Machine Learning Model Outcomes

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

Overview

Designing an efficient machine learning model to predict whether a customer will tip generously for a ride or not for NYC Taxi & Limousine Commission.

Problem

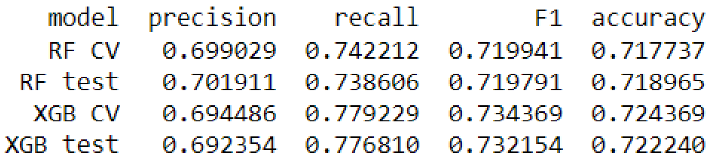
We as tasked to predict “generous” tippers—those who tip ≥ 20%. Cab drivers depend on generous tip to have a living wage.

Solution

We use train two different machine learning models for predicting whether a customer will be generous or not. After testing the performance of the two models, we pick the one that is better and efficient, in this case XGBoost.

Details

- Trip’s itinerary, predicted fare amount, and time of day were assumed to be the important features affecting tipping. New feautures were engineering based on these information too.
- The XGBoost model’s F_1 score was 0.732 which shows that the features provided and engineered are indeed impacting tipping.
- VendorID was found to be the most important feature in both models.



```
      model  precision    recall        F1  accuracy
      RF CV   0.699029  0.742212  0.719941  0.717737
    RF test   0.701911  0.738606  0.719791  0.718965
     XGB CV   0.694486  0.779229  0.734369  0.724369
   XGB test   0.692354  0.776810  0.732154  0.722240
```

F1 scores for random forest and XGBoost models

Future model suggestions

- Collect/add more granular driver and user-level data, including past tipping behavior.
- Identify key differences in different VendorID types given how it hugely impact tip amount.

Results Summary

The resulting algorithm has reasonale prediction capabilities for predicting whether a rider would generously tip or not. Our XGB model performamce was also found to be decent on test data seen in table above.

Next Steps

Moving forward, the Automatidata data team could engage with the New York City Taxi and Limousine Commission to present the model outcomes and propose its potential utility as a predictor of tip amounts. Nonetheless, enhancing the model would necessitate acquiring supplementary data for substantial improvement.