Predicting Taxi Tips in NYC

GROUP NO. 24

Presentation link - Presentation, Youtube_Presentation

Velunde Swapnil
2260842
School of Computing
Dublin City University
Dublin, Ireland
swapnil.velunde2@mail.dcu.ie

Patil Nandkishor
22261985
School of Computing
Dublin City University
Dublin, Ireland
nandkishor.patil3@mail.dcu.ie

Khot Chaitanya
22262858
School of Computing
Dublin City University
Dublin, Ireland
chaitanya.khot2@mail.dcu.ie

***Abstract* - In order to create a more accurate tip prediction model, this paper will look at the significant variables affecting tipping behavior in the yellow cab sector in New York City (NYC). In order to assist both passengers and taxi drivers, it is crucial to understand the factors driving tipping behavior, given that over 97% of customers leave tips. Numerous variables, including the passengers' demographics, the trip's length, and the taxi driver's conduct, impact tipping behavior. The study of the literature looked at numerous research that attempted to forecast tipping patterns in the NYC taxi business. The findings suggested that various variables, including standard tip options, estimated journey length and pricing, and regional and temporal trends, may be used to forecast tip amounts. However, these studies have areas for improvement, such as possible data biases and little control over outside variables. For the benefit of taxi drivers and other stakeholders, this study paper underlines the significance of identifying the critical variables that affect tipping behavior and constructing a more accurate tip prediction model.**

Keywords - Tipping behavior, Tip prediction model, Missing Data Handling, Linear Regression, Random Forest

I. INTRODUCTION

The New York City (NYC) yellow taxi industry has been an integral part of the city's transportation system, serving millions of passengers annually. Tipping is an essential aspect of the NYC taxi industry, and it has been observed that tipping behavior varies widely among passengers. Therefore, understanding the factors influencing tipping behavior is crucial for passengers and taxi drivers. This report aims to investigate the key factors influencing tipping behavior in the NYC taxi industry and how this knowledge can aid in developing a more reliable tip prediction model. The variability in tipping behavior has been influenced by various factors, such as passenger demographics, ride duration, and taxi driver behavior. A better understanding of these factors could help develop a more reliable tip prediction model, which would benefit taxi drivers and stakeholders. Our main goal is to identify key factors that influencing tipping behavior and how this knowledge will aid in development of a more reliable tip prediction model.

II. RELATED WORK

As per the work of O. H. Azar in [1], tipping is customary in numerous sectors, including the New York Yellow Cab Company. Over 97% of travellers pay with a credit card tip, indicating the existence of a norm. Road conditions and the holiday season can impact norms and compliance, whereas the bystander effect may reduce tipping when numerous passengers travel in a cab.

Research conducted by K. Haggag and G. Paci in [2] discovered that default tip selection considerably influenced tipping behavior in the restaurant industry. In a restaurant setting, they discovered that when guests were offered default tip selections of 15%, 18%, or 20%, the group with the 20% default tip choice had a higher average tip amount.

Another research by S. Devaraj and P. C. Patel in [3] discovered a link between sunshine and the magnitude of tips in NYC cabs. The study gives insights into the psychological variables that impact tipping behavior by considering the weather to enhance tips. However, the study only addressed one external factor and did not examine other potential factors that may impact tipping behavior. More research is needed to properly comprehend the complexity of tipping patterns in the NYC taxi business.

A study by W. Tan and J. Zhang in [4] examined the relationship between stock market fluctuations and tipping decisions in the NYC taxi industry over two years, using data from 10 million taxi trips and statistical analysis techniques to investigate differences in tipping behavior across rider demographics and times of day. The study's limitations include its sole emphasis on the stock market as the only external economic element and its short duration.

The work done by J. Zhang in [5] investigates large-scale taxi travel data in New York City utilizing outlier identification methods and machine learning approaches. The study used a dataset of over 170 million cab journeys to find trends in the data, such as popular sites and weather influence, and to compute the shortest path for travel using the MoNav-CH method. However, it does not directly address tipping behavior and uses an older dataset from one year ago, limiting its applicability to current taxi usage trends and tipping behavior.

