Modelling Tipping Behavior in Urban Transportation

Final Report

# Group Members

* Archanakumaree Dineshbhai Bhusara - abhusara -50560481
* Nihar Madasu - niharmad – 50559501
* Navaneethakannan Arumugam - arumuga3 – 50560498
* Yiyuan Wang – yiyuanwa – 50311193

# Introduction:

Our project aims to build a comprehensive model to analyze tipping behavior in urban transportation systems, more specifically the New York City Taxi system. Encompassing variables such as trip details, fare amounts, tipping amounts, weather conditions, demographic attributes, and socioeconomic indicators, the goal is to develop a predictive model capable of analyzing tipping behavior. This involves employing advanced statistical and machine learning techniques to identify correlations and patterns between the variables and the tip amounts paid by customers. The objective is to gain insights into the factors influencing tipping behavior in urban transportation systems, facilitating a deeper understanding of consumer behavior. Using variables employed across various research experiments involving the same topic, we plan to find the top predictors that influence tipping behavior.

# Literature Survey

1. Tang (2019) - Big Data Analytics of Taxi Operations in New York City:

Tang employs big data analytics to investigate taxi operations in New York City, leveraging vast datasets to uncover insights into various aspects of service provision and customer satisfaction. Methodologically, Tang's study involves the collection and analysis of large-scale transactional data from taxi services, encompassing variables such as trip duration, route optimization, and passenger feedback. This study helps our research by pulling us in the right direction in terms of variable selection and EDA.

1. Elliot et al. (2017) - Tippers and Stiffer: An Analysis of Tipping Behavior in Taxi Trips:

Elliot and colleagues conducted a nuanced analysis of tipping behavior in taxi trips, distinguishing between different types of passengers based on their propensity to tip. The study employs a mixed- method approach, combining surveys, interviews, and observational data to capture the multifaceted nature of tipping decisions.

1. Correa (2017) - Exploring the Taxi and Uber Demands in New York City: An Empirical Analysis and Spatial Modeling:

Correa's research adopts an empirical analysis and spatial modelling approach to examine the demands for both traditional taxi services and ride-hailing platforms in New York City. It provides us with significant insight into using spatial analysis in our research as well.

1. Judd Cramer and Alan B. Krueger - Disruptive Change in the Taxi Business: The Case of Uber:

Cramer and Krueger conducted a comparative study to analyze the disruptive impact of ride-hailing platforms like Uber on the traditional taxi business. The quantitative analysis of market data to examine shifts in consumer behavior provides us with guidance towards using another predictor in our model.

1. Riascos and Mateos (2020) - Networks and Long-range Mobility in Cities: A Study of More Than One Billion Taxi Trips in New York City:

Riascos and Mateos employ network analysis techniques to study long-range mobility patterns in urban areas using taxi trip data. This information provides us with an analysis of mobility which provides us with another possible predictor of customer tipping behavior.

1. Xie et al. (2021) - Revealing Spatiotemporal Travel Demand and Community Structure Characteristics with Taxi Trip Data: A Case Study of New York City:

Xie and colleagues conducted a case study utilizing taxi trip data to reveal spatiotemporal travel demand patterns and community structure characteristics in New York City. Methodologically, the study involves the collection and analysis of detailed trip data from taxi services, encompassing variables such as trip origin-destination pairs, travel times, and passenger demographics.

1. Hu and Du (2024) - Passenger Group Size and Tipping: An Empirical Study of 50 million NYC Yellow Taxi Rides:

Hu and Du undertook an empirical study to investigate the relationship between passenger group size and tipping behavior using a dataset of 50 million NYC yellow taxi rides. It focuses on one of the predictors we plan to add to our model, namely the passenger count.

1. Aydin and Acun - An Investigation of Tipping Behavior as a Major Component in Service Economy: The Case of Taxi Tipping:

Aydin and Acun conduct an in-depth investigation of tipping behaviors as a major component of the service economy, focusing specifically on taxi tipping. The study employs a mixed-method approach, combining surveys, experiments, and observational data to explore the psychological, social, and economic factors shaping tipping decisions.

