NYC Taxi Tipping Pattern Ananlysis

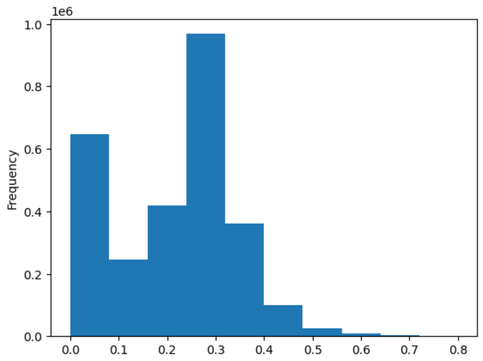
Navaneethakannan Arumugam; Nihar Madasu; Archana Bhusara; Yiyuan Wang

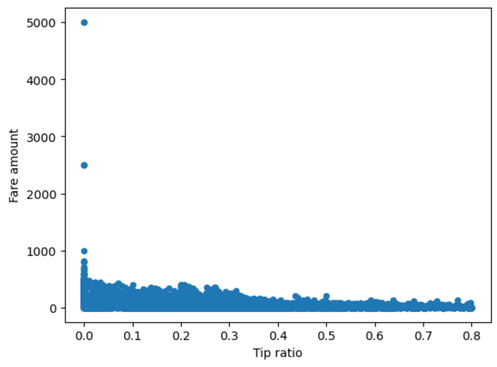
INTRODUCTION

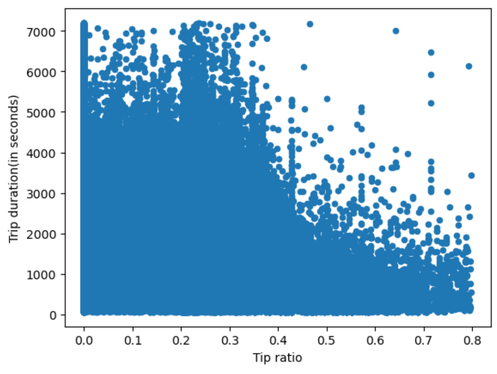
Our project aims to build a comprehensive model to analyze tipping behavior in urban transportation systems, more specifically the New York City Taxi system. Encompassing variables such as trip details, fare amounts, tipping amounts, weather conditions, demographic attributes, and socioeconomic indicators, 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 aforementioned 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.

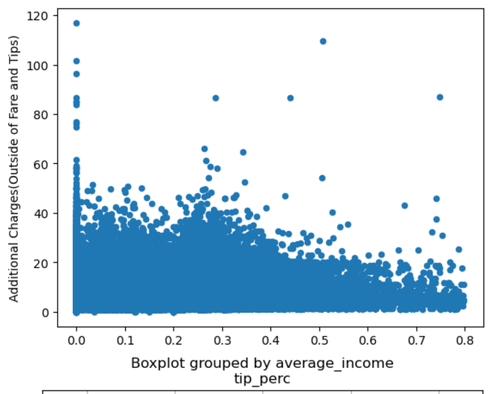
Exploratory Data Analysis

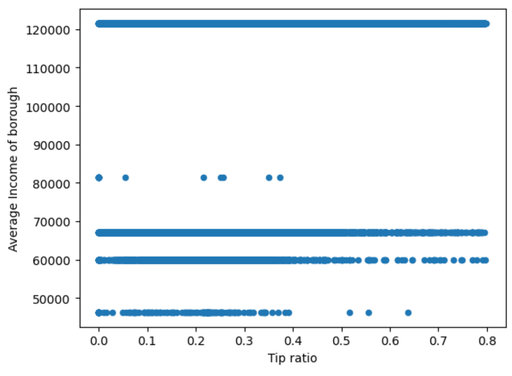
We performed EDA on the variables to assess their relationship with the predictor using correlation plots and box plots. We got the following results:

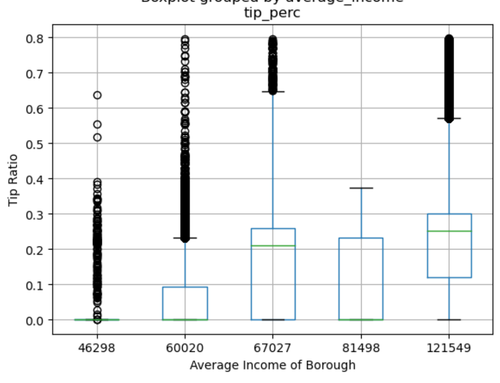


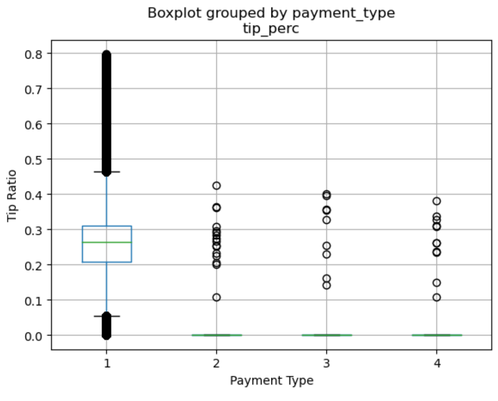












EXPERIMENTS AND RESULTS

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 RandomForestRegression, 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

CONCLUSION

By applying feature selection techniques like LASSO regression, we were able to find features from the dataset that had the strongest correlation with the response variable i.e. tip ratio. Since the r2 values of both models, with and without using non-selected variables from LASSO regression, we can conclude that the rest of the variables have a very negligible relationship with the response variable.

The top variables we were able to find through our research were:

Type of payment – A code associated with the payment type used for the taxi trip.

Fare amount – The overall fare amount for the taxi trip

Trip duration – The duration of the trip, converted to seconds

Additional charges – Charges outside of fare and tip that may be included in the trip, like airport fee and mta tax

Average Income of Borough – The average income of the borough where the trip is initiated

Although we were not able to find a strong correlation between the predictors and response, with the r2 score coming out to around 0.47 for linear regression, we were able to pick out some variables that can be interpreted to have some form of relation with the tipping behavior of a customer in a taxi trip.

We were able to achieve better r2 score using Random Forest Regression on training data, though this model seemed to be slightly overfitted and didn't perform as well on test data.