Executive Summary Report for TLC

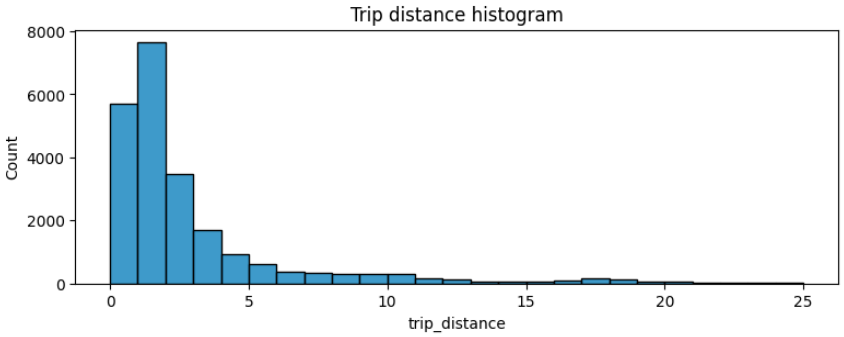
Overview / Problem

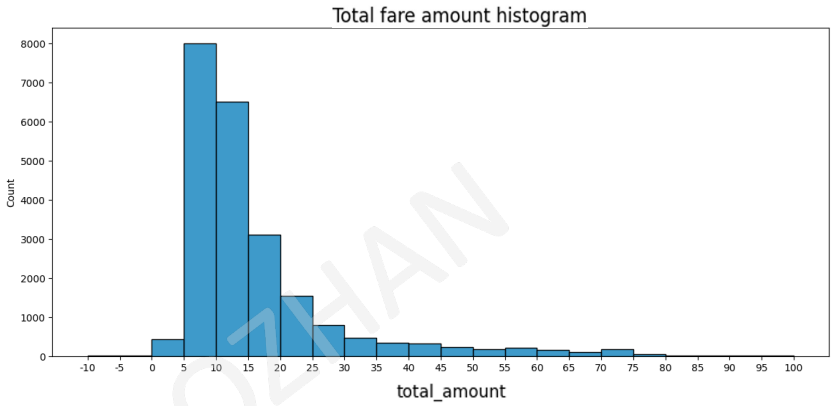
Before a regression model can be built, a preliminary inspection of data is needed. What insights the data could potentially inform us about is the main focus for generating meaningful results.

trip_distance	total_amount
33.96	150.30
33.92	258.21
32.72	179.06
31.95	131.80
30.83	111.38
30.50	119.31
30.33	73.20
28.23	62.96
28.20	70.27
27.97	63.06
27.88	89.16
27.34	88.56
27.20	88.55
26.86	94.75
26.54	86.76

Trip distances and total amounts in descending order

Key variables visualised





Initial Findings

- The information that this dataset contains is likely to satisfy the needs of a predictive model.
- Total amount and trip distance can be two of the key variables that could help build a predictive model for taxi fare as they should have direct influence on the pricing.
- Majority of trips were 1-3 miles and most taxi fares ranged between $5 and $15, as seen from the histograms above.
- Longest trips do not necessarily attract the highest fares as shown on the left handside table (See the provided notebook for more findings and details).

Next Steps

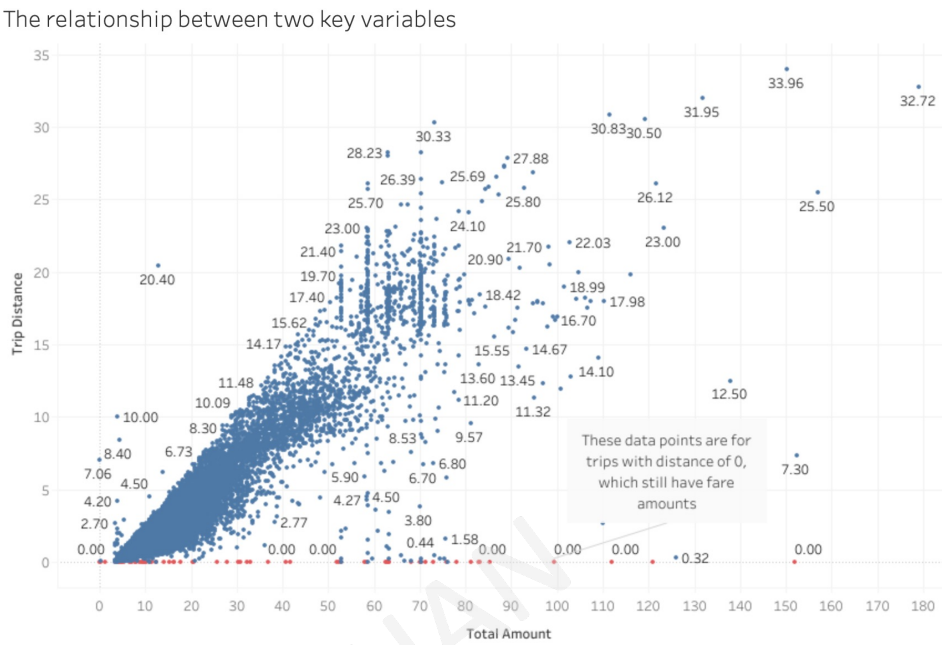
- Further EDA (Exploratory Data Analysis) will be conducted and the results will be interpreted.
- Data analysis and cleaning will be performed to get a deep understanding of the variables as well as any anomalies relating to them (e.g. outliers that could skew the data and results).
- Some variation between the total amounts for almost the same distances is also observed. Any potential outlier issue will be analysed further and imputations will be made where necessary.
- Descriptive statistics will be used to get deeper insight about the data so that it can be structured according to the needs of the models that are going to be built.
- Granularity of date/time data will be ensured and further analysis will be conducted to explore potential patterns.