1. Tan and Zhang (2021) - Good Days, Bad Days: Stock Market Fluctuation and Taxi Tipping Decisions:

Tan and Zhang examine the impact of stock market fluctuations on taxi tipping decisions, employing econometric techniques to analyze transactional data from taxi services. This research helps us in identifying another possible key predictor to add to our model.

1. Zhan et al. (2016) - A Graph-Based Approach to Measuring the Efficiency of an Urban Taxi Service System:

Zhan and colleagues propose a graph-based approach to measuring the efficiency of an urban taxi service system, utilizing network analysis techniques to assess service quality and performance. Using

efficiency as a factor to add to our prediction model is also an approach we considered after looking at this paper.

1. Lee and Sohn - A Large-scale Data-based Investigation on the Relationship between Bad Weather and Taxi Tipping:

Lee and Sohn undertook a large-scale data-based investigation to examine the relationship between adverse weather conditions and taxi tipping behaviour. The study involves the collection and analysis of transactional data from taxi services, along with meteorological data on weather conditions. This research provided us guidance in using weather data in conjunction with the dataset, to add it as a predictor in our model.

1. Ferreira Neto et al. - Do Tourists Tip More Than Local Consumers? Evidence from Taxi Rides in New York City:

Ferreira Neto and colleagues investigate differences in tipping behaviour between tourists and local consumers using transactional data from taxi rides in New York City. Methodologically, the study involves the segmentation of passengers based on residency status and the analysis of tipping amounts and frequencies.

## Proposed Method

The proposed method is to perform analysis on the relationship of each variable with the response (tip ratio) and to perform feature selection to pick out the most highly correlated variables with the response. We will also run a regression model using these features to assess the quality of these observations. A more detailed rundown of the experiments is given in the ‘experiment’ section. We believe this approach is novel because it serves to give us a meta-analysis on all the existing research by finding the strongest predictors for tipping behavior among each one that has been used in existing research, rather than perform a detailed analysis on its relationship with any one or more of these variables.

## Experiments

Questions we want to answer:

1. What are the factors influencing tipping behavior for customers of the taxi rides involved in this dataset i.e. under what conditions can we expect a customer to pay a percentage of tip above or below average?
2. Based on existing research on the subject, which of the predictors that have already been studied are useful in predicting this behavior, and what are the top predictors among these?
3. Can we possibly bring in environmental factors, and see how much they influence the daily number of trips across the dataset?

To answer the above questions, we started by introducing useful data from external datasets into the data which we could use for further analysis. We brought in Borough-level income data, and daily weather data including precipitation, avg temp, avg humidity, avg windspeed. Using this accumulated data, we performed EDA on the consolidated dataset.

First, we tried to analyze the daily number of trips in each borough and the relationship of that variable with other aggregated data points from the dataset:

A graph with blue dots

Description automatically generatedA graph with blue dots

Description automatically generated

A graph of blue dots

Description automatically generated

Then we created a variable called “tip ratio” which is nothing, but a factor of the tip amount divided by the fare amount of each ride. We performed EDA for this variable and analyzed its relationship with potential predictors:

A graph with numbers and lines

Description automatically generated

The above diagram gives use the distribution of our response variable i.e. tip ratio. The most common ratio is between around 0.25 and 0.35.

A graph with blue dots

Description automatically generated

Looking at the scatterplot between tip ratio and fare amount, we can see that there is a slight negative correlation between the 2 variables i.e. for higher fare amounts, the tip ratio tends to be lower.

A graph with blue dots

Description automatically generatedA graph of blue dots

Description automatically generated

It’s a similar story for the relationship between tip ratio and the variables trip duration and additional charges (charges outside of fare and tips).

A graph of a number of boxes

Description automatically generated

The distribution for tip ratio seems to be similar for all values of passenger count, with median values around the same for all values of passenger count except for values above 6.

A graph of a number of boxes

Description automatically generated with medium confidence

For each borough, the distribution of the tip ratio variables seems to be drastically different, suggesting there might be some use to including borough-level information like average income.

A graph with lines and dots

Description automatically generated

Similarly, the payment type ‘1’ seems to have a drastically different distribution for the tip ratio variable as compared to other payment types.

