UNIVERSITÀ DEGLI STUDI FIRENZE

Heart Failure Prediction

Super Learner

Mattia Bennati

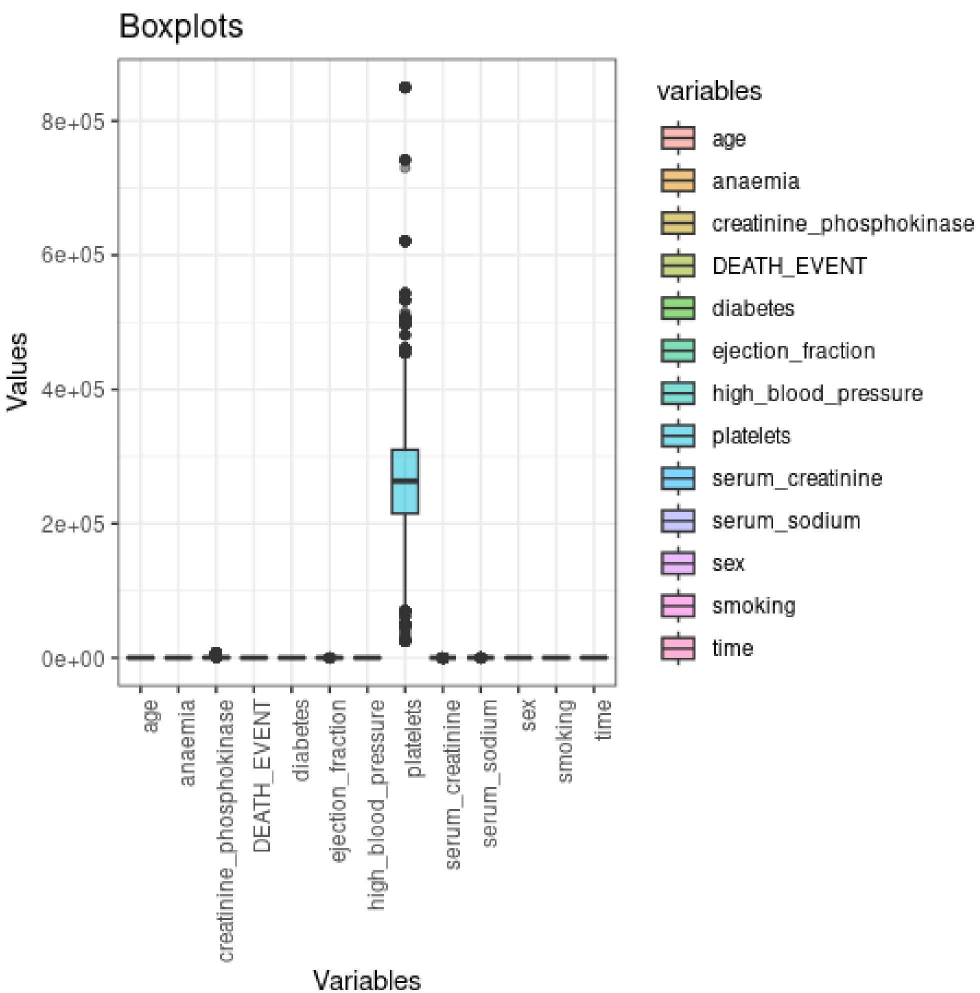
Florence 11/06/24

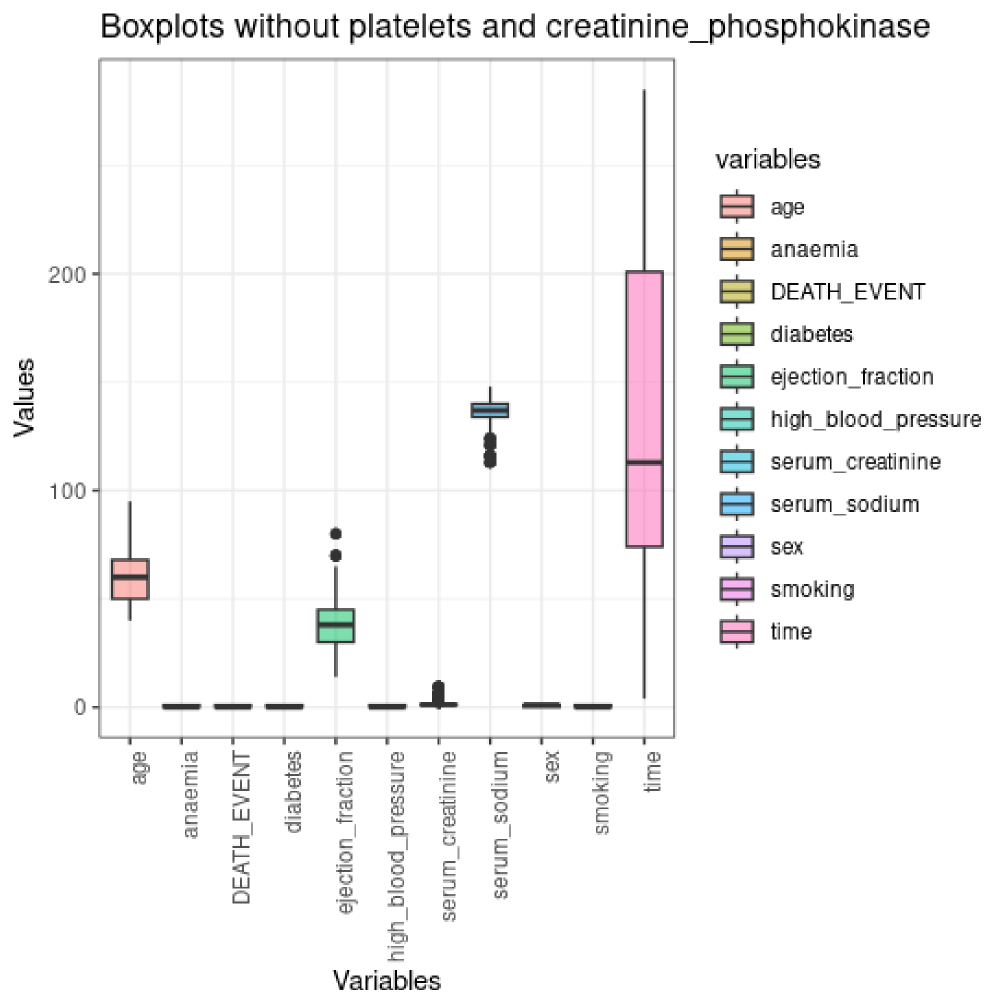
Project

Overview

1. Objective:
Training a super learner model while leveraging the strenghts of multiple predictive algorithms, achieving greater accuracy and a lower RMSE value.

2. Predictive models:
- **RandomForest** and **Ranger** (Ensambles based on decision trees)
- **XGBoost** (Bagging based decision trees)
- **GLMNet** and **Bayesian GLM** (General linear regressions)
- **SVM**
- **Neural Networks**
- **Polymars** (Piecewise-Polynomial, splines based regression)





Boxplots representations of all the variables. Highlighting the presence of some outliers.

10-fold cross-validation (parallel)

```
# cross-validation
cv_control <- SuperLearner.CV.control(V = 10)  # 10-fold cross-validation

# Setting up parallel processing
cluster <- makeCluster(detectCores() - 1)  # use all cores except one
registerDoParallel(cluster)
registerDoRNG(seed = 123)  # Ensure reproducibility
```

Model training with 70% of the dataset

```
# Executing the cross-validation phases in parallel thanks to cv_control
clusterExport(cluster, c("Y_train", "X_train", "learners", "cv_control"))
super_learner <- SuperLearner(Y = Y_train, X = X_train, family = binomial(),
                              SL.library = learners, method = "method.NNLS",
                              cvControl = cv_control, verbose = TRUE)
```

ROC and AUC

Evaluation

ROC (Receiver Operating Caracteristhic)

- Shows the performance of a binary classification model at different threshold levels

- Trade-Off between specificity and sensibility

Sensitivity: proportion of the true positives correctly identified by the model

$$sensitivity = \frac{True\ positives\,(TP)}{True\ positives\,(TP) + False\ negatives\,(FN)}$$

Specificity: proportion of the true negatives correctly identified by the model

$$specificity = \frac{True\ negatives\,(TN)}{True\ negatives\,(TN) + False\ positives\,(FP)}$$

ROC and AUC

Evaluation

AUC (Area Under The Curve):