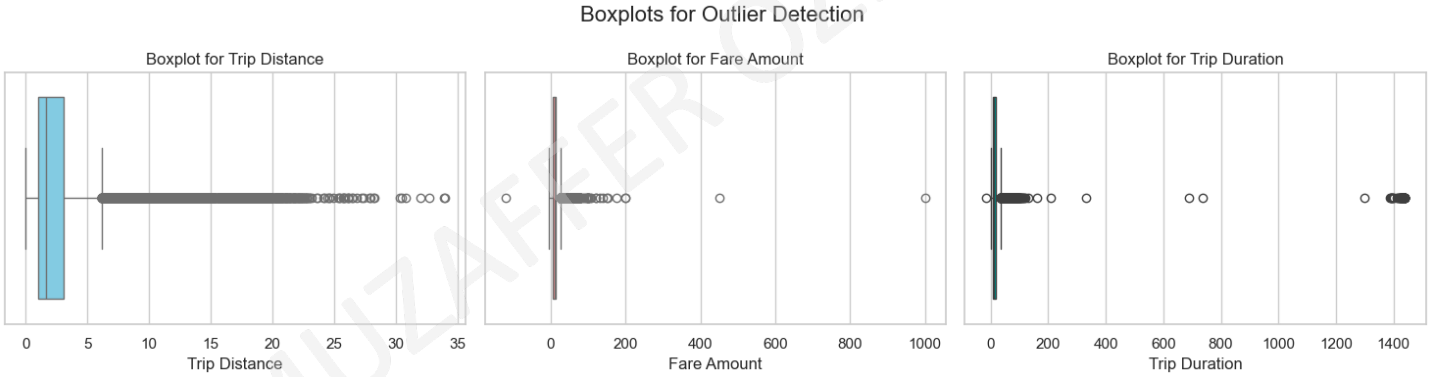
Overview / Problem

Before a model can be built, the data needs to be analysed, explored and cleaned. This stage is when initial actions regarding outliers are taken following the initial phase of exploratory data analysis.

It has now become evident that certain data entries may pose a hindrance to achieving accuracy in predictions of taxi ride fare amounts. Specifically, these are instances where the total amount is recorded, but the total distance is registered as "0" The analysis at this stage points to such occurrences as potential irregularities or data outliers that require consideration within the algorithm or potential removal.

Outlier Analysis





Solution

- The outliers in trip distance seem normal.
- Zero values in the distance column could have been easily removed as they did not make much sense. However, deeper analysis into data showed that they are actually values that were recorded with precision, meaning they were values close to 0 and then rounded down to 0. Nevertheless, they will be ignored considering their proportion (148) to the total sample size of 22699.
- Negative values in duration column are replaced with zero.
- Upper threshold for duration is set at 88.78 whilst fare amount is capped at $62.5 following a series of calculations (see the provided notebook for further details).

Next Steps

For any future models to be accurate and robust, any erroneous/duplicate data and outliers that could skew the data unreasonably should be analysed thoroughly. There should be a consistent and reasonable approach towards handling of any ungrounded outliers. For instance, with current data, we know that trips with the highest fare amounts are not necessarily the longest trips. How unusual data points are recorded in the first place can be investigated so that data collection process can be streamlined and any anomaly that could be a problem for future analysis and modelling can be prevented.