In the paper [6], the authors gave insights into machine learning techniques for predicting taxi trip time and may be used to construct a solid tip prediction model for NYC yellow cabs. The research employs feature engineering approaches as well as MLP and XGBoost algorithms. However, the quality of the prediction model may only sometimes apply to forecasting tip amounts since the factors driving tip amounts may differ from those influencing trip duration. The research emphasizes the significance of specific parameters, such as the distance between pickup and dropoff sites, but does not explicitly investigate the factors driving tipping behavior.

In her study M. Lynn [7], blends ideas of social trade, equality, emotive events, and social identity to present a motivational framework for tipping behavior. The model considers several variables: situational considerations, service quality, societal norms, and client and server characteristics. Understanding tipping behavior in various service scenarios is easier with this framework's help. The study is, however, constrained by a need for more empirical data and a Western cultural focus.

In the work done by B. Chandar, U. Gneezy, J. A. List, and I. Muir [8], the authors explored the factors that motivate tipping behavior in the US through a nationwide field experiment. The study finds that social preferences, such as altruism and reciprocity, play a significant role in tipping behavior and highlight potential implications for the taxi industry. A limitation of this study includes using a controlled experiment that may not fully capture real-world tipping behavior.

The work of S. Jain and A. See [9] reviewed many studies on forecasting tipping patterns in the NYC taxi sector. Field experiments, significant data approaches, and data exploratory analysis were some of the methodologies used in the investigations. The findings suggested that several variables, including standard tip options, estimated journey length and fee, and regional and temporal patterns, could help estimate tip amounts. Potential data biases and the absence of control over outside factors are two problems that limit these investigations. This study by Jain et al. used location binning and averages to forecast tip amounts, and the outcome was 54.62%. The study did not consider the influence of other variables, such as customer characteristics and service quality, on tipping behavior.

Using exploratory data analysis methods, the study by N. Natashia and S. Velu [10] offers insightful information on NYC taxi data. Over 1 billion cab trips are included in the dataset, spanning 2009 through 2015. The analysis's findings indicate that Manhattan is home to most pickup and drop-off locations, with Midtown accounting for most trips. The study also shows a connection between fee amount and trip length. However, the paper does not mention tipping practices in the NYC taxi sector. Despite its shortcomings, the study provides a basis for additional examinations of tipping behavior, research, and creating a trustworthy model for tip prediction.

In order to promote the feasibility of electric taxis in New York City, the study [11] proposes a data-driven way to improve the taxi service strategy. The authors look at much data on taxi trips and electric vehicle charging to build a model that optimizes taxi service operations. According to the study, it is possible to significantly increase the viability of electric taxis by optimizing taxi service operations to reduce downtime for charging and increase the number of trips that can be performed on a single charge. The authors also touch on the challenges of implementing such a paradigm. The research provides important details regarding the potential for data-driven methods to improve the overall viability of electric taxis.

The study conducted by D. Elliott, M. Tomasini, M. Oliveira, and R. Menezes [12] looks into the factors that affect tipping behavior in the New York City (NYC) taxi industry using a dataset of cab journeys. The study aims to identify the key factors impacting tipping behavior in order to develop a tip prediction system that is more accurate. According to the author's study of the data using descriptive statistics and regression analysis, the journey distance, time of day, and payment method are just a few of the variables significantly affecting tipping behavior. The research provides helpful information for cab drivers and companies aiming to boost customer happiness and income. The focus on New York City and dependence on self-reported data are two issues with the study. Future research should consider a broader range of factors and data from multiple cities and countries to understand tipping behavior in the taxi industry comprehensively.

III. DATA MINING METHODOLOGY

The experimental design is disseminated into four subsections - Data Collection, Data Preparation, Feature Selection.

