Title # New York City TLC Machine Learning Project

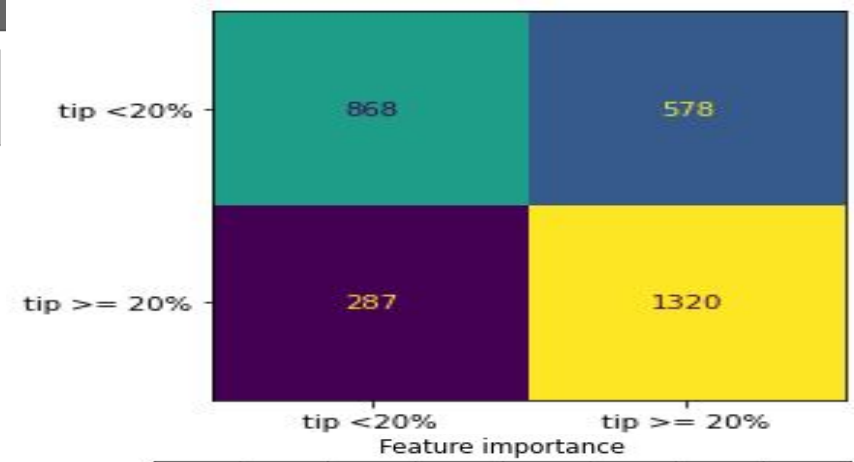
Project Overview

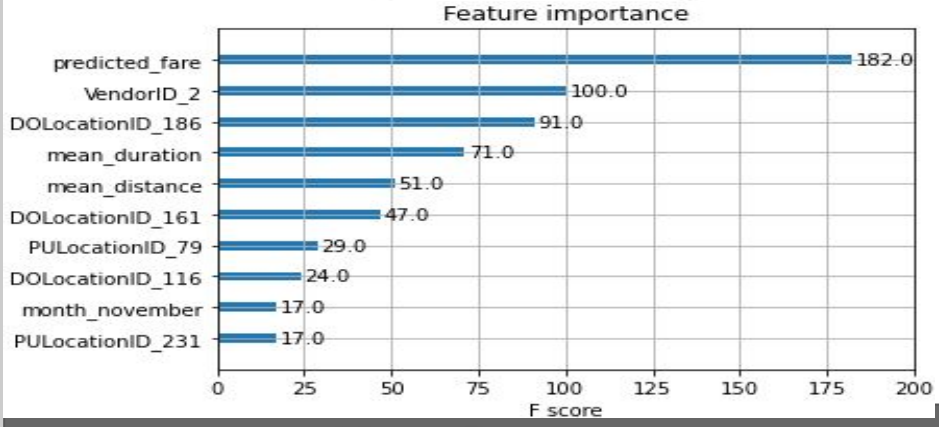
This project goal is to develop the machine learning model(s) to predict whether riders will be the generous tippers (by tipping 20% or more). By successfully developing the model has the good performance and predictions, we'll determine which features might be best influence to whether the riders will give a tip. Also, it'll be tremendously helpful to the drivers to boost their revenue.

Key Insights

- Both the **Random Forest** and **XGBoost** models were developed. **XGBoost** was chosen as the best model with a small margin higher in metrics. Specifically, the F1 score (main metric score) of **XGBoost** is higher than the **Random Forest** model (0.753 vs. 0.752 respectively).
- From the given test data for the model to predict, there are 1607 riders who give a tip >=20%, and 1446 riders with <20%.
- Of those 1607 riders, the model correctly predicts 1320 (82%) and 287 incorrect (18%). With 1446 riders, the model corrects 868 (60%) and 578 incorrect (40%) (refer to the matrix plot).
- From the model's incorrect predictions, it might be less desirable to account for the higher proportion of the riders will leave tip <20% but the drivers will expect >=20%, and that will lead to the frustration and disappointment.
- On the other hand, the proportion of model predicts the riders will tip <20%, but it's >=20% could be acceptable. This will make the drivers surprised by the unexpected tip they could receive that they wouldn't know beforehand.
- On the plot is the top 10 features that accounted for the most influence to the target variable. This **XGBoost** model, as well as the **Random Forest** are powerful and highly accurate in predictions. However, they won't be easy to explain (i.e. how the model predicts). For instance, we have no clue on how these features account for the most important to the target variable.

Details





Next Steps

- Even though there might be concerns of model's incorrect predictions, there wouldn't account too much in severity, and can be acceptable.
- We might recommend the New York TLC to deploy this XGBoost model. Continuing monitoring how its works, and get the feedback and evaluation from the drivers for further improvement and enhancement.
- The good approach is to request the NYC TLC to collect more rides data to boost the model's performance.