Overview / Problem

New York City Taxi & Limousine Commission needs a model to that predicts taxi fares. In this part of the project, whether type of payment method has an impact on the amount of a fare needs to be investigated, as at this stage, the focus is to explore how TFL can generate more revenue.

Solution

Whether there is a statistically significant difference between the total fare amounts paid by credit card or cash will be investigated running a T-test.

Results Summary

The two-sample t-test conducted reveals the statistically significant difference in the average total fare amount between riders paying by credit card and those paying in cash.

When more customers pay by credit card, revenues could hypothetically increase. But we should note that there could be other potential factors causing this difference.

Descriptive Statistics and T-test Outcomes

Details

Payment Method	Average Fare Amount
Credit Card	$13.43
Cash	$12.21

Average Fare Amount by Payment Method

Descriptive statistics reveal that the average amount of fare payment differs by $1.22. To validate that this difference is not just by pure chance, a two-sample t-test is conducted:

T-statistic	P-value	Degrees of freedom
6.87	0	16675.49

T-test scores

It is evident from the T-statistic that there is a great difference between two groups. With p-value of 0, we confirm that, the occurrence of this difference is not by chance so there is a statistically significant difference in the mean of the two groups.

Next Steps

If customers are encouraged to make their taxi fare payments with credit cards, TLC's revenues can be boosted. As a way of promotion, taxi drivers can advise the customers that the preferred payment method is credit card. Similarly, signs that show credit card as preferred payment method can also be placed in the car.

Overview / Problem

New York City Taxi & Limousine Commission needs a model to predict fare amounts for rides. The deliverable for this request is a regression model.

Solution

A multiple linear regression (MLR) model is built, leveraging the characteristics and distribution of the provided data. The MLR model demonstrated its effectiveness in accurately predicting taxi fares before the ride commences.

The model exhibits strong performance on both the training and test datasets, indicating a well-balanced model that avoids being over-biased and overfit. The model's performance was even better when evaluated on the test data.

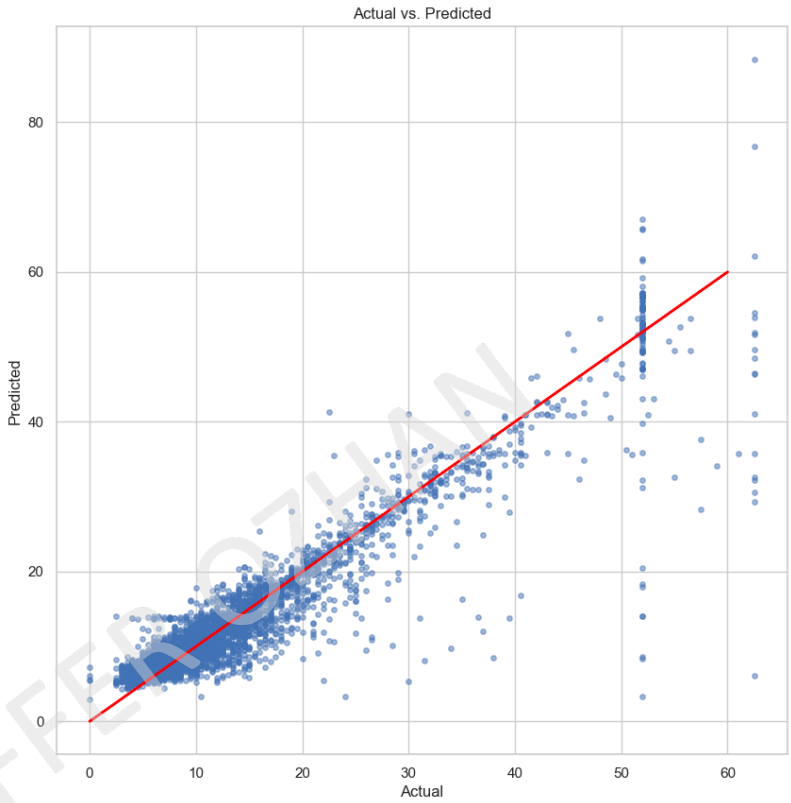
Results Summary

Addressing outliers through imputation resulted in model optimisation, notably with respect to the fare amount and duration variables.

The linear regression model offers a robust foundation for estimating taxi ride fares with accuracy.

