the variability of the response variable (Figure 2).

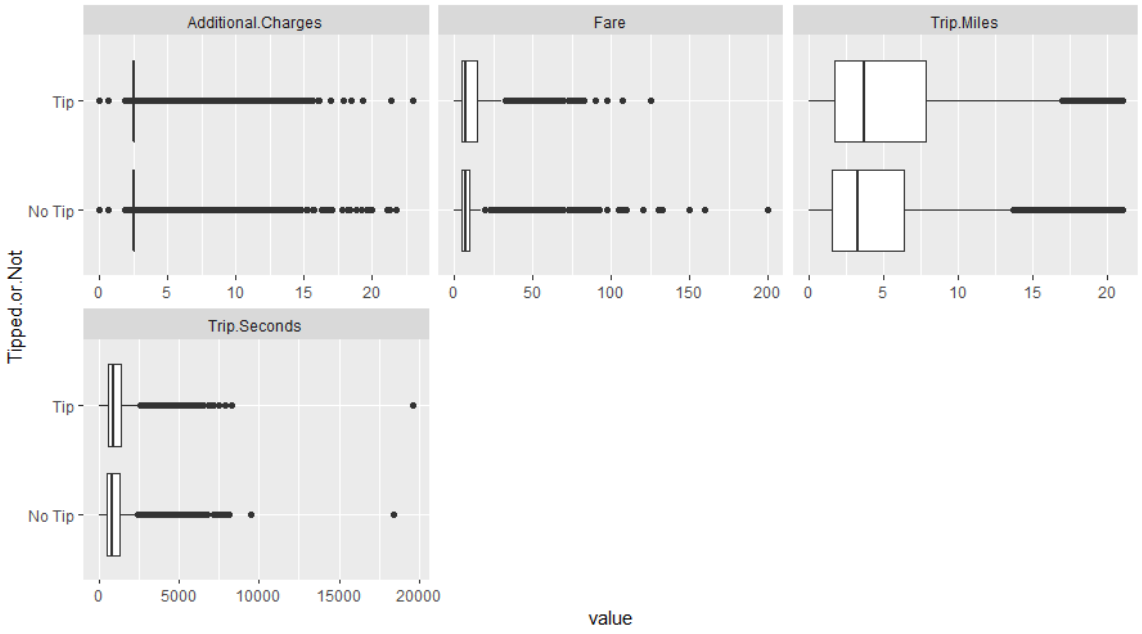



*Figure 3: Box plots of Tip or no tip response vs. various predictors*

After considering the inconclusive box plots we decided to fit a model for each predictor. Due to the small number of considered predictors, we thought it appropriate to fit all of them. These values indicated how much the response variable changed from a one-unit change in the predictor(s).

To compare the models the Akaike Information Criterion (AIC) was calculated. The AICs of the models are presented in Table 1. Tip_Odds is the odds ratio for a rider tipping or not tipping. The AIC measures the model's relative quality due to its similarity to $R^2$adj in accounting for the effects of additional predictors[iv]. Model 4, shown in the table below, had the lowest AIC. We wanted to investigate whether the inclusion of additional predictors would improve the model.

*Table 1: Response vs. Predictor AIC values of models initially considered.*

| | Response | Predictor(s) | AIC |
|---|---|---|---|
| Model 1 | Tip_Odds | Trip.Miles | 381408 |
| Model 2 | Tip_Odds | Trip.Seconds | 382659 |
| Model 3 | Tip_Odds | Fare | 378432 |
| Model 4 | Tip_Odds | Additional.Charges | 377277 |
| Model 5 | Tip_Odds | Additional.Charges+Fare | 376019 |

Two models were chosen for additional consideration, model 4 and model 5. These two models had the lowest AIC values.

```
Call:
glm(formula = Tipped.or.Not ~ Additional.Charges, family = "binomi
al",
    data = train.udata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8586  -0.5905  -0.5905  -0.4849   2.0971

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.081414   0.007501 -277.50   <2e-16 ***
Additional.Charges  0.165881   0.002091   79.32   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 383146  on 425544  degrees of freedom
Residual deviance: 377273  on 425543  degrees of freedom
AIC: 377277

Number of Fisher Scoring iterations: 4
```

```
Call:
glm(formula = Tipped.or.Not ~ Additional.Charges + Fare, family =
"binomial",
    data = train.udata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3307  -0.5940  -0.5636  -0.5293   2.1050

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.1805293  0.0080040 -272.43   <2e-16 ***
Additional.Charges  0.1201787  0.0024292   49.47   <2e-16 ***
Fare                0.0228790  0.0006363   35.96   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 383146  on 425544  degrees of freedom
Residual deviance: 376013  on 425542  degrees of freedom
AIC: 376019

Number of Fisher Scoring iterations: 4
```

*Figure 4: Model 4 and 5 Regression summary tables.*

The summary equations for these models are shown in Table 2.

