CA683: Data Analytics and Data Mining

A report submitted to Dublin City University, School of Computing for module CA683: Data Analytics and Data Mining. We hereby certify that the work presented, and the material contained herein is our own except where explicitly stated references to other material are made.

Student Name	Student ID	Qualification	Module	Email Address
Romil Sakariya	21264095	MCM	CA683	romil.sakariya2@mail.dcu.ie
Khilti Dedhia	21264200	MCM	CA683	khilti.dedhia2@mail.dcu.ie
Sharmistha Sawant	21263197	MCM	CA683	sharmistha.sawant2@mail.dcu.ie
Romain Bernard	21114595	Erasmus	CA683	romain.bernard3@mail.dcu.ie

GitHub Link:
https://github.com/romilskr3/CA683-Data-Analytics-and-Data-Mining-Assignment/blob/main/CA683_DM_Assignment.ipynb

Youtube Link for Video Presentation:
https://drive.google.com/file/d/1LKA_8nvI2THnkVAJeKWHWk4YvIfkHwrP/view?usp=sharing

Drive link for all the files:
https://drive.google.com/drive/folders/1BVQo5WNjQBBoen-8HU1tZDr0GG0-IyXh?usp=sharing

Analysis Of New York City Taxi Dataset To Predict Tipping Behaviour Of Passengers

Romil Sakariya
Department of Computing
Dublin City University
Dublin, Ireland
romil.sakariya2@mail.dcu.ie
Student ID: 21264095

Khilti Dedhia
Department of Computing
Dublin City University
Dublin, Ireland
khilti.dedhia2@mail.dcu.ie
Student ID: 21264200

Sharmistha Sawant
Department of Computing
Dublin City University
Dublin, Ireland
sharmistha.sawant2@mail.dcu.ie
Student ID: 21263197

Romain Bernard
Department of Computing
Dublin City University
Dublin, Ireland
romain.bernard3@mail.dcu.ie
Student ID: 21114595

Abstract— As an NYC taxi driver, incoming revenue from rides can vary substantially from one day to another. There are some locations that offer a higher rate of fares such as densely populated residential areas or office areas. But are there certain trip characteristics such as average trip speed or trip distance that correspond to higher tip percentage from passengers? We explore this question by analyzing the fares and tips of all yellow taxi trip records that took place in the month of July 2021. We extended this analysis to implement classification and regression models by using feature engineered trip characteristics as dependent variables to predict their corresponding tip percentage. Upon analyzing the results produced by these models, we can conclude that although trip characteristics are not the primary factors that influence a passenger's tipping behaviour, they do hold some significance in the aspect and can be leveraged to the taxi driver's financial benefit.

Keywords—prediction, statistical analysis, New York City taxi dataset, regression modeling

I. Introduction

In a city as densely populated and widely distributed as New York City, public transportation can be a key factor to most residing individuals. New York City public transport offers a wide array of services ranging from metros to buses to cabs and even limousines. As a resident of NYC, one is over familiar with the standard yellow taxis that run around the city in large numbers. Approximately 13,587 yellow taxis assist commuters in NYC to navigate various parts of the city. Although this number may seem adequate, cab drivers are regularly overworked to keep up with the ever-growing demand of this ever-growing city. Taxi drivers understand the location parameters that can help them perform better financially but there are some time slots such as the evening and morning rush hour where cab drivers can considerably increase their fare frequency. In this paper we intend to statistically analyse the trip data from July 2021 to pinpoint factors contributing to higher earnings from tips for taxi drivers.

II. Related work

Tipping behaviour of taxi passengers can be affected by many factors. The climate, hygiene of the taxi, income of the passenger, quality of service provided by the driver are just some such factors that may have an influence on the tipping behaviour of passengers.

A lack of trustworthy data in the past has led to many misleading studies that incorrectly indicate various factors affecting the tipping behaviour of passengers. The study conducted by Elliot et al. criticize the findings of past researchers indicating that race and income have an impact upon the tipping behaviour of people. They back up this criticism by explaining how poor sampling practices and lack of consideration of income backgrounds lead to incorrect results. They then conclusively prove that a passenger's tipping behaviour is independent of the income of their pick-up and drop-off locations. [3].

Since the introduction of more reliable data gathering techniques for taxi trips began early in the decade, there have been more reliable and trustworthy studies conducted in this area of study. Researchers have used taxi trip data to analyse location and routing based data to estimate and predict various factors about a neighbourhood or locality. Zhan et al. attempt to estimate urban link travel time [4] whereas Chang, Tai and Hsu [5] use taxi location data to predict potential request hotspots in a city. Deri and Moura can pinpoint locations in NYC that exhibit co-behaviour at a certain time of day using the location data in their dataset [7]. There have also been a few studies that focus on the taxi drivers themselves. Leng et al. tried to analyse taxi drivers' psychological behaviour shift before and after the introduction of online applications to reserve taxis [6]. And then there are some studies that focus on the financial aspect of taxi trips data such as Dong et al. developed methods to reveal taxi drivers' operation patterns to make more revenue [8].

