JADBio Description of Performed Analysis

Setup

JADBio version **1.4.69** ran on dataset **small_synthetic_binary** with **1000** samples and **80** features to create a predictive model for outcome named **feature0**. The outcome was continuous leading to a **regression** modeling.

The preferences of the analysis were set to **false** for feature selection and **false** for full feature models tried.
The **R2** 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 **Preliminary**.
The number of CPU cores to use for the analysis was set to **1**.
The execution time was **00:00:09**.

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	Mode imputation		
	Mean imputation		
	Contant Removal		
	Standardization		
Feature Selection	Test-Budgeted Statistically Equivalent Signature (SES)	alpha	0.05
		maxk	2
	LASSO	penalties	1.0
	FullSelector		
Modeling	Linear Regression	lambdas	1.0
	LinearSVR	costs	1.0
	PolynomialSVR	gammas	1.0
		costs	1.0
		epsilons	0.1
		degrees	3
	RBFSVR	gