JADBio Description of Performed Analysis

Setup

JADBio version **1.4.118** ran on dataset **german** with **1000** samples and **20** features to create a predictive model for outcome named **Creditability**. The outcome was discrete leading to a **classification** modeling.

The preferences of the analysis were set to **true** for feature selection and **false** for full feature models tried.
The **AUC** metric was used to optimize for the best model.
The maximum number of features to select was set to **25**.
The effort to spend on tuning the algorithms were set to **Quick**.
The number of CPU cores to use for the analysis was set to **1**.
The execution time was **00:01:38**.

Configuration Space

JADBio's AI decide to try the following algorithms and tuning hyper-parameter values:

Algorithm Type	Algorithm	Hyper-parameter	Set of Values
Preprocessing	Mean Imputation		
	Mode Imputation		
	Constant Removal		
	Variable Normalization		
Feature Selection	Epilogi	stoppingThreshold	0.001
		stoppingCriterion	Independence Test
		equivalenceThreshold	0.01
	Test-Budgeted Statistically Equivalent Signature (SES)	maxK	2.0
		alpha	0.05
	LASSO	penalty	1.0
Modeling	Classification Decision Tree with Deviance splitting criterion	alpha	0.05
		minLeafSize	3
	Ridge Logistic Regression	lambda	1.0
	Classification Random Forest with Deviance splitting criterion	nTrees	100
		minLeafSize	3.0
	Support Vector Machines (SVM) of type C-SVC with Linear Kernel	cost	1.0
	Support Vector Machines (SVM) of type C-SVC with Polynomial Kernel	cost	1.0
		degree	3
		gamma	1.0
	Support Vector Machines (SVM) of type C-SVC with Gaussian Kernel	cost	1.0
		gamma	1.0

Leading to **25** combinations and corresponding configurations (machine learning pipelines) to try. For the full configurations tested see the Appendix.

Configuration Estimation Protocol

JADBio's AI system decided to estimate the out-of-sample performance of the models produced by each configuration using **Incomplete 10-fold CV without dropping.** Overall, 225 models were set out to train.

JADBio Results Summary

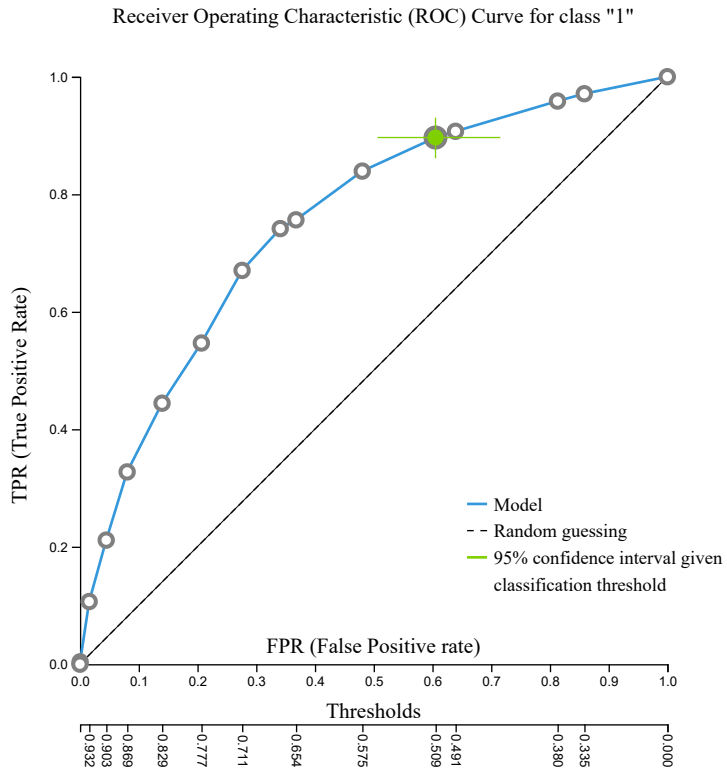
Overview

A result summary is presented for analysis optimized for Performance. The model is produced by applying the algorithms in sequence (configuration) on the training data:

Preprocessing	Feature Selection	Predictive algorithm
Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi algorithm with hyper-parameters: equivAlpha = 0.01, and stopping criterion = Independence Test with threshold: 0.001.	Ridge Logistic Regression with penalty hyper-parameter lambda = 1.0

The **Area Under The Curve** is **0.755** with 95% confidence interval being [**0.703,0.803**].
The **Mean Average Precision (a.k.a. Average Area Under the Precision-Recall curve)** is **0.742** with 95% confidence interval being [**0.695,0.788**].

