Predicting Taxi Tips in New York City

Overview

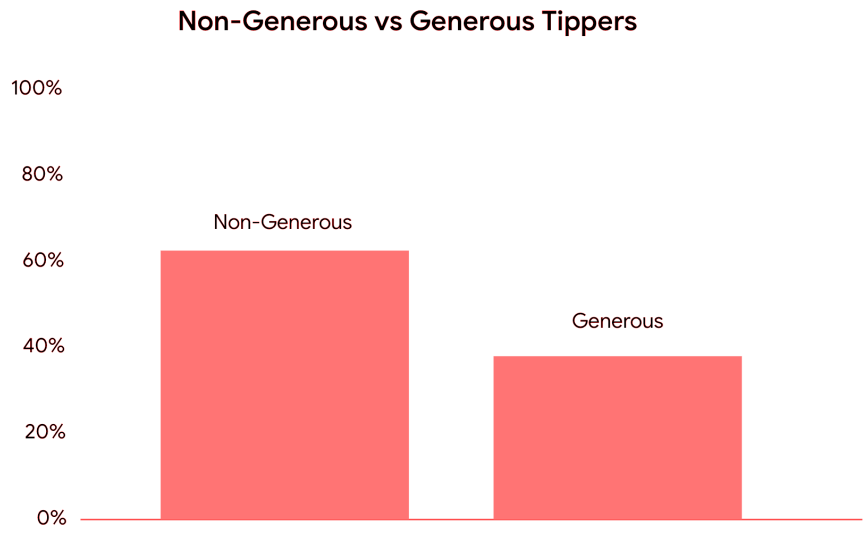
Using multiple linear regression and random forest methods, the purpose of this project was to predict whether a client would pay or not pay a gratuity during yellow cab taxi trips in NYC in 2017. The accuracy was at 86% and the precision was at 72% in the ultimate random forest model which showed the most important features that determined tippers from non-tippers. The tipper was at over 20% and the non- tipper at below 20% were shaped by these variables: duration, distance, and cost of trip.

Business Understanding

For a NYC taxi driver, the average salary is estimated to be $45,000, as listed on salary.com. The median monthly rent in NYC is $6,500 which is quite higher than a person can afford on an average taxi driver's salary. This is crucial to understand why it is important to encourage riders to leave a tip so that drivers can earn a living wage.

Data Understanding

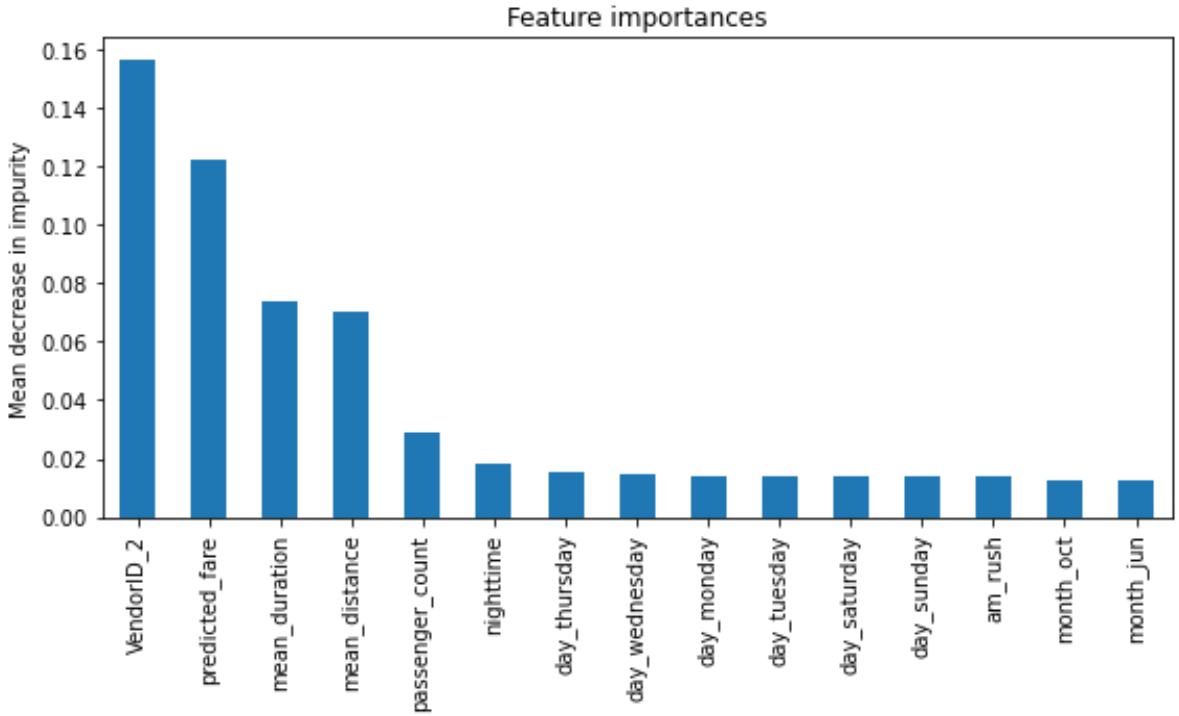
The data for this analysis is from NYC.gov's NYC Taxi and Limousine Commission and is made up of 408,000 trips and 18 features, which are created of details such as trip duration, destination, vendor, toll information, and payment method. The data breakdown of generous to not so generous tippers are shown in the bar chart below.



To determine if a ride occurred during rush hour, a feature was engineered. Data was cleaned to remove duplicates and reformatted to correct data types.

Modeling and Evaluation

To determine the feature importance of who would generously tip or not, a random forest model of 100 decision trees was created. The top three important factors that would determine a tipper from a non-tipper were VendorID, predicted_fare, and mean_duration, as shown in the chart below. The model was highly predictive with the F1 score at 0.723 and with an overall accuracy of 0.685. 78% of the data was correctly identified, which is better than 48% random guesses.



Conclusion

This model has a high predictive rate and can assist taxi drivers to anticipate the high chance of being tipped or not. It would be recommended to create a parametric model to consider which variables would contribute to the actual tip amount. Future analysis of increasing a rider's tipping history could help the stakeholder with this business problem.