Regression Model Outcomes

Details



- As high as 87% of the variance in the target variable (fare_amount) is explained by the model as evident from the R^2 of 0.87. *(Further analysis and insights available in the provided notebook)*
- Error metrics are also at acceptable levels:

  (MAE: 2.1 MSE: 14.36 and RMSE: 3.8)

	passenger_count	mean_distance	mean_duration	rush_hour	VendorID_2
0	0.030825	7.133867	2.812115	0.110233	-0.054373

Regression coefficients

Regression coefficients reveal that for every 1 mile travelled, fare amount increased by a mean of $2 (after standardised data converted back to miles). Similarly, for each additional unit of mean duration, the fare amount is expected to increase by around 2.80 units.

Collecting additional data for routes that are currently under-represented in the dataset is likely to enhance the model's accuracy and coverage. Also, location coordinate data in format of latitude and longitude can help design an improved model.

The insights gleaned from this analysis can be harnessed by the New York City Taxi and Limousine Commission to develop a mobile application, empowering TLC riders to access estimated fares prior to the commencement of their rides.

The model offers a robust and dependable fare prediction capability, making it a valuable resource for subsequent modelling endeavours.

Overview / Problem

New York City Taxi & Limousine Commission needs a machine learning model that predicts whether a taxi rider will be a generous tipper.

Solution

Two different modeling architectures (Random Forest and XGBoost) are used. Both models performed successfully and yielded accurate predictions regarding when the tip will be a generous one (>20%). Both models can be used by taxi drivers and further feedback can be received.

Results Summary

The resulting model is practical for identifying passengers who are likely to give generous tips, demonstrating reasonably high scores in terms of precision, recall, F1 score, and overall accuracy. For further recommendations, please consult the "next steps" section. More details can also be found in the provided notebook.

Future model suggestions

Past tipping behaviour can be included in the model, as those patterns could improve predictions, for which data collection at user-level and driver-level is needed.

Riders' behaviours to tip can and will evolve constantly due to external factors. Making an RF or XGBoost model fetch and fit new data on a real-time basis can help model keep maintain its high predictive power, which could otherwise get lower as the gap between the time of the historical data and present time widens.

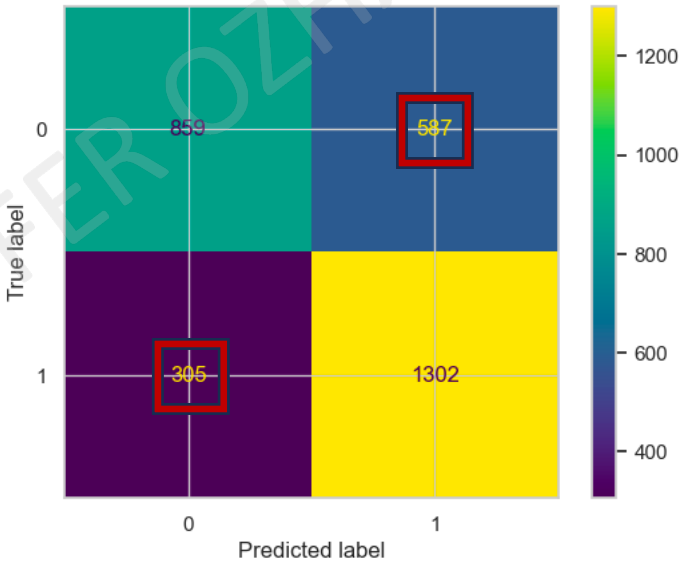
This will require careful consideration of data quality, model stability and computational resources. Also, it is essential to strike a balance between keeping the model updated and ensuring it continues to provide accurate predictions.

Machine Learning Model Outcomes

Details

- A trip's itinerary, predicted fare amount and time of day are assumed to have a relationship with tip amount.
- In line with the assumption, two models are built and fit to the data and tests are performed. It is clear from the below results that these factors allow us to predict tipping as both models have F_1 scores of over 0.74.

	model	precision	recall	F1	accuracy
0	RF CV	0.692214	0.813474	0.747954	0.711432
0	RF test	0.690021	0.808961	0.744772	0.708156
0	XGB CV	0.691387	0.812694	0.747141	0.710449
0	XGB test	0.689254	0.810205	0.744851	0.707828



The confusion matrix of XGBoost model above reveals that the model exhibits a higher probability of making false positive predictions compared to false negatives, making Type I errors more prevalent. This situation is suboptimal, as it is more favourable for a driver to experience a positive surprise, receiving a generous tip unexpectedly, rather than encountering disappointment when they had anticipated a generous tip. Nevertheless, the model's overall performance remain acceptable.

Next Steps

Either of the models can be shared with TLC and be used as a predictive model for tip amounts. Notably, potential inclusion of user-level and driver-level data into the model will significantly improve its predictive power.