XGBClassifier - v1

XGBoost, or Extreme Gradient Boosting, is an algorithm for gradient boosting on decision trees. It trains many decision trees sequentially, the additional tree always trying to mitigate the error of the whole model. XGBoost was the first gradient boosting algorithm to be implemented and is currently widely adopted in the ML world.

Model Performance

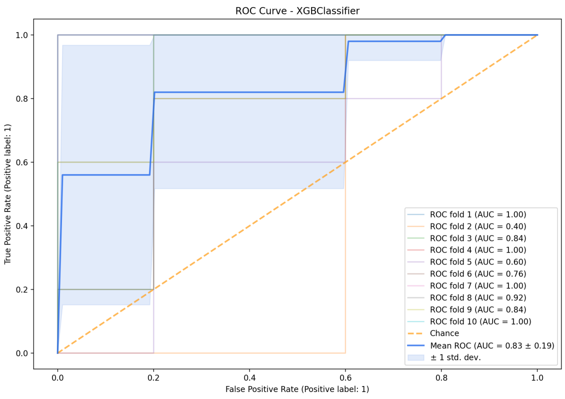
Model performance is analysed by various metrics. This model has been selected based on the neg_log_loss score.

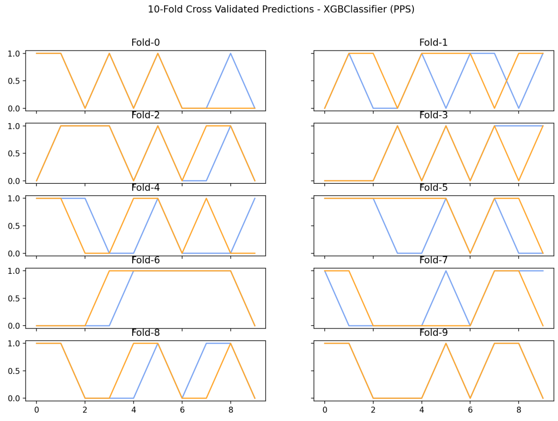
Metric	Score
Accuracy	80.00 ± 13.42 %
Precision	80.13 ± 15.73 %
Sensitivity	84.00 ± 14.97 %
Specificity	76.00 ± 21.54 %
F1 Score	80.99 ± 12.46 %

Confusion Matrix

		True Class	
		Faulty	Healthy
Prediction	Faulty	42.00 ± 7.48 %	12.00 ± 10.77 %
	Healthy	8.00 ± 7.48 %	38.00 ± 10.77 %

Area Under Curve & Cross Validation Plots





Validation Strategy

All experiments are cross-validated. This means that every time a model's performance is evaluated, it's trained on one part of the data, and test on another. Therefore, the model is always test against data it has not yet been trained for. This gives the best approximation for real world (out of sample) performance. The current validation strategy used is StratifiedKFold, with 10 splits and with shuffling the data.

Model Parameters

Parameter	Value
booster	**gblinear**
lambda	**2.9720e-02**
alpha	**1.2804e-06**
learning_rate	**1.7793e-02**
verbosity	**0**

Features

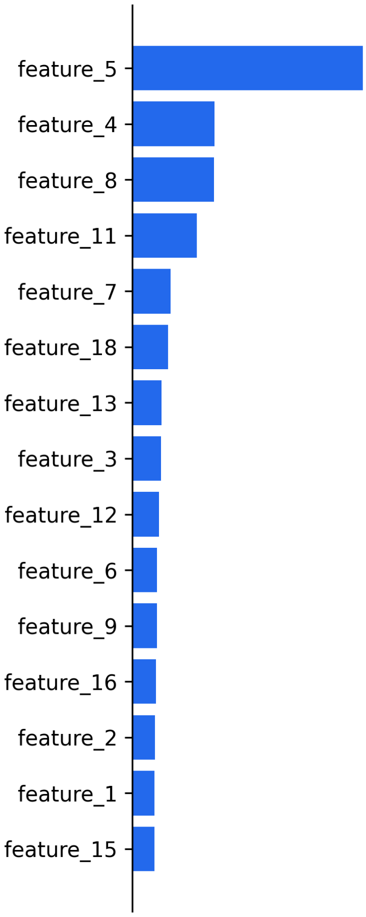
Feature Extraction

First, features that are co-linear (a * x = y) up to 99.0 % were removed. This resulted in 0 removed features:
Subsequently, the features were manipulated and analysed to extract additional information. All promising combinations are analysed with a single shallow decision tree.

Sort Feature	Quantity
Linear Features	0
Arithmetic Features	0
Trigonometric Features	0
Inverse Features	0
K-Means Features	0
Lagged Features	0
Differentiated Features	0

Feature Selection

Using a Random Forest model, the non-linear Feature Importance is analysed. The feature importance is measured in Mean Decrease in Gini Impurity. The feature importance is used to create two feature sets, one that contains 85% of all feature importance, and one that contains all features that contribute more than 1% to the total feature importance.
The top 15 feature with their mean decrease in Gini impurity are visualized on the right.



Data Processing

The following data manipulations were made to clean the data:

1. Removed 0 duplicate columns and 0 duplicate rows
2. Removed 0 outliers with clip
3. Imputed 0 missing values with zero
4. Removed 0 columns with constant values

Model Score Board

Not only the XGBClassifier has been optimized by the AutoML pipeline. In total, 13 models where trained. The following table shows the performance of the top 10 performing models:

Model	neg_log_loss
XGBClassifier	-0.3315 ± 0.0465 %
RandomForestClassifier	-0.3453 ± 0.0414 %
RandomForestClassifier	-0.3851 ± 0.0048 %
CatBoostClassifier	-0.4104 ± 0.0619 %
RandomForestClassifier	-0.4330 ± 0.0470 %
XGBClassifier	-0.4286 ± 0.0520 %
CatBoostClassifier	-0.4114 ± 0.0867 %

XGBClassifier	-0.4601 ± 0.0593 %
XGBClassifier	-0.5931 ± 0.0316 %
CatBoostClassifier	-0.5002 ± 0.1651 %