We then performed feature selection for all these features and tried to run regression to see how much the feature selection influences the model.

## Results

After running Lasso regression on the dataset with alpha = 0.01, the features that were selected were ‘payment\_type’,’fare\_amount’,’trip\_duration\_sec’,’additional\_charges’,’avg\_income\_borough’.

Comparing linear regression models using all variables and using only the feature selected models, the result we got is:

Linear Regression with Lasso Feature Selection:

Train RMSE: 0.09806526710844032

Test RMSE: 0.09816917755189361

Train R2-Score: 0.4726056413430354

Test R2-Score: 0.4709899116217966

Linear Regression without Feature Selection:

Train RMSE: 0.09801447443335352

Test RMSE: 0.09811728022933713

Train R2-Score: 0.47315182520119714

Test R2-Score: 0.471549088158001

Since the R2-score for both models is the same, we can conclude that the variables other than the ones selected by Lasso regression have negligible effect on the model, meaning they don’t have much of a relationship with the response variable i.e. tip ratio.

The LR co-efficients of each selected variable is given below:

payment\_type(Code associated with type of payment): -1.39434607 \*10^-1

fare\_amount(The base fare amount of trip): -3.02825516e \*10^-4

trip\_duration\_sec(Trip duration in seconds): -4.40815814\*10^-05

additional\_charges(Additional charges outside of fare and tip): 6.18297282\*10^-03

avg\_income\_borough(Average income of borough where pickup location of trip is located): 5.28595681\*10^-07

In terms of model accuracy, we found better accuracy using Random Forest Regression, although there was slight overfitting on training data:

Random Forest with feature selection:

Train RMSE: 0.07950652131108234

Test RMSE: 0.08795268318614785

Train R2-Score: 0.6533345560899135

Test R2-Score: 0.5753688642464024

## Effort Statement

Everyone in the group contributed equally, with tasks ranging from data mining, cleansing, model selection, model evaluation and preparation of documentation.

# References:

1. Tang, Y. (2019) Big Data Analytics of Taxi Operations in New York City. *American Journal of Operations Research*, **9**, 192-199
2. Elliott, David & Tomasini, Marcello & Oliveira, Marcos & Menezes, Ronaldo. (2017). Tippers and Stiffers: an Analysis of Tipping Behavior in Taxi Trips.
3. Correa, Diego, Exploring the Taxi and Uber Demands in New York City: An Empirical Analysis and Spatial Modeling (August 1, 2017).
4. Judd Cramer, Alan B. Krueger. Disruptive Change in the Taxi Business: The Case of Uber.
5. Riascos, A.P., Mateos, J.L. Networks and long-range mobility in cities: A study of more than one billion taxi trips in New York City. *Sci Rep* 10, 4022 (2020).
6. Xie C, Yu D, Zheng X, Wang Z, Jiang Z (2021) Revealing spatiotemporal travel demand and community structure characteristics with taxi trip data: A case study of New York City. PLoS ONE 16(11): e0259694.
7. Hu, Ye and Du, Rex Yuxing, Passenger Group Size and Tipping: An Empirical Study of 50 million NYC Yellow Taxi Rides (January 25, 2024).
8. Asli Elif Aydin, Yüksel Acun,An investigation of tipping behavior as a major component in service economy: The case of taxi tipping,Journal of Behavioral and Experimental Economics
9. Weiqiang Tan and Jian Zhang, Good Days, Bad Days: Stock Market Fluctuation and Taxi Tipping Decisions, Management Science 2021
10. X. Zhan, X. Qian and S. V. Ukkusuri, "A Graph-Based Approach to Measuring the Efficiency of an Urban Taxi Service System," in IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 9, pp. 2479-2489, Sept. 2016
11. Won Kyung Lee, So Young Sohn,A large-scale data-based investigation on the relationship between bad weather and taxi tipping,Journal of Environmental Psychology
12. Amir B. Ferreira Neto, Adam Nowak, and Amanda Ross, Do Tourists Tip More Than Local Consumers? Evidence from Taxi Rides in New York City, International Regional Science Review 2019