*Table 2: Summary Table of Model Equations.*

| Model Number | Equation |
|---|---|
| Model 4 | Log(Tip_Odds) = -2.081 + 0.166*Additional.Charges |
| Model 5 | Log(Tip_Odds) = -2.181 + 0.120*Additional.Charges + 0.023*Fare |

From inspection of the AIC we considered model 5 to be the better model. To confirm we did not overfit the model, we conducted likelihood ratio tests. We tested if all the predictors are significant or not for model 5 and we tested if model 5 is better than model 4. The p-value was less than 0.05 for both the tests. The likelihood ratio tests compare the deviance between the two models and evaluate if the presence of additional predictors improve the model fit

enough to justify making the model more complex. As shown in Figure 5, we can see that it the predictors are significant and adding fare to additional charges model is slightly better. Therefore, we chose model 5, with predictors additional charges and fare, as our model of interest.

```
> ##Model with Additional Charges - checking deltaGsquared to see if
 it is better than a model without any predictors
> #h0:beta1=0; h1:atleast one not zero
> #p-value =0; Therefore this model is useful
> 1-pchisq(mod_train_adch$null.deviance-mod_train_adch$deviance,1) #0
[1] 0
> #Which model is better?
> #h0:beta2=0; h1:beta2 not equal to zero
> #p-value =0; Therefore we can go with Fare and Additonal Charges
> 1-pchisq(mod_train_adch$deviance-mod_train_fare_adch$deviance,1)#0
[1] 0
```

*Figure 5: R Output of Model Hypothesis Test*

For a unit increase of additional charge, the estimated log odds of getting tipped increases by exp(0.120)=1.127, while holding the fare constant. For a unit increase of fare, the estimated odds of getting tipped increases by exp(0.023)=1.02, while holding the additional charges constant.

To further validate our model, we generated a Receiver Operator Characteristic (ROC) curve and calculated the Area under the ROC curve (AUC). ROC and AUC are important evaluation metrics for calculating the performance of any classification model. We generated ROCs for both the models 4 and 5 (Figure 6). From inspection it appears that model 5 is a better model than model 4. Since the model curves are in the top-left side of the diagonal line, our models perform better than “random guessing” or a “useless classifier”.

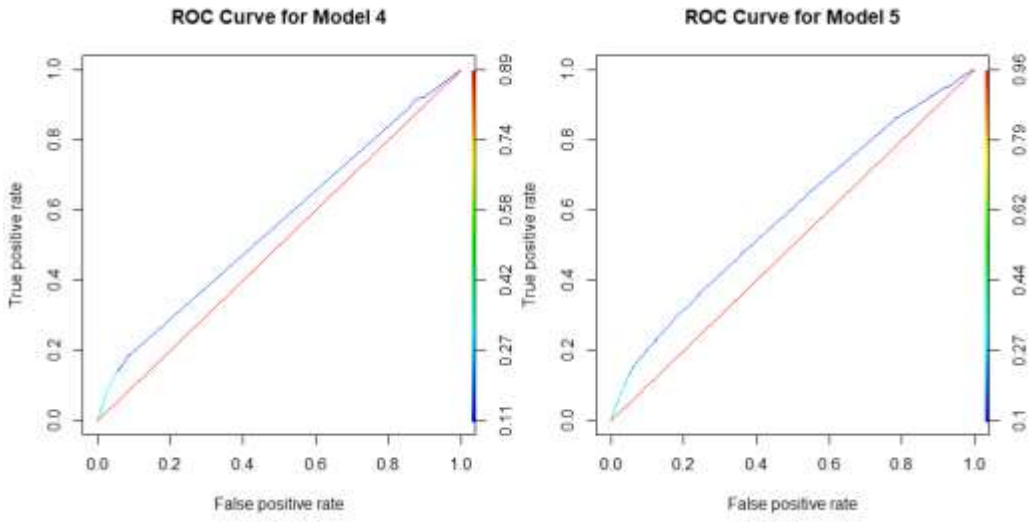



*Figure 6: Model 4 and 5 ROC curves.*

After the ROC curves were generated the area under each curve (AUC) was calculated for comparison of the predictive power of models 4 and 5. The results are presented in Table 3. The AUC of random guessing is defined to be 50%. The calculated AUC for model 4 was 55.9% and model 5 was 58.7%. From the results, our model is improved compared to random guessing.

