

Searching for Ensemble Methods in Credit Rating System

Atacan Kavdır & Sinem Balkuvar

Retail Credit Risk Departments in Banks seek an automatic system to give credit approval or rejection decisions. Statistical models come into play to meet the need. However, success may not be always the outcome. Model's predictive performance is highly important in reaching it. Here lies the project's main goal. The project compares and contrasts several advanced Machine Learning algorithms to reach high performance. Higher predictive performance of the ensemble models may be at the expense of interpretability of the models. However, the advanced algorithms will provide insights for the business line by putting forth the most important features in better predicting default probability.

Pipeline

After data wrangling and feature engineering, we went feature elimination process with 150+ features for each datasets(i.e. Bank Customer and Credit Bureau Data)

With RFE method, we ended up with ~15 features for each dataset

ROC-AUC score is used as the metric to evaluate overall performances of the trained classification algorithms

Cross validation and hyper parameter tuning methods implemented to find and select the best performing models

Soft Voting is applied to achieve higher performance with ensembles than individual algorithms for each data set

Two datasets are combined with voting algorithm and reached to the higher performance metric

Model-I

Bank Customer Data

Feature Engineering

Null Treatment

Feature selection

Correlation Check

Model Trials

Grid Search

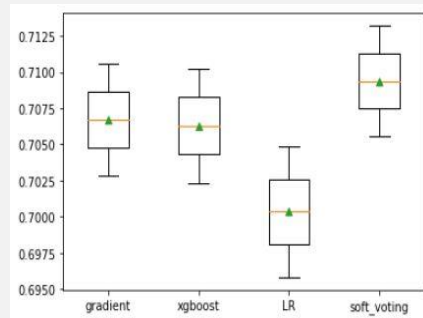
Cross Validation

Gradient Boosting

XGBoost

Logistic Reg

Soft Voting



CV scores of different models

ROC AUC

Train
74.1

Test
71.9

Model-II

Credit Bureau Data

Feature Engineering

Null Treatment

Feature selection

Correlation Check

Model Trials

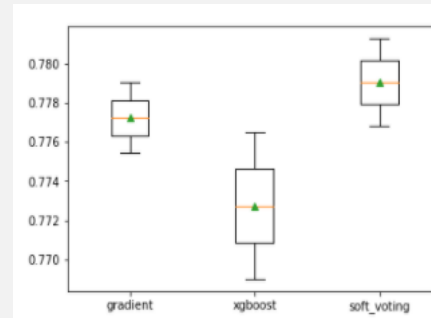
Grid Search

Cross Validation

Gradient Boosting

XGBoost

Soft Voting



CV scores of different models

ROC AUC

Train
82.8

Test
75.9

Ensembles

Model 1 Bank Customer Data Based Voting Model

- Gradient Boosting, XGBoost and Logistic Regression algorithms are used with their best performing hyperparameters.
- After evaluation of train and test scores of both individual and voting algorithms, final model of the dataset decided to be ensemble voting algorithm of 3 models.

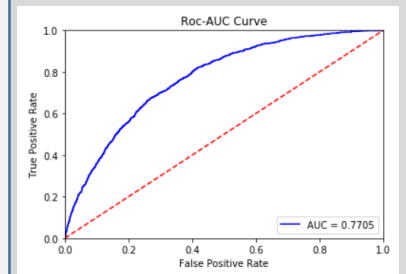
Model 2 Credit Bureau Data Based Voting Model

- Numerous ML Algorithms have been trained for the model. As the inputs of the final model Gradient Boosting & XGBoost algorithms are used with their best performing hyperparameters.
- After evaluation of both individual and voting algorithms, final model of the dataset decided to be ensemble voting algorithm of 2 models.

Final Model

Final Model Ensemble of Model 1 and Model 2

- Yet another voting algorithm is used to combine the two ensemble model based on two datasets, reaching the higher performance metric, i.e. AUC score of 77.1
- In the training process of final model, soft voting applied.



Test
77.1

