Executive summary report for the TLC

Prepared by Wei Zhang

Overview

TLC has contracted the data team to build a machine learning model to predict whether a 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. Both models performed acceptably, with a random forest architecture yielding slightly better predictions. As a result, the team would recommend beta testing with taxi drivers to gain further feedback.

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 is clear that these factors do indeed help predict tipping. The model's F_1 score was 0.7235.

	model	precision	recall	F1	accuracy
0	RF CV	0.674919	0.757312	0.713601	0.680233
0	RF test	0.675297	0.779091	0.723490	0.686538
0	XGB CV	0.673074	0.724487	0.697756	0.669669
0	XGB test	0.675660	0.747978	0.709982	0.678349

Image Alt-Text: F1 scores for random forest and XGboost models

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

Results Summary

The resulting algorithm is usable to predict riders who might be generous tippers, with reasonably strong precision, recall, F_1, and overall accuracy scores.

Next Steps

As a next step, the data team can consult the TLC to share the model results and recommend that the model could be used as an indicator of tip amount. However, additional data would be needed to realize significant improvement to the model.