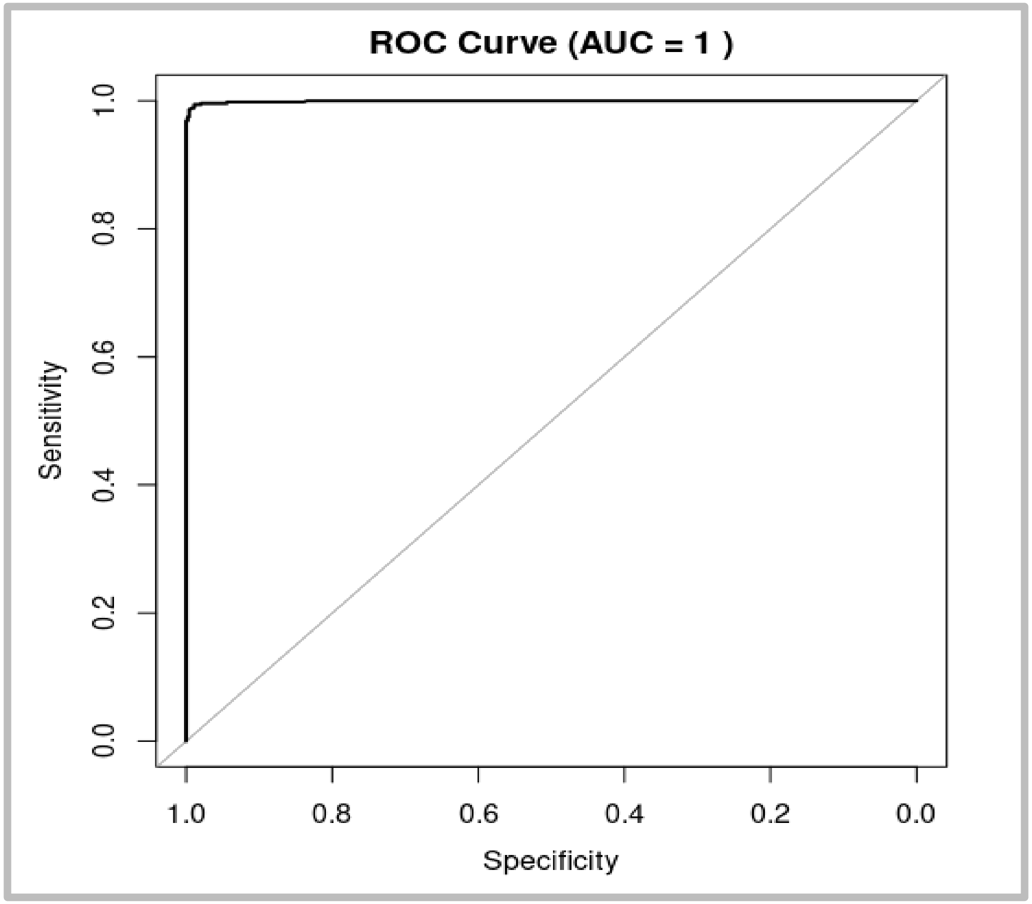
- It's the area under the ROC curve

- Summarises the performance of the model and its in the range [0,1]

AUC = 1:
The model is able to correctly distinguish between positive and negative classes