A. Data Collection

Data was gathered on the NYC Taxis for the year 2019 from the NYC Open Data Repository. The dataset consists of 84 million rows and 19 features. The data is available on a monthly basis and in parquet format. Using the entire dataset introduced the challenge of computing cost and efficient memory management. Since this was not the intended scope of study, sampling approach was taken into consideration. Stratified sampling approach was selected for sampling this dataset as it is known to produce samples which are more representative of the population []. Four strata were nominated namely, pickup day, dropoff day and pickup hour, dropoff_hour. For stratified sampling, it is essential that the chosen strata had skewed distribution. It was observed that pickup day and dropoff_day followed normal distribution while pickup hour and dropoff_hour followed left-skewed distribution. This aided in narrowing down the options to pickup_hour and dropoff_hour. High collinearity was perceived between them and pickup_hour fixed as the basis for performing stratified sampling on monthly data as it provided with more insights on a day-to-day basis. The margin of error was calculated which determines whether a sample is an appropriate representation of the population. The accepted range for appropriate sampling is 2% and calibration recorded 1.2% which established that the sample was significant.

B. Data Preparation

After sampling was done, initial check was performed to check if all the data belonged to the year 2019. This constraint was kept as the NYC Taxi Dataset is a massive big dataset consisting of more than 6 million rows for every month. Therefore, any data related to any other year, is not going to be our intended scope for the dataset. But we have kept the data where the pickup year and dropoff year was 2019, as there may be trips occurring near the midnight time of 31st December. Preprocessing consisted of identifying missing information and handling, checking for duplicate data, understanding and resolving negative behaviour of data and making insights on zero data. For the first issue, the absence of airport fees in the total amount column was detected. Despite the metadata specifying a standard airport fee, no airport fees were found in the total amount for any airport or the specified airport rate code. Therefore, the decision was made to drop the airport fee column from the analysis to avoid potential bias. Furthermore, removing missing values for passenger count also resulted in the removal of null values for RatecodeID, store_and_forward_flag, and zero data for payment_type, indicating that the same rows consisted of missing information. In case of congestion_surcharge, using Kolmogorov-Smirnov Test it was found that the column possessed Not Missing at Random data. The possible values of congestion_surcharge were imputed using knowledge of the dataset, by calculating the difference between total amount and the sum of all other charges, and using them to replace null values of congestion_surcharge. Negative behaviour was observed in fare charges. However, a pattern was discovered in which all charges in a particular row were negative and their absolute values equalled the total amount. The minus sign before the amount is assumed to be due to a technical failure, as the total amount calculation would not have been correct otherwise. To obtain positive charge values, absolute value was used for numeric data. Total amount data was found to be zero in some cases when trip distance was zero, which could be due to trip cancellations or no trips taken. As per metadata dictionary provided, payment type 3 was considered valid if trip distance was not zero because it could be for personal use by the driver. However, where trip distance was specified but no payment information was available, the data was removed from the analysis because it represented a small portion of the dataset and was unlikely to significantly impact model performance.

C. *Feature Selection*

The features were recognized as per NOIR framework into nominal, ordinal, interval and ratio variables. Statistical tests were performed on them as per their NOIR type. For categorical variables, the Chi-Square Test was used to determine strong relations with the target variable. In case of continuous data, Pearson's correlation coefficient was performed. Finally, the multicollinearity check was performed using the Variance Inflation Factor to avoid the high collinear data. It was discovered after analysing the data that fare amount had a strong positive correlation with total amount and a strong negative correlation with mta_tax. To avoid multicollinearity issues, it was decided to remove fare amount from the analysis, as removing either total amount or mta_tax would still leave one variable with high multicollinearity. At the end, we got twelve features after undergoing statistical test verification. After conducting statistical tests for feature selection, we arrived at a final set of twelve features that were deemed statistically significant for our analysis.

IV. EVALUATION/RESULTS

The Multiple Linear Regression model was chosen as the baseline model with selected features. R-Squared, Mean Squared Error (MSE), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were used for evaluation. Other metrics like Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) along with error distribution plots namely, Residual and Q-Q plots were used. Upon applying the model, R-squared was recorded to 0.51, MSE at 3.75, MAE at 0.85 and RMSE at 1.94. The residual plot displayed a random scattering of points around the horizontal line at y=0. Moreover, the residuals followed a straight line on the Q-Q plot which indicates that they are normally distributed. For checking overfitting, Ridge Regression was implemented along with GridSearchCV for selecting the optimal alpha value. The next model selected was Random Forest Regression. For this normality was checked on the tip amount variable. After satisfying the criteria, Random Forest with 100 trees was carried out. We got an improved score of R-squared at 0.88, MSE at 0.94, MAE at 0.22 and RMSE at 0.97. The error distribution plots displayed confirmed normal distribution as well.

