Executive Summary of Regression Model Building for New York TLC Data

Commission Prepared by Automatidata

Project Overview

The NYC Taxi & Limousine Commission has consulted with Automatidata to build a random forest model to predict taxi cab tips. The Automatidata team built a random forest classification model with a 70.2% f1 score on the test data.

Key Insights

EDA

- Feature columns had data types converted
- Feature engineering to produce new features
- Target produced from data transformation

Preprocessing

- Irrelevant features were removed
- Categorical columns were encoded

Modeling

- A random forest model and a xgboost model were trained on the dataset with hyperparameter tuning performed with GridSearchCV.
- Random forest model was declared the winner with better overall scores.

Challenges

- Need to balance ethics between disappointing drivers and fair customer representation

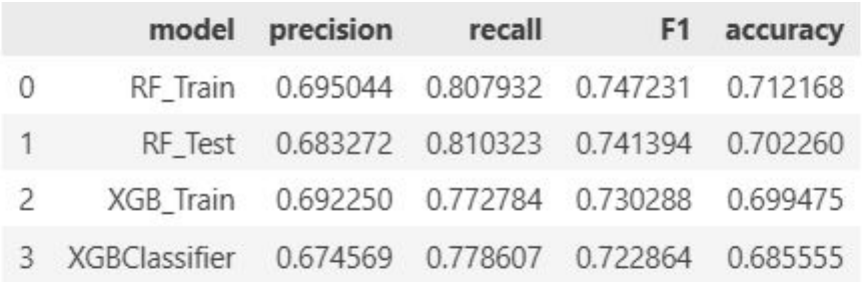
Details

EDA and Preprocessing

- The dataset concatenated with another dataset
- Target produced from transformation applied the dataset features.
- Irrelevant columns and columns that would cause data leakage were removed.

Modeling

- A train and test set were created were 80% of the dataset was used for training
- Categorical variables were encoded.
- Visualizations of feature importances



	model	precision	recall	F1	accuracy
0	RF_Train	0.695044	0.807932	0.747231	0.712168
1	RF_Test	0.683272	0.810323	0.741394	0.702260
2	XGB_Train	0.692250	0.772784	0.730288	0.699475
3	XGBClassifier	0.674569	0.778607	0.722864	0.685555

Next Steps

- Encourage TLC to produce app with random forest model as f1 score is stable
- Encourage TLC to gather more features representative of target (customer surveys, events for specific time periods)