BINARY CLASSIFICATION USING TAIWANESE BANKRUPTCY PREDICTION DATESET

Jakub Dziurka, Paweł Hermansdorfer

PROJECT OBJECTIVE

Main Goals:

- Analyze Taiwanese Bankruptcy Prediction dataset,
- Build a set of classifiers and select the best one.

Scope of Work:

- Understand the dataset,
- Select appropriate classification models,
- Apply proper validation techniques and optimize metrics for evaluation,
- Find the best approach to solve the classification problem.

UNDERSTANDING THE DATASET

Dataset Overview

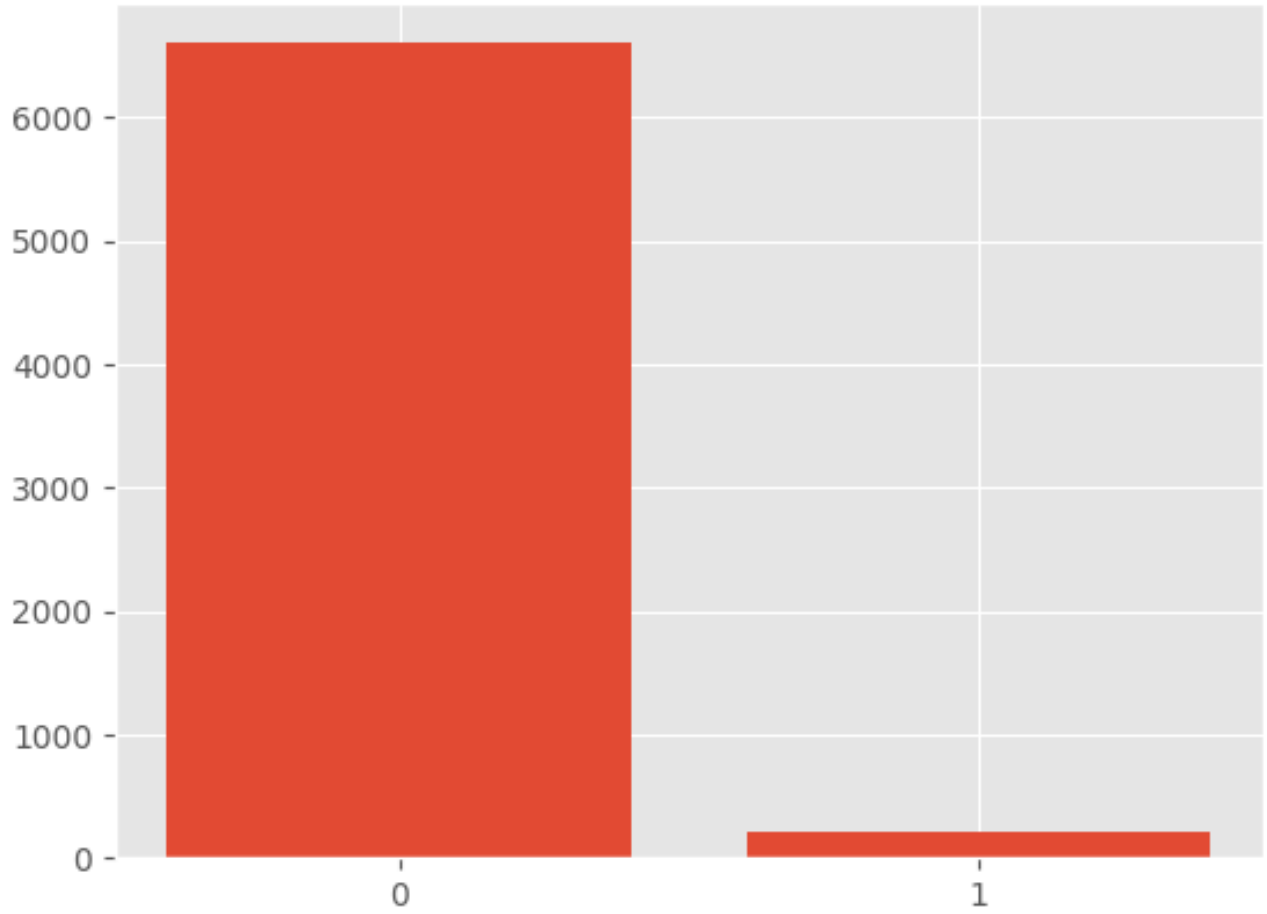
- The dataset contains 6819 records with 95 features related to financial metrics.

