Machine Learning Model Outcomes

Executive summary report for the New York City Taxi and Limousine Commission

Prepared by Automatidata

Overview

New York City Taxi & Limousine Commission has contracted the Automatidata data team to build a machine learning model to predict whether a NYC TLC taxi cab rider will be a generous tipper.

Problem

After rejecting the initial modeling objective (predicting non-tippers) out of ethical concern, it was decided to predict "generous" tippers—those who tip ≥ 20%. This decision was made to balance the sometimes competing interests of taxi drivers and potential passengers.

Solution

The data team used two different modeling architectures and compared their results. Unfortunately, neither approach delivered strong predictions. As a result, the team would recommend using this model as a tool to derive deeper business insight, or at best a very rough guide to taxi drivers. The next steps section offers suggestions for additional analysis that could improve the usability of the results.

Details

Behind the data

- The data team's assumption was that a trip's itinerary, predicted fare amount, and time of day may have a strong enough relationship with tip amount that we could accurately predict generous tipping.
- After the data team built the identified models and performed the testing, it became clear that there was not as strong a correlation as anticipated, with an F_1 score of just 0.350.

model	precision	recall	F1	accuracy
RF CV	0.452301	0.271088	0.338523	0.631182
RF test	0.461000	0.277000	0.346000	0.635000
XGB CV	0.445599	0.299513	0.357846	0.625696
XGB test	0.457000	0.284000	0.350000	0.633000

F1 scores for random forest and XGboost models

Results Summary

The resulting algorithm is usable to predict riders who might be generous tippers, but has serious limitations to its usability in informing business decisions.

Future model suggestions

- Collect/add more granular driver and user-level data, including past tipping behavior.
- Cluster with K-means and analyze the clusters to derive insights from the data.

Next Steps

As a next step, the Automatidata data team can consult the NYC TLC to share the model results and recommend that the model could be used as an indicator of tip amount with a very high degree of uncertainty. However, additional data is needed to realize significant improvement to the model.