Raymond and Cramer have observed a reciprocity effect when customers hails taxis instead of when they are dispatched for them [9]. They statistically found that passengers tend to tip hailed taxi rides twice the percentage they would generally tip dispatched taxi rides. Lee and Sohn also observed a reciprocity effect when they analysed tipping behaviour in extreme weather conditions. [10] Their study observed a higher tipping tendency by passengers travelling in a cab in extreme weather. This is understandable as passengers feel grateful for the service, they receive even in extreme weather conditions and correspondingly tip generously.

All the above researchers have attempted to analyse and predict the tipping behaviour of passengers based on intuitively related factors such as quality of service and income of passenger, but not a lot of work has been conducted in analysing the effect that trip characteristics such as average speed and time of the day when trip was initiated. These trip characteristics may or may not have a higher effect on passengers' tipping behaviour if analysed effectively.

III. Dataset, Operations and Exploratory Analysis

A. Dataset

The NYC government has enabled TLC (Taxi & Limousine Commission) to publish trip records from all the taxis that run in NYC [1]. This data provided by TLC is categorized into different types of taxi that commute in NYC. This large open-ended data is organized by months and years and can be accessed accordingly. We decided to select the month of July 2021 for this paper. The downloaded raw data consists of the following attributes: VendorId, tpep_pickup_datetime, tpep_dropoff_datetime, passenger_count, trip_distance, store_and_fwd_flag, PULocationID, DOLocationID, RatecodeID, payment_type, fare_amount, extra, mta_tax, tip_amount, tolls_amount, improvement_surcharge, total_amount, and congestion_surcharge. There were 2,821,515 recorded yellow taxi trips taken in the month of July 2021 as stated by this dataset. After going through related work and analyzing our dataset, we decided to use the following attributes to understand the tipping behaviour of passengers: passenger count, trip distance, speed of trip, time of day of the trip, day of week of the trip.

B. Data Processing

Our first data processing step was to only consider trips that were paid for using credit cards. The data dictionary provided by TLC along with the dataset states that tips are only recorded for trips that are paid for using a credit card. Hence, we only considered trips that were paid for using a credit card.

I. Handling NULL values

As our next step, we checked if there were any null values in the dataset. We found some null values for the trip distance and passenger count variables. These missing values seemed to be missing completely at random (MCAR) and since imputing the number of passengers or the trip distance for a taxi trip seemed counterintuitive, we dropped the rows that contained null values. We also checked if there were any duplicate trip entries in the dataset and found no duplicate values in the entire dataset.

II. Handling Negative values

The financial data in the dataset has some negative values for fare amount, total amount, and tip amount. These negative values may have been caused by faulty equipment while data collection step by TLC. These negative values do not make logical sense in a financial data context. Since we cannot just consider the absolute values because the data is untrustworthy, we decided to drop these rows from our dataset.

III. Feature Engineering

a) Calculating Tip Percentage

The tip amounts provided for each trip have a large variance depending on the trips. To normalize this variable, we calculated the tip percentage for each trip. We considered that tip percentage of a trip is the ratio of the tip amount over the total fare amount. As expected, the tip percentage for all the trips was somewhere in the range of 0 to 100.

b) Calculating Trip Duration

Using the pick-up timestamp and drop-off timestamp, the trip duration of each trip can be calculated as a difference between these timestamps.

c) Calculating Average Speed of each Trip

Having obtained the time duration of each trip, we can calculate the average speed of the trip by taking a ratio of the total trip distance with respect to the trip duration.

d) Determining Day-of-Week for each Trip

We used the pick-up timestamp in the dataset to obtain the day of the week when each trip was initiated. We represented each day of the week numerically in the range 0 to 6 starting from Sunday.

e) Creating Tip Percentage Buckets

We created two buckets – one for all trips earning a tip percentage lower than or equal to the median tip percentage and another one for all trips having tip percentage higher than the median. These buckets or bins would be used as categories to be predicted by our classification model.

f) Creating Buckets for Time-of-Day

We created four buckets to classify each trip based on the period of the day it was taken. The four buckets were created at equal boundaries of 04:00 to 10:00, 10:00 to 16:00, 16:00 to 22:00

and 22:00 to 04:00. Each of this bucket represents a different slot of a day – Morning, Afternoon, Evening, Night.

