Machine Learning Model Results for TLC Data

Executive Summary from Automatidata

Overview

New York City TLC stakeholders have been impressed with the data analytical work completed by the Automatidata team in this project. As a result, they have reached out once again for assistance in creating a machine learning model that can help predict whether or not a rider will be a generous tipper.

Objective

- Build a random forest model from the New York City TLC dataset
- Build an XGBoost model from the New York City TLC dataset
- Evaluate both models

Results

- For the random forest model:
  - Best hyperparameter combination: max_depth = None, max_features = 3, max_samples = 1.0, min_samples_leaf = 4, min_samples_split = 2, n_estimators = 200
  - Test data scores: precision = 0.526575, recall = 0.998755, F1 = 0.689581, accuracy = 0.526695
- For the XGBoost model:
  - Best hyperparameter combination: colsample_bytree = 0.7, learning_rate = 0.2, max_depth = 4, min_child_weight = 5, n_estimators = 200, subsample = 0.7
  - Test data scores: precision = 0.584438, recall = 0.630989, F1 = 0.606822, accuracy = 0.569604

Next Steps

- Add more values into hyperparameter tuning for more optimized hyperparameters in both models.
- Utilize the XGBoost model version to predict generous tippers.