The Area Under the ROC Curve is shown in the figure below:



Selecting to classify as class: 1 any sample with predicted probability to be in this class above **0.5092**, the model achieves:

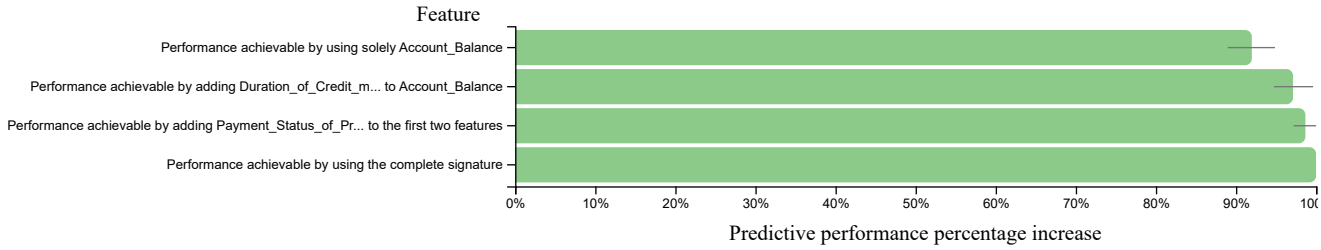
Metric	Mean estimate	CI
Accuracy	0.746	[0.704, 0.785]
Balanced Accuracy	0.646	[0.594, 0.694]
F1 Score	0.830	[0.797, 0.860]
Matthews correlation criterion (phi coefficient)	0.339	[0.235, 0.441]
Precision	0.776	[0.730, 0.818]
True Positive Rate (a.k.a. Sensitivity, Recall. Hit Rate)	0.896	[0.861, 0.930]
Specificity	0.395	[0.289, 0.494]
True Positive Ratio	0.628	[0.580, 0.668]
True Negative Ratio	0.118	[0.085, 0.153]
False Positive Ratio	0.182	[0.143, 0.224]
False Negative Ratio	0.072	[0.048, 0.098]

Feature Selection

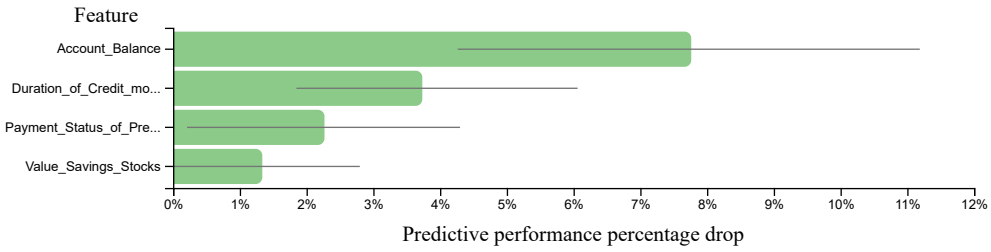
There were **4** features selected out of the **20** available.

The selected features consist of the following subset called a signature. **There was a single signature identified.** The first signature identified by the system is the set: **Account_Balance, Duration_of_Credit_monthly, Payment_Status_of_Previous_Credit, Value_Savings_Stocks** in order of importance. The following features cannot be substituted with others and still obtain an equal predictive performance: **Account_Balance, Duration_of_Credit_monthly, Payment_Status_of_Previous_Credit, Value_Savings_Stocks**.

The performance achieved by adding each feature in sequence to the model relative to the performance of the final model with all selected features is shown below. The features are added in order of importance:

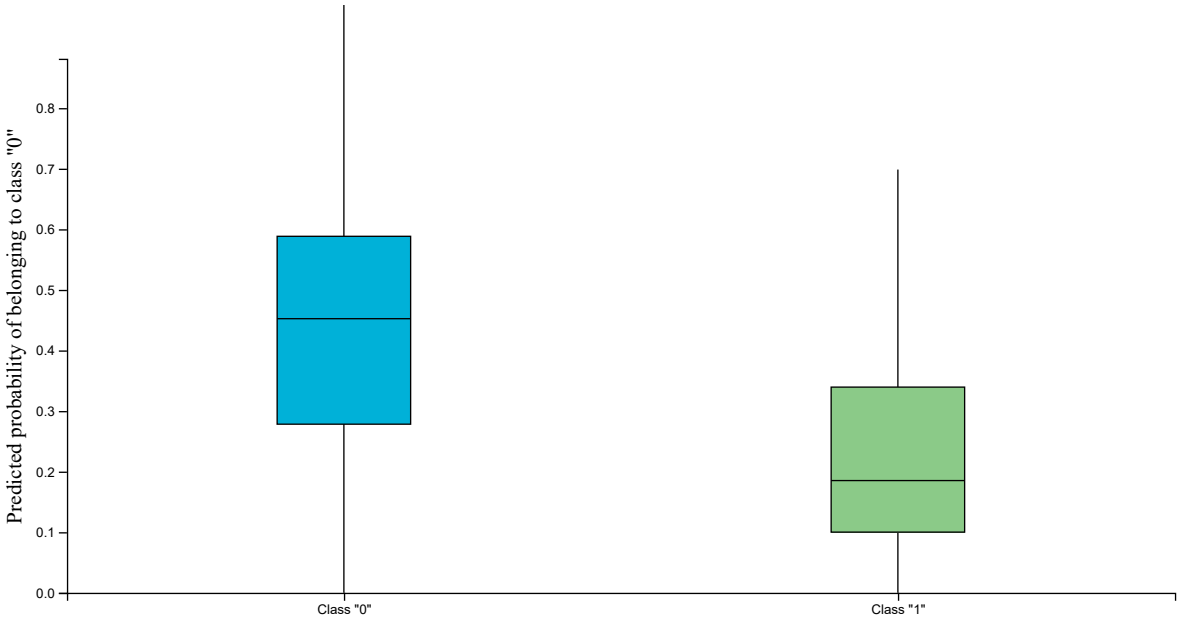


Some features may not seem to add predictive performance to the model; however, the feature selection algorithms include them as an effort to make the final model more robust to noise. The performances achieved by a model that contains all features except one, relative to the performance achieved when the feature is removed is shown below:



For some features there is no noticeable drop in performance when they are removed because they carry predictive information that is shared by other features selected.

The separation of the predictions of the classes achieved by the model is shown in the box-plots below. These are the out-of-sample predictions made by model produced by the same configuration as the final model when the sample was used for testing (e.g.., during cross-validation) and was not used to train the model.



Appendix

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
1	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi	equivThresh = 0.01, stopping criterion = Independence Test, stopping threshold = 0.001	Classification Decision Tree with Deviance splitting criterion	minimum leaf size = 3, alpha = 0.05	0.7234656084656085	00:00:00.300	false
2	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Ridge Logistic Regression	lambda = 1.0	0.7596296296296298	00:00:00.087	false
3	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi	equivThresh = 0.01, stopping criterion = Independence Test, stopping threshold = 0.001	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.7576719576719578	00:00:00.291	false
4	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.7536772486772487	00:00:00.755	false
5	IdentityFactory	FullSelector	-	Trivial model	-	0.5	00:00:00.000	false
6	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Support Vector Machines (SVM) of type C-SVC	kernel = 'Linear Kernel', cost = 1.0	0.7575661375661377	00:00:00.112	false
7	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi	equivThresh = 0.01, stopping criterion = Independence Test, stopping threshold = 0.001	Support Vector Machines (SVM) of type C-SVC	kernel = 'Polynomial Kernel', cost = 1.0, gamma = 1.0, degree = 3	0.568994708994709	00:00:00.290	false
8	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Support Vector Machines (SVM) of type C-SVC	kernel = 'Linear Kernel', cost = 1.0	0.7598412698412699	00:00:00.740	false
9	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi	equivThresh = 0.01, stopping criterion = Independence Test, stopping threshold = 0.001	Support Vector Machines (SVM) of type C-SVC	kernel = 'Linear Kernel', cost = 1.0	0.7633597883597882	00:00:00.280	false
10	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Support Vector Machines (SVM) of type C-SVC	kernel = 'Polynomial Kernel', cost = 1.0, gamma = 1.0, degree = 3	0.6133862433862435	00:00:00.754	false

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
11	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Support Vector Machines (SVM) of type C-SVC	kernel = 'Polynomial Kernel', cost = 1.0, gamma = 1.0, degree = 3	0.5967724867724867	00:00:00.115	false
12	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.7425396825396826	00:00:00.293	false
13	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Ridge Logistic Regression	lambda = 1.0	0.7615343915343917	00:00:00.725	false
14	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.7535978835978837	00:00:00.144	false
15	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Classification Decision Tree with Deviance splitting criterion	minimum leaf size = 3, alpha = 0.05	0.7168253968253967	00:00:00.126	false
16	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Support Vector Machines (SVM) of type C-SVC	kernel = 'Gaussian Kernel', cost = 1.0, gamma = 1.0	0.7097619047619048	00:00:00.760	false
17	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.7408730158730158	00:00:00.765	false
18	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.7483597883597883	00:00:00.170	false
19	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Support Vector Machines (SVM) of type C-SVC	kernel = 'Gaussian Kernel', cost = 1.0, gamma = 1.0	0.6715873015873016	00:00:00.170	false
20	Mean Imputation,	Epilogi	equivThresh = 0.01, stopping	Ridge Logistic Regression	lambda = 1.0	0.7636507936507936	00:00:00.279	false

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
	Mode Imputation, Constant Removal, Standardization		criterion = Independence Test, stopping threshold = 0.001					
21	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Classification Decision Tree with Deviance splitting criterion	minimum leaf size = 3, alpha = 0.05	0.7157407407407407	00:00:00.743	false
22	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.7415608465608464	00:00:00.762	false
23	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi	equivThresh = 0.01, stopping criterion = Independence Test, stopping threshold = 0.001	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.7421693121693121	00:00:00.305	false
24	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi	equivThresh = 0.01, stopping criterion = Independence Test, stopping threshold = 0.001	Support Vector Machines (SVM) of type C-SVC	kernel = 'Gaussian Kernel', cost = 1.0, gamma = 1.0	0.7167195767195768	00:00:00.325	false
25	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi	equivThresh = 0.01, stopping criterion = Independence Test, stopping threshold = 0.001	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.7421693121693121	00:00:00.312	false