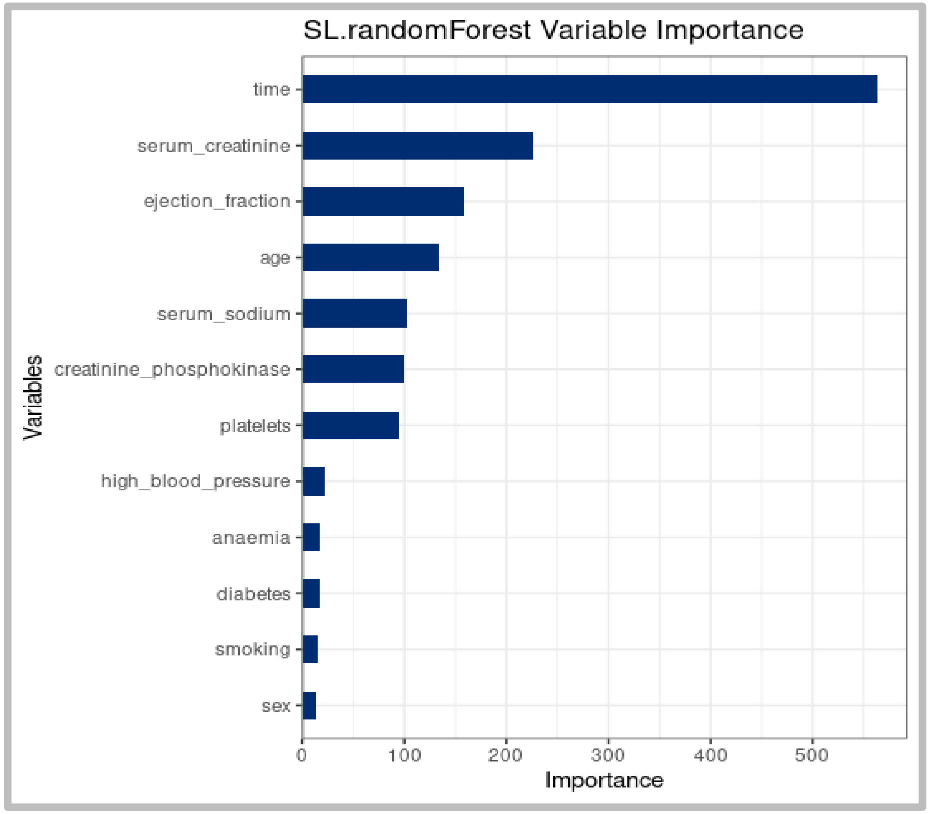
AUC = 0.5:
The model is unable to distinguish the classes

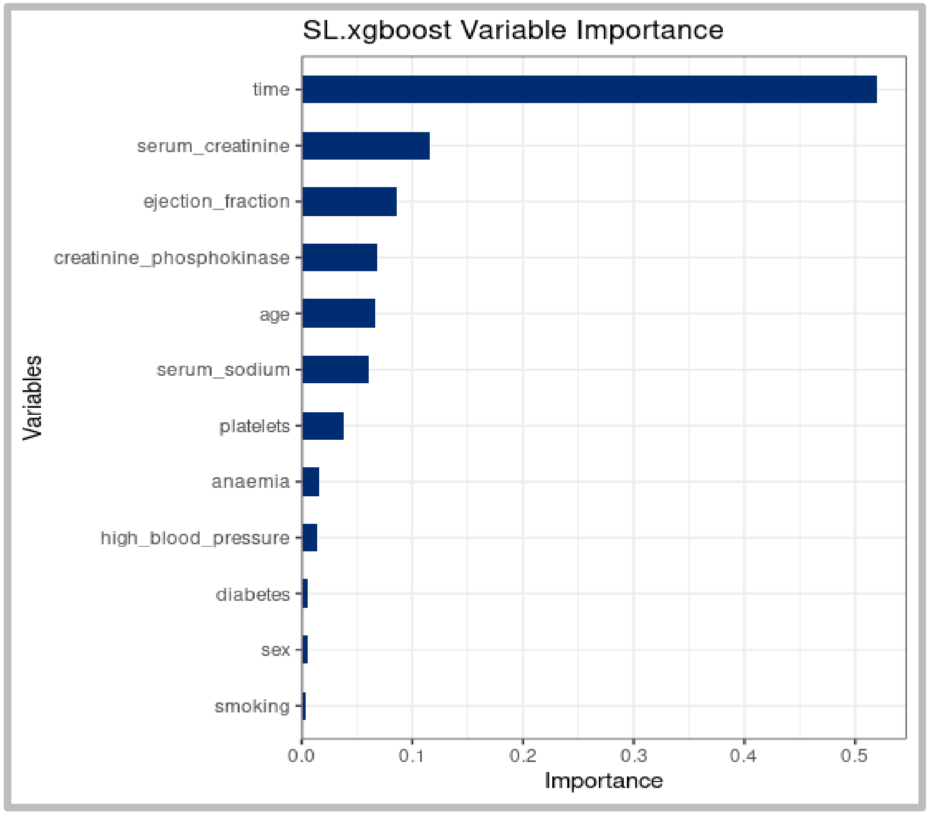
AUC < 0.5:
The model performs worse than a casual classification



The graph shows that the model is able to correctly identify true positives and avoid false positives.

