Predicting Taxi Gratuities in New York City

Project Summary

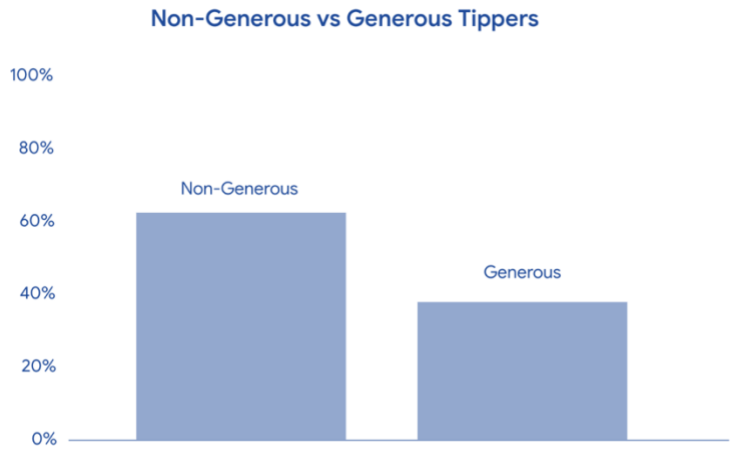
The objective of this initiative was to develop predictive models using multiple linear regression and random forest techniques to ascertain the likelihood of receiving substantial gratuities from taxi riders. Analyzing 2017 data from yellow taxi trips in New York City, the study found that trip length, distance, and fare amount were key indicators in predicting whether a rider would be a large tipper (over 20%) or not (under 20%). The robustness of the random forest model was evident with an accuracy of 86% and a precision of 72%.

Business Context

Considering the average taxi driver in New York earns about $45,000 annually, as reported by salary.com, and faces median monthly rent costs of $6,500, understanding the dynamics of tipping is crucial for drivers striving for a sustainable income.

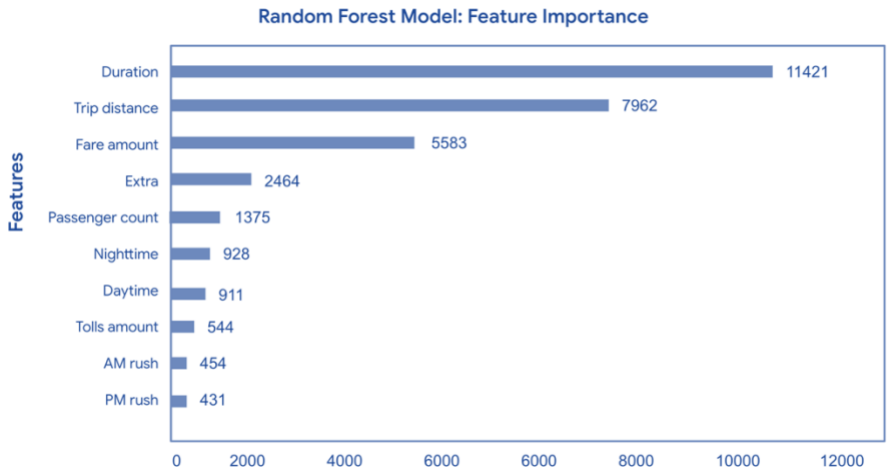
Data Insights

The research utilized data from the NYC Taxi and Limousine Commission, comprising roughly 408,000 distinct trips with 18 different attributes, accessible through NYC.gov. This data set was refined by removing non-essential columns and converting data into the correct formats. Figure below shows the proportion of generous tippers and non-generous tippers.



Model Development and Results

A random forest model with 100 trees was crafted to identify the most predictive factors for high tipping. As visualized by the plot below, the model's performance was strong, with key indicators being trip duration, distance, and fare cost.



Conclusion

The findings provide insights for taxi drivers on the likelihood of receiving significant tips. Further analysis, such as a parametric model, could refine understanding of how different variables affect the tipping amount. Future improvements could include incorporating data on individual riders' previous tipping habits to better support the financial well-being of taxi drivers.