NYC TLC Machine Learning Model Summary

Prepared by Automatidata

Overview

- To create a suite of models and algorithms to improve the operations efficiency and revenue of the NYC TLC and its drivers.
- Initially, the proposal was to create a model that identified those who would not tip. Due to ethical concerns, this exclusive objective was rejected in favor of an inclusive one.

Objective

Of this phase, to create a model that predicts those riders who are more likely to tip generously, defined as >= 20% of total amount less tip amount.

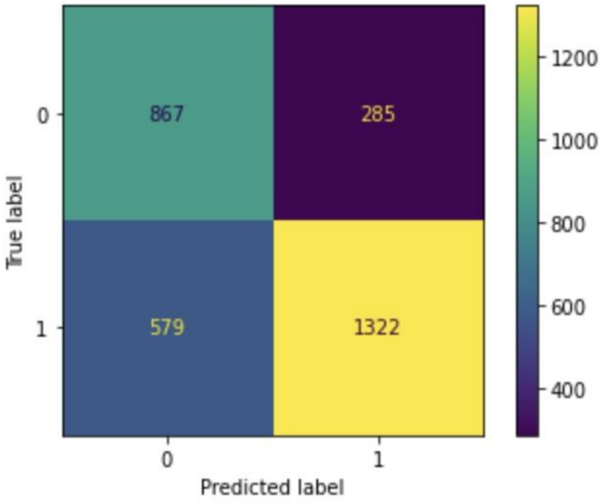

Fig 1. Confusion matrix of the XGBoost model

Results

- Multiple models were created, and the best performing model was using XGBoost methodology. How it arrives at these specific conclusions are unclear, but two features stand out as containing the most predictive signal.
- The model correctly predicts 82.3% of those riders who will tip generously based on these features. Overall, the model was 71.7% accurate in its classifications.
- The model had about double the number of false negatives than false positives, meaning this model favors taxi drivers by minimizing lost revenue, but excludes more riders who would tip generously.
- The model was fit to the f1 metric, balancing precision and recall. This was done because the stakes of a FP and a FN are relatively balanced. This model scored a 75.4%.

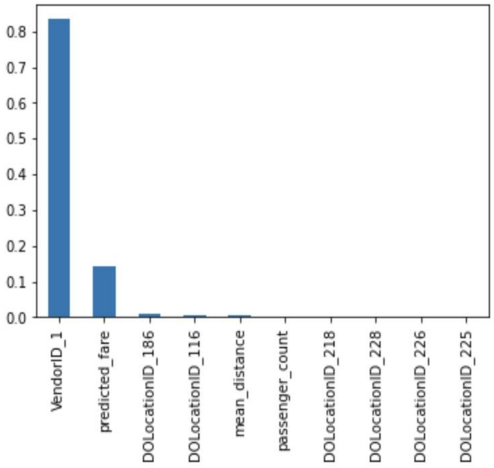

Fig 2. Plot of top 10 feature importances of XGBoost model

Next Steps

- Test the model with a select group of taxi drivers to determine its effectiveness in the field.
- Begin tracking cash tips, tipping history and riding history. These would add a significant amount of relevant data, improving predictive performance of any future models.
- Use caution with its implementation. This model by design excludes some riders from the service, and this model is not perfect. Many riders who depended on the service may no longer be picked up.