V. CONCLUSION AND FUTURE WORK

In conclusion, this research paper investigated the key factors influencing tipping behavior in the NYC taxi industry and their implications for developing a reliable tip prediction model. The findings of this study contribute to the existing literature on tipping behavior by considering a comprehensive set of factors and employing advanced statistical techniques. The results highlight the importance of passenger count, trip distance, fare amount, and RatecodeID in influencing tipping behavior. Surprisingly, the congestion surcharge which is extra charge for traffic also played a role in increasing the tip received. Identifying these key factors can aid in developing a more accurate and reliable tip prediction model, which can benefit taxi drivers, passengers, and stakeholders in the NYC taxi industry. Further research and analysis can be conducted to refine and validate the tip prediction model and explore other factors that may influence tipping behavior in the industry. The findings of this study can also have implications for other industries where tipping is a common practice.

To delve deeper into the impact of psychological factors, such as happiness, on taxi tipping behavior, we could analyze the data by comparing it with the performance of New York's most popular sports teams, such as the New York Yankees. If there is a significant increase in the tipping percentage on the days when the Yankees win, it could support the hypothesis that people tend to tip more generously when they are happy due to the success of their favourite team. Another factor not included in this data was the car's condition. Factors such as the car's cleanliness could significantly impact customer satisfaction, and invariably the tip amount is not present in the dataset. Including such actors will aid in accurate prediction

Overall, this research contributes to understanding tipping behavior in the NYC taxi industry and provides insights for future research and practical applications in the field.

REFERENCES

[1] O. H. Azar, "The economics of tipping," *J. Econ. Perspect.*, vol. 34, no. 2, pp. 215–236, 2020.

[2] K. Haggag and G. Paci, "Default Tips," *Am. Econ. J. Appl. Econ.*, vol. 6, no. 3, pp. 1–19, 2014.

[3] S. Devaraj and P. C. Patel, "Taxicab tipping and sunlight," *PLoS One*, vol. 12, no. 6, p. e0179193, 2017.

[4] W. Tan and J. Zhang, "Good days, bad days: Stock market fluctuation and taxi tipping decisions," *SSRN Electron. J.*, 2017.

[5] J. Zhang, "Smarter outlier detection and deeper understanding of large-scale taxi trip records: A case study of NYC," in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, 2012.

[6] M. Poongodi *et al.*, "New York City taxi trip duration prediction using MLP and XGBoost," *Int. J. Syst. Assur. Eng. Manag.*, vol. 13, no. S1, pp. 16–27, 2022.

[7] M. Lynn, "Service gratuities and tipping: A motivational framework," *J. Econ. Psychol.*, vol. 46, pp. 74–88, 2015.

[8] B. Chandar, U. Gneezy, J. A. List, and I. Muir, "The drivers of social preferences: Evidence from a nationwide tipping field experiment," *SSRN Electron. J.*, 2019.

[9] S. Jain and A. See, "Predicting Taxi Tip-Rates in NYC [A predictive model focused on location binning and averages]," *Ucsd.edu*. [Online]. Available: https://cseweb.ucsd.edu//classes/sp15/cse190-c/reports/sp15/050.pdf. [Accessed: 15-Apr-2023].

[10] N. Natashia and S. Velu, ‘DATA EXPLORATORY ON TAXI DATA IN NEW YORK CITY’, 04 2020.

[11] C.-M. Tseng, S. C.-K. Chau, and X. Liu, “Improving viability of electric taxis by taxi service strategy optimization: A big data study of New York city,” *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 3, pp. 817–829, 2019.

[12] D. Elliott, M. Tomasini, M. Oliveira, and R. Menezes, “Tippers and stiffers: An analysis of tipping behavior in taxi trips,” in *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 2017, pp. 1–8.

[13] Link for the dataset: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page