*Table 3: Model 4 and 5 AUC Values*

| ROC Curve | AUC |
|---|---|
| Model 4 | 55.89% |
| Model 5 | 58.7% |

Confusion matrices were generated in R for evaluation of model performance to supplement the ROC curves. After discussion with our local expert and review of online resources[v] we decided to use a lower threshold value of 0.3 to increase the sensitivity of our model and avoid the situation of an unbalanced prediction. As we are trying to predict if the driver will get tipped or not, it is equally important to predict whenever a driver gets tipped and

doesn't get tipped.  Our main goal is to predict when drivers are getting tipped. Therefore, in order to reduce the number of false negatives we chose to reduce the threshold value to 0.3. We understand that by reducing the threshold we increase the number of false positives as well. However, we consider this tradeoff to be acceptable due to the low AUC of our models. The confusion matrix for our models are presented as four-fold plots in the figure below.

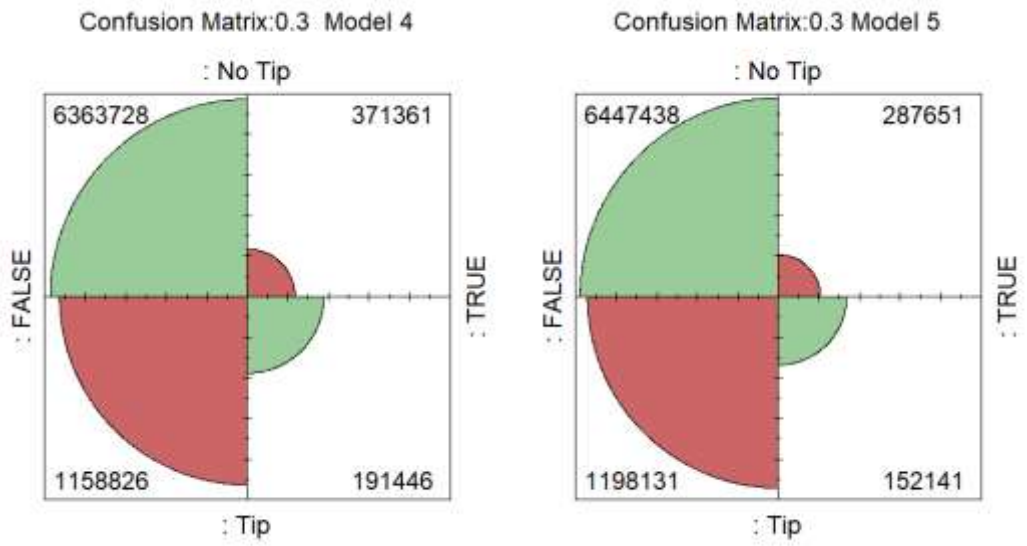



*Figure 7: Model 4 and 5 0.3 Threshold Confusion Matrix as a fourfold plot.*

Our test dataset included 8,079,561 trips of which model 5 predicts 152,141 trips result in a tip and 6,447,438 trips result in no tip. The accuracy of our model is (True Positive + True Negative)/Total which is 81.62%. The sensitivity (True Positive / (False Negative + True Positive)), meaning our model predicted that a driver would receive a tip and in reality they did, was 11.2%. The specificity (True Negative / (True Negative + False Positive)), meaning our model predicted that a driver would not receive a tip and in reality they did not, was 99.9%. Due to the preponderance of no tip samples in our data set, we expect the specificity to be a large value. In the same manner, we accept the low sensitivity based on the distribution of the dataset.

The after consultation with our local expert we chose not to conduct a “goodness of fit” test. In practice, the ROC and AUC are considered sufficient for model analysis as the resulting plots indicated that the model was improved compared to random guessing. To further validate, we used test data and predicted the probability of getting tipped. The training model calculated a 29% probability of getting tipped for a fare of $27 with an additional charge of $6. With the same criteria, our model predicted a 30% probability of getting tipped. Therefore, our model is somewhat better than random guessing.