Key Observations

- The dataset is highly imbalanced, with significantly fewer bankrupt companies.

Preprocessing Steps

- Perform feature selection to remove irrelevant variables.



Did the company go bankrupt?

UNDERSTANDING THE NATURE OF THE DATA

Key observations:

- The original distribution, the training distribution and the test distribution show a strong class imbalance.

This imbalance must be addressed to ensure the model is evaluated fairly and does not favor the majority class.

Training distribution: *0: 4619, 1: 154;*

Test distribution: *0: 1980, 1: 66;*

Original distribution: *0: 6599, 1: 220;*

PRECISION AND ACCURACY IN IMBALANCED DATASETS

Model:

- *KNeighborsClassifier(n_neighbors=5)*

Traditional accuracy metrics fall short when dealing with imbalanced datasets. To address this, metrics like precision, recall and F1 score are used. They offer a deeper insight into how well a model performs on both classes.

***Accuracy**: 96.52981427174976*

***Precision**: 27.27272727272727*

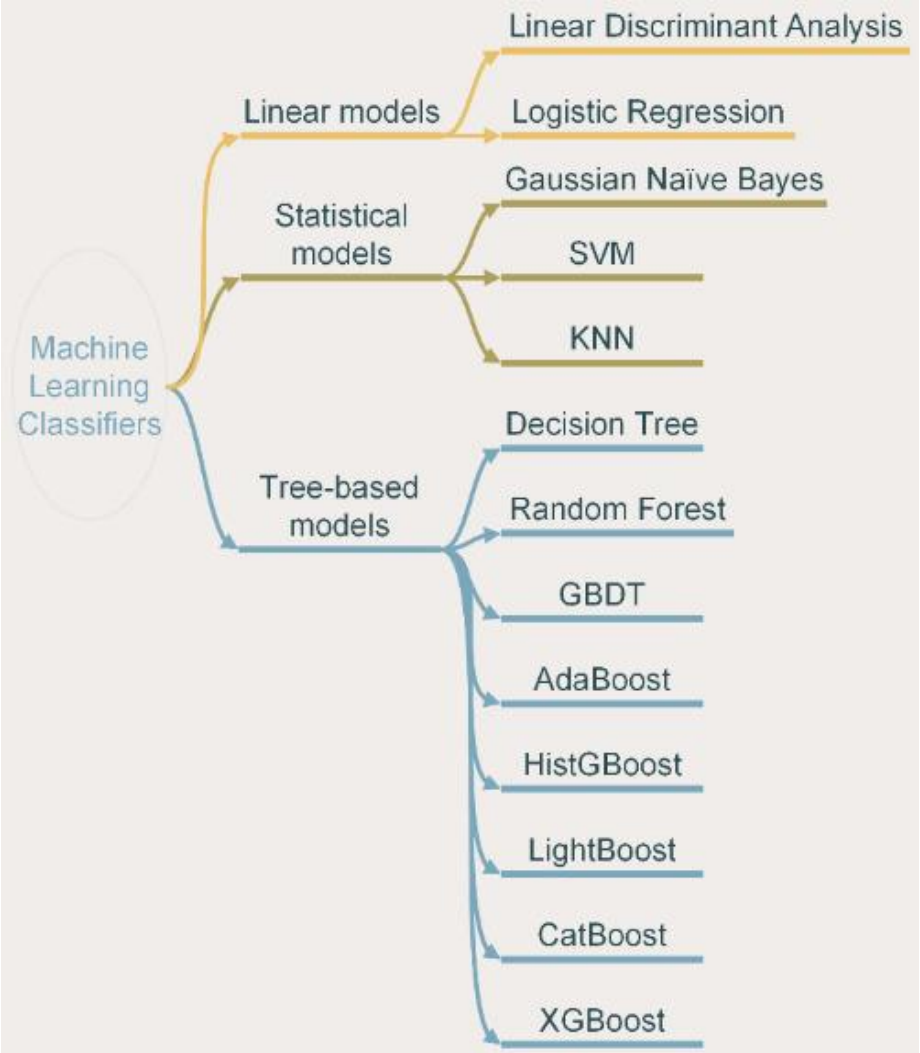
***F1-score:** 7.792207792207792*

DIFFERENT TECHNIQUES TO ADDRESS THE CLASS IMBALANCE ISSUE