IV. *Handling Outliers*

a) Trip Distance

There were several outliers for trip distance variable. There were trips with a trip distance value of 0 miles, and some trips had a trip distance value of greater than 1000 miles. We handled these outliers by dropping any trips with trip distances less than 1 mile or more than 850 miles. This range is a logical bound for a yellow taxi that operate in downtown NYC.

b) Passenger Count

Passenger counts are entered manually by drivers for every trip. Since, it is manually added there is a high possibility that it may be incorrectly entered for some trips. The maximum number of passengers allowed to travel in a NYC yellow taxi by law is 6. There are also several trips that have a passenger count of 0 which can be considered incorrectly entered data since a taxi trip can only be considered if there was at least 1 passenger. Therefore, we consider all trips with passenger count outside the range of 1 to 6 as outliers and dropped them since the data for these trips cannot be trusted.

c) Trip Duration

There are some trips that have a trip duration of 0 seconds. These trips have the same pick-up and drop-off timestamp values. Upon investigating further, we found that this error in recording timestamp is caused by faulty equipment. We consider a lower bound of 1 minute for a trip to be considered as a valid trip since any trip lasting less than a minute may be caused by faulty equipment recordings.

d) Average Speed of Trip

The legal speed limit of operation for cars in NYC is 25 miles-per-hour. This restriction also applies to NYC yellow cabs and was introduced in 2014 [11]. After calculating speed for each trip, we considered trips with an average speed of greater than 50 miles per hour to be inconsistent data. Such high average speed was caused by either very large trip distances or very small trip duration or in some cases both. To ensure our data for average speed of trips is accurate, we only considered trips that had average speed of less than 50 miles per hour.

C. *Exploratory Analysis*

For our analysis, we did not require any location-based data or detailed breakdown of the amount of taxes imposed on a particular trip. So, while exploring, we dropped these irrelevant attributes since they would only contribute to computational time and not provide any relevant insights for our analysis. Initially, we explored the distribution of average speed of trips. Upon plotting a histogram of the average speeds for all the trips, we found the data to be left skewed. A similar skewness is observed in the other variables as well. Figure 1 represents the histogram of the average speed for all trips. To rectify this skewness, we decided to take a natural logarithm of the variables. This handles the skewness and checks the influence of any outliers in the variable. Figure 2 gives the distribution of the natural logarithmic values of the tip percentage of all trips.

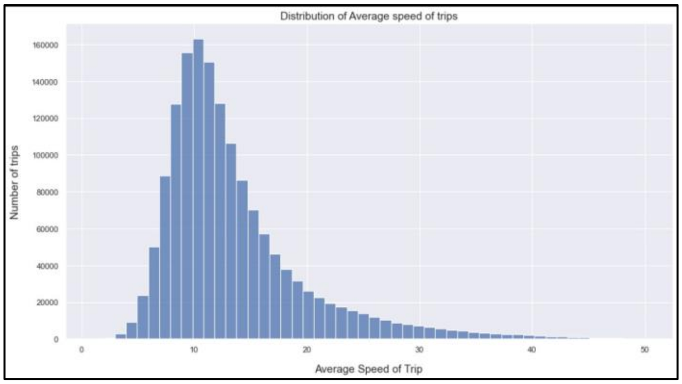

Figure 1. Histogram of the Average Speed for all trips.

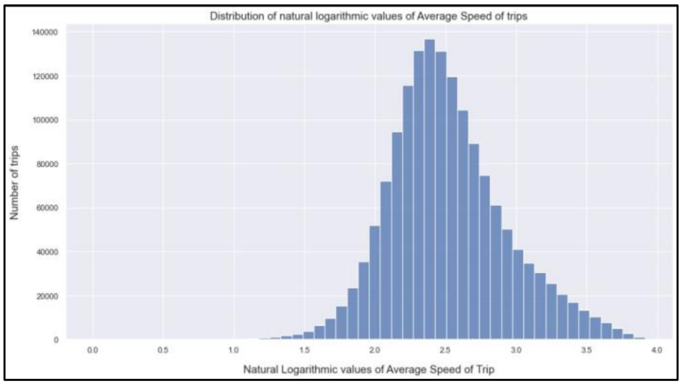

Figure 2. Histogram of the natural logarithmic Average Speed for all trips.

Another exploratory factor was the number of rides taken during weekdays versus weekends. We simply analysed all the trips in December based on whether they were initiated on a weekday or on a weekend. One unique finding was the difference in the number of rides taken on a weekday versus the weekend. In a city like New York, one would expect more trips being taken on weekends rather than weekdays, but the data reveals exactly the opposite. Figure 3 depicts the comparison on the number of rides on a weekday versus number of rides on a weekend.