Future work could include, considering additional predictors or further limiting the scope of the dataset. Given the relatively low AUC of the model selected, there is a possibility that a better predictor is not in our dataset. A better predictor may be in demographic and socioeconomic data about the riders which is not in the data provided to the City of Chicago by the various NTCs. Review of the distribution of the datasets in a histogram (Figure 8), indicates that the majority of the rides in the City of Chicago are less than 10 miles. Other than during “surge” fares, when additional fees are applied due to high demand for rides, the vast majority of the fares are under $50. By further restricting these predictors, the utility of the model may be increased within those constraints due to the removal of non-typical rides. These items can be seen in the log tail of these predictors in data histogram presented in the figure below. There is an opportunity for further refinement of the threshold value though an automated search for improved model performance however it is recommended that the potential predictors be reviewed before developing the automated search.

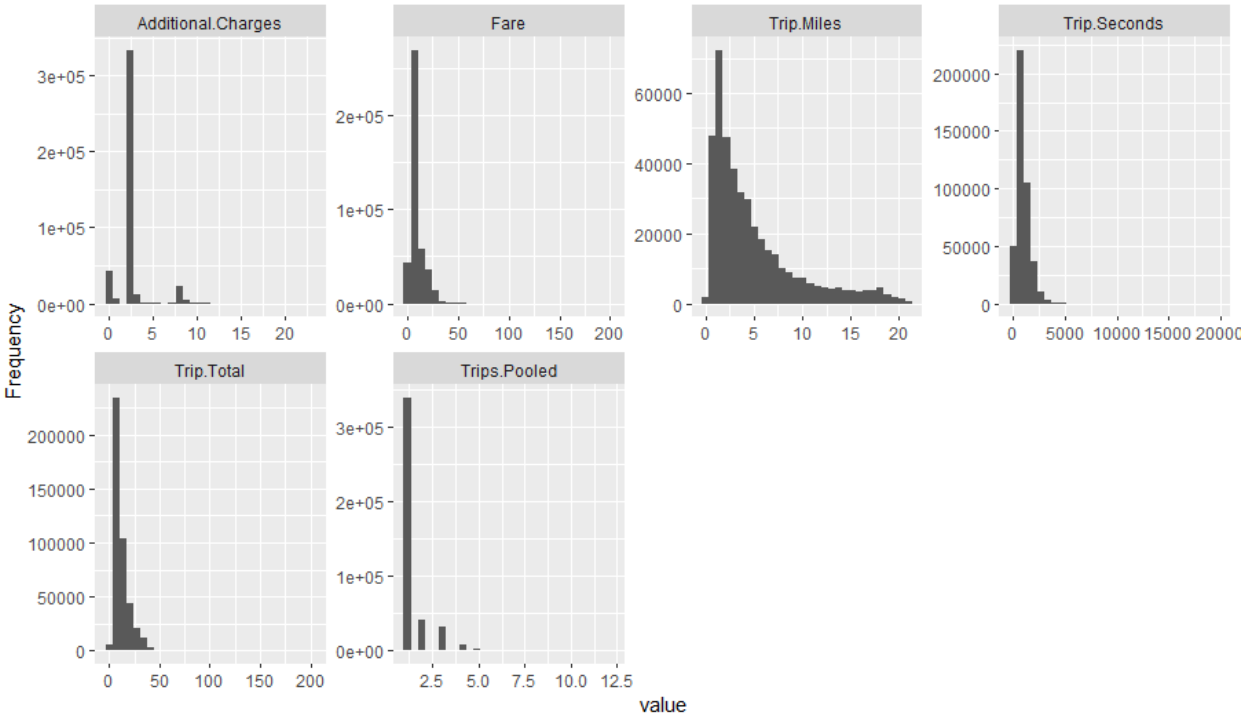



*Figure 8: Histogram of predictors in "Trips" dataset.*

Based upon our statistical analysis, we conclude that model 5 does have use in predicting whether a driver can expect a tip from their rider for rides under 20 miles in the City of Chicago. However, we believe that it can be further improved by researching and adding other significant predictors.

## References


[i] Data Source - https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p

[ii] Nationwide Tipping Field Experiment - https://www.nber.org/papers/w26380.pdf

[iii] Nearly two-thirds of Uber customers don't tip their driver - https://www.theverge.com/2019/10/21/20925109/uber-tipping-riders-drivers-percentage-gender-nber-study

[iv] Simplifying the ROC and AUC metrics. - https://towardsdatascience.com/understanding-the-roc-and-auc-curves-a05b68550b69

[v] A Guide to Machine Learning in R for Beginners: Logistic Regression - https://medium.com/analytics-vidhya/a-guide-to-machine-learning-in-r-for-beginners-part-5-4c00f2366b90