TECHNIQUE	ACCURACY	PRECISION	RECALL	F1-SCORE
Normal	*96.52981427174976*	*27.27272727272727*	*4.545454545454545*	*7.792207792207792*
Random Over-Sampling	*93.10850439882698*	*19.51219512195122*	*36.36363636363636*	*25.39682539682539*
SMOTE	*89.05180840664711*	*15.94827586206896*	*56.06060606060606*	*24.83221476510066*
K-Folds Cross-Validation (5 splits)	*96.43206256109482*	*18.18181818181818*	*3.030303030303030*	*5.194805194805195*

Resampled dataset shape: 0: 4619, 1: 4619

THE FOLLOWING CLASSIFICATION MODELS THAT ARE USED:



LinearDiscriminantAnalysis

Normal	*0.33870967741935487*
Random Over-Sampler	*0.2874251497005988*
SMOTE	*0.29411764705882354*
K-Folds cross validation	*0.32786885245901637*

CatBoostClassifier

Normal	*0.3478260869565217*
Random Over-Sampler	*0.4320000000000005*
SMOTE	*0.4507042253521127*
K-Folds cross validation	*0.3448275862068965*

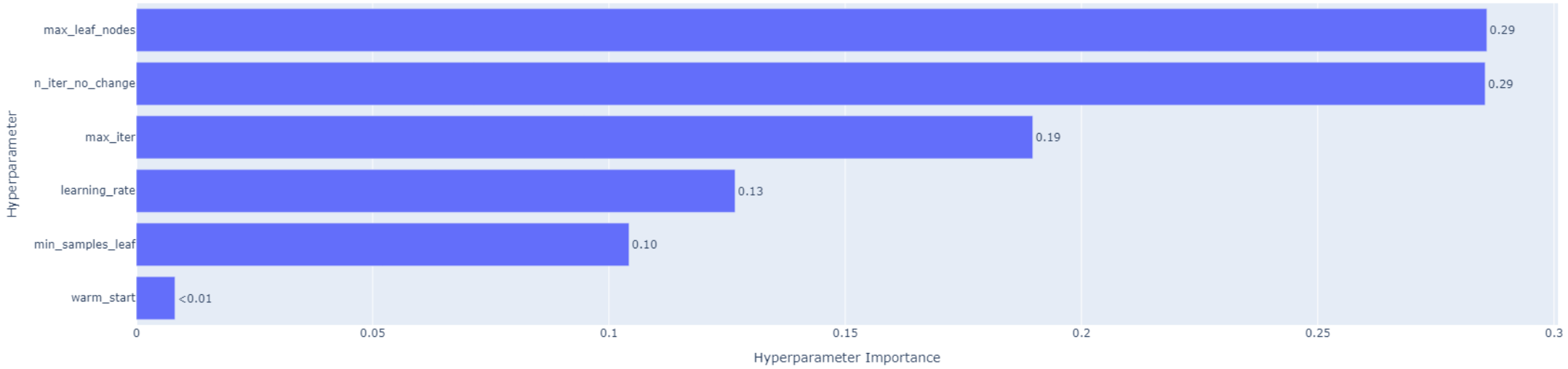
KNeighborsClassifier

Normal	*0.07792207792207792*
Random Over-Sampler	*0.25396825396825395*
SMOTE	*0.24832214765100669*
K-Folds cross validation	*0.05194805194805195*