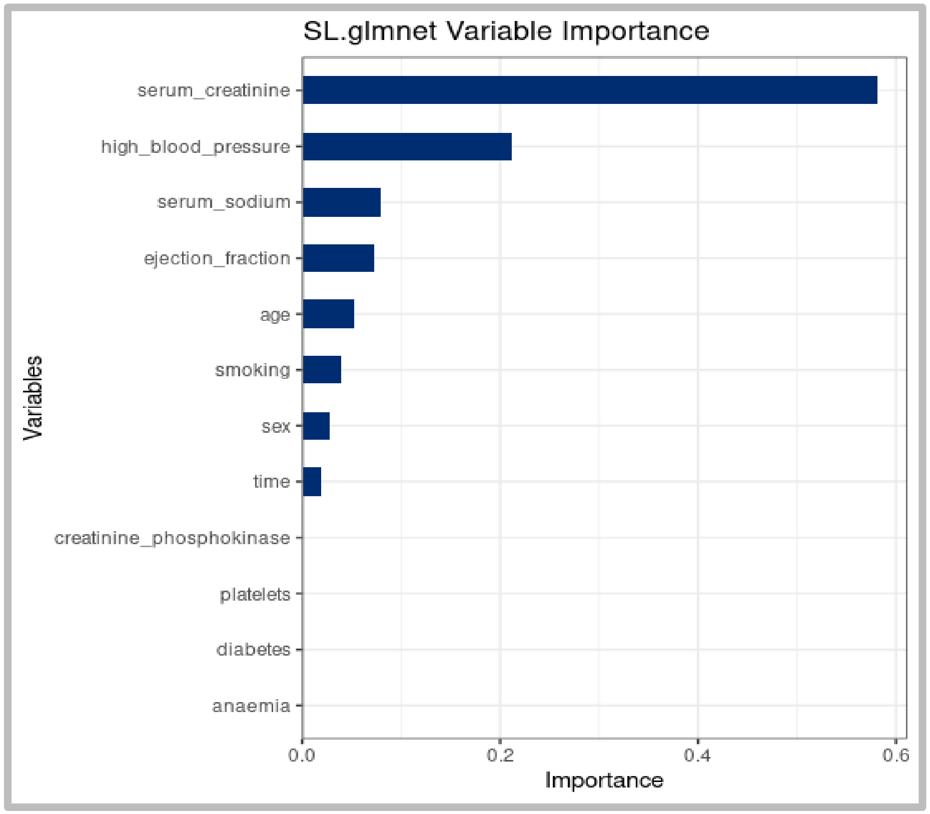
Variable importance

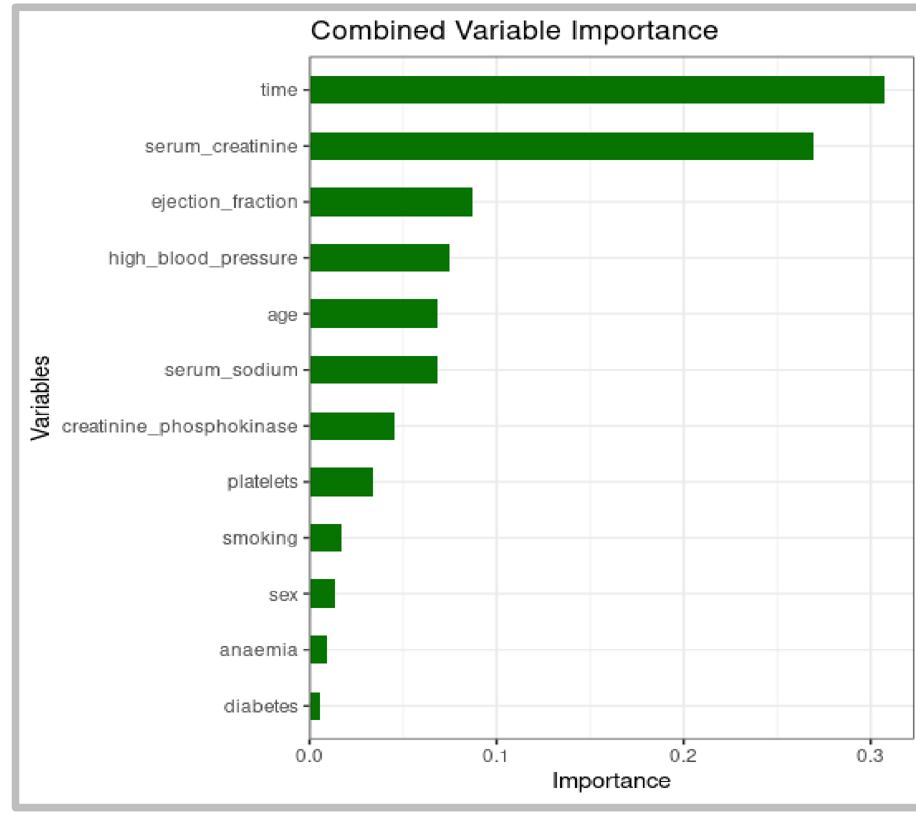




Graphs showing the varabile importance identified by the Random Forest and the XGBoost models