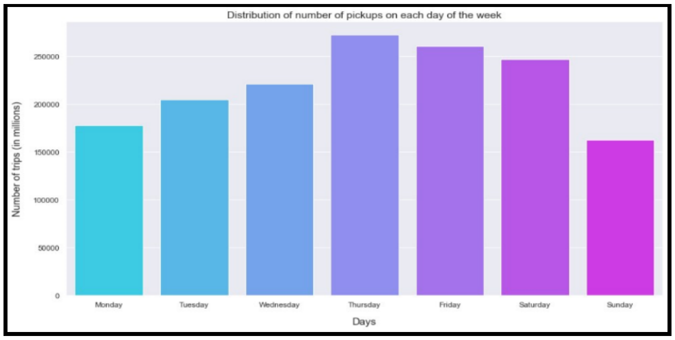

Figure 3. Distribution of number of pickups on each day of the week.

IV. Data Mining Methologigies

Once we had cleaned and processed the raw data into usable features, we moved on to create models to predict tip percentage given by passengers. While exploring the dataset, we observed that the relationship between tip percentage and other independent variables in the dataset was not exactly linear. This observation was then solidified once we constructed the correlation matrix. There was very little correlation between tip percentage and other independent variables. The range of the correlation coefficient was -0.001 to 0.008. Other variables did have some high correlation coefficients - as expected, average trip speed was highly correlated with trip distance and trip duration. We then checked the variables for any multi-collinearity using variance inflation factor or VIF. The VIF values denoted a high multicollinearity caused by two features: the natural logarithmic values of average speed of trips and natural logarithmic values of trip duration. Upon dropping these two variables, the VIF factors came down into the acceptable limits.

I. Decision Tree Classification Model

By this point we deduced that no form of linear regression would be able to fit this dataset effectively. The assumptions for linear regression would not hold true for this dataset. Therefore, we moved to non-linear regression methods, namely Decision Trees and Random Forest. Decision Tree is a supervised learning model for classification and regression. The model predicts the value of the dependent or target variable from some learning decisions of the available features or independent variables [12]. We first started by using a single decision tree to classify all the tips in the dataset into two classes: lesser or more than the tip percentage median. We had already created buckets and classified all trips that obtained a tip percentage of lesser than the median to be in bucket 1 and the rest to be in bucket 2. Then we split the dataset into two parts - training and testing data in a 70:30 ratio respectively.

II. Random Forest Regression Model

A random forest is an estimator that fits several classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting [13]. We used Random Forest Regression model to accomplish two tasks. We used the model to determine the importance of all the independent features. We also used the regression model to predict tip percentage values for the test dataset.

V. Evaluation & Results

I. Decision Tree Classification Results:

We provided the training dataset to the Decision Tree Classifier model and calculated the accuracy of the predictions made by the classifier against the testing dataset. Figure 4 represents the confusion matrix of the Decision Tree Classifier model.

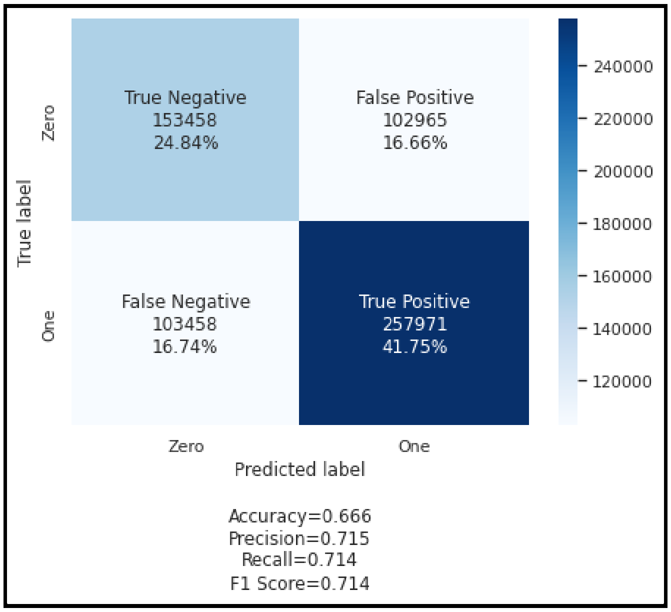

Figure 4. represents the confusion matrix of the Decision Tree Classifier model.

II. Random Forest Regression:

We compared these predicted values with the actual values in the test component of the dataset. This comparison gave us the errors or residuals value which was in turn used to calculate the root-mean-squared-errors (RMSE). Table 1 describes the importance of various variables as calculated by the Random Forest Regression model.