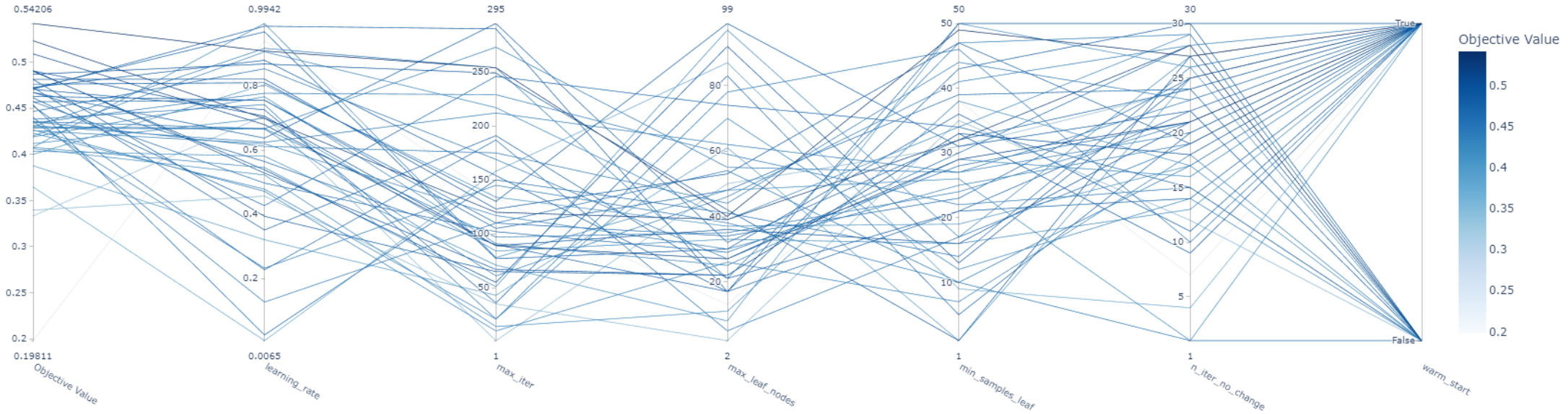
HistGradientBoostingClassifier

Normal	*0.3555555555555555*
Random Over-Sampler	*0.5087719298245614*
SMOTE	*0.4852941176470588*
K-Folds cross validation	*0.3516483516483516*

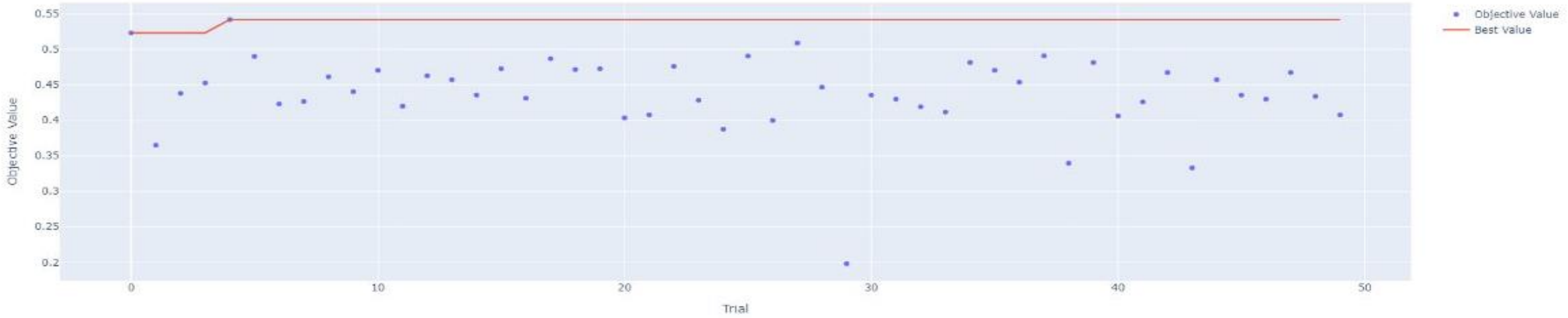
HYPERPARAMETER IMPORTANCES



PARALLER COORDINATE PLOT



OPTIMIZATION HISTORY PLOT



BEST OF THE BEST

model: *HistGradientBoostingClassifier*

method: *Random Over-Sampler (RO)*

Best hyperparameters:

- ***learning_rate:*** *0.9084578238481625,*
- ***max_iter:*** *254,*
- ***max_leaf_nodes:*** *40,*
- ***min_samples_leaf:*** *49,*
- ***warm_start:*** *True,*
- ***n_iter_no_change:*** *27*

accuracy: *0.976050830889540*6

precision: *0.7073170731707317*

recall: *0.4393939393939394*

f1_score: *0.542056074766355*1

THANK YOU FOR YOUR ATTENTION