Variable importance



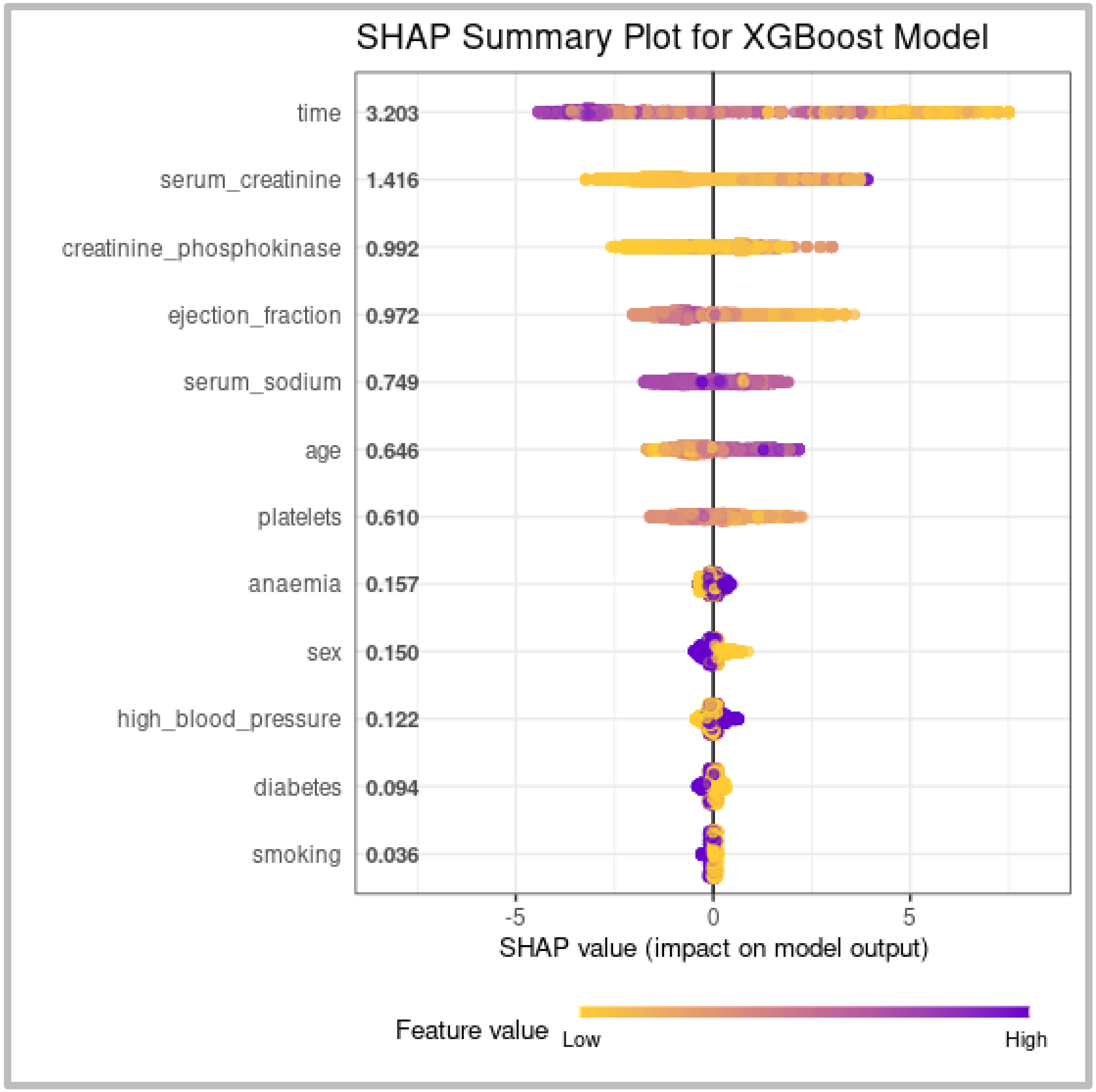


Graphs showing the varabile importance of the GLMNet model alongside the average variable importance of all the models combined.