Variables	Importance of Variables
passenger_count	10.25%
log_trip_distance	10.47%
pickup_day_number	19.34%
pickup_hour_quater	59.94%

Table 1: Importance of various variables with respect to tip percentage as given by Random Forest Regression model.

The root mean squared error (RMSE) for this model was 4.70. Therefore, we can conclude that random forest regression model was able to fit this dataset effectively.

VI. Conclusion & Future Work

The task of predicting tipping behaviour of passengers from trip characteristics without considering any specifications about other conditions, was very challenging. We created some new features that represent the trip characteristics in a more intuitive way. Average speed of trip, Trip duration, and Tip Percentage were some such features that we created. Upon analyzing the relationship between the trip characteristics, we discovered that most of the variables and created features are not correlated linearly. Therefore, we did not use any form of linear regression because it would not have fit the data correctly and would've produced incorrect untrustworthy results. We discretized our prediction variable into two buckets to use Decision Tree classification model which produced fair results considering the amount of information it was provided for training. We then used Random Forest Regression Model to fit a sample of our dataset. This model was able to produce good results along with providing us a value of importance for each variable with respect to calculating tip percentage. Using Decision Tree Classification model, we can categorize whether a trip would yield more or less than the median (16.67%) given the following trip characteristics: passenger count, trip distance, day of the week, time of the day. Using Random Forest Regressor on a sample of 10,000 randomly chosen trips run over 100 trees in the forest, we can predict the tip percentage with a root mean square error of 4.70 given the following trip characteristics: passenger count, trip distance, day of the week, time of the day. A good scope of future work on this paper by considering actual subjective opinion and passenger characteristics while trying to predict tipping behaviour of passengers. This task would entail a tedious task of data collection but might assist in backing our findings above.

Another scope of future work on this paper is to consider service characteristics alongside the trip characteristics to analyse the effect of different types of taxi services have on passengers.

VII. References

[1] "TLC Trip Record Data," Nyc.gov. [Online]. Available: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

[2] "Yellow Cab," Nyc.gov. [Online]. Available: https://www1.nyc.gov/site/tlc/businesses/yellow-cab.page.

[3] D. Elliott, M. Tomasini, M. Oliveira and R. Menezes, "Tippers and stiffers: An analysis of tipping behavior in taxi trips," 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, doi: 10.1109/UIC-ATC.2017.8397523.

[4] X. Zhan, S. Hasan, S. V. Ukkusuri, and C. Kamga, "Urban link travel time estimation using large-scale taxi data with partial information," Transp. Res. Part C Emerg. Technol., vol. 33, pp. 37–49, 2013.

[5] H. W. Chang, Y. C. Tai, and J. Y. J. Hsu, "Context-aware taxi demand hotspots prediction," Int. j. bus. intell. data min., vol. 5, no. 1, p. 3, 2010.

[6] B. Leng, H. Du, J. Wang, L. Li, and Z. Xiong, "Analysis of taxi drivers' behaviors within a battle between two taxi apps," IEEE Trans. Intell. Transp. Syst., vol. 17, no. 1, pp. 296–300, 2016.

[7] J. A. Deri and J. M. F. Moura, "Taxi data in New York city: A network perspective," in 2015 49th Asilomar Conference on Signals, Systems and Computers, 2015, pp. 1829–1833.

[8] Y. Dong, Z. Zhang, R. Fu, and N. Xie, "Revealing New York taxi drivers' operation patterns focusing on the revenue aspect," in 2016 12th World Congress on Intelligent Control and Automation (WCICA), 2016, pp. 1052–1057.

[9] Amber Raymond, B.S.W. and Cramer, K., 2020. Taxi Tipping in New York City (2014-2017): Reciprocity in Hailed vs. Dispatched Cab Fares. OpRpRp, p.61.

[10] Lee, W.K. and Sohn, S.Y., 2020. A large-scale data-based investigation on the relationship between bad weather and taxi tipping. Journal of Environmental Psychology, 70, p.10145

[11] *Mayor de blasio signs new law lowering New York city's default speed limit to 25 MPH* (2014) *The official website of the City of New York*. Available at: https://www1.nyc.gov/office-of-the-mayor/news/493-14/mayor-de-blasio-signs-new-law-lowering-new-york-city-s-default-speed-limit-25-mph#/0.

[12] *1.10. Decision Trees* (no date) *scikit-learn*. Available at: https://scikit-learn.org/stable/modules/tree.html.

[13] *Sklearn.Ensemble.RandomForestRegressor* (no date) *scikit-learn*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.