SHapley Additive exPlanations:

- Help identifying the most important features and their effects on the predictions by quantifying their impact

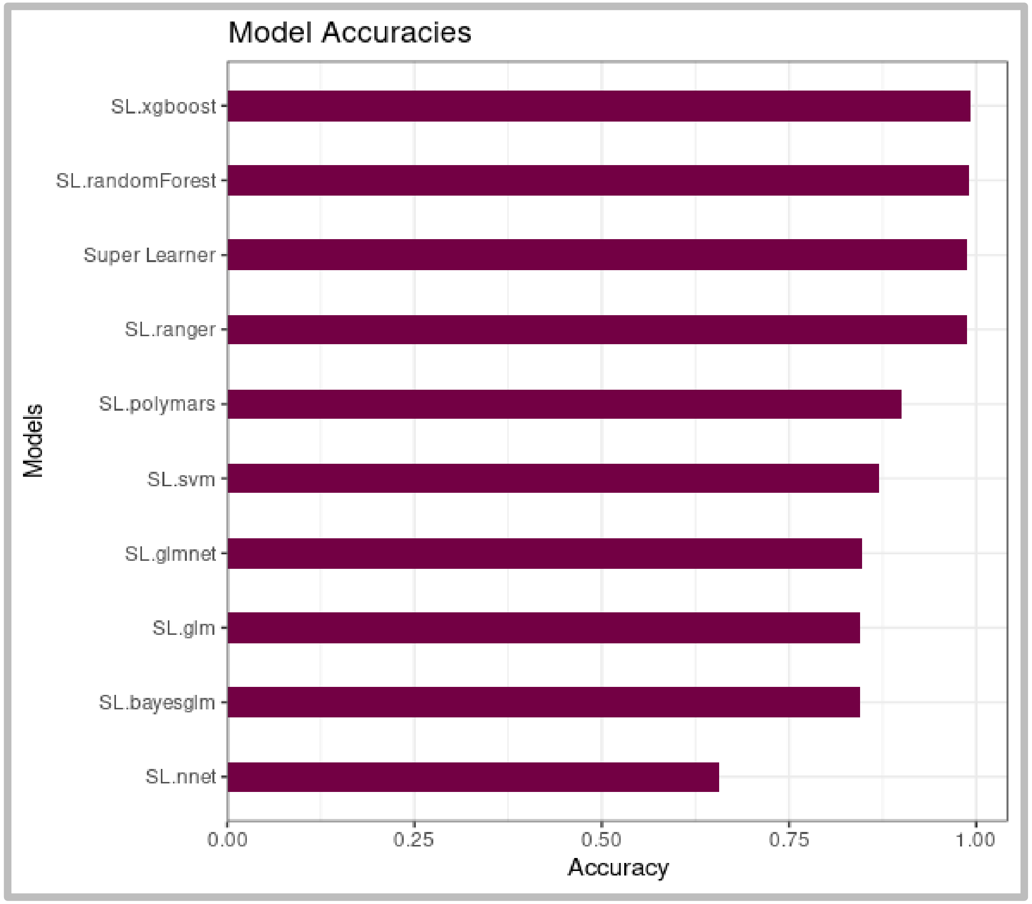
- Show how and how much the feature shifts the prediction in relation to the average



Descriptive SHAP summary showing the impact of each feature over the target variable

Performance evaluation

Model	Accuracy	RMSE
SL.xgboost	0.9927	0.0845
SL.randomForest	0.9893	0.0877
SL.ranger	0.9887	0.0866
Super Learner	0.9887	0.0848
SL.polymars	0.8993	0.2688
SL.svm	0.8713	0.2986
SL.glmnet	0.8480	0.3467
SL.glm	0.8460	0.3474
SL.bayesglm	0.8460	0.3474
SL.nnet	0.6573	0.4764



Descriptive bar plot showing the level of accuracy for each base model, compared to the SuperLearner

Resources

Dataset:
https://www.kaggle.com/datasets/aadarshvelu/heart-failure-prediction-clinical-records

Project source code:
https://github.com/Scrayil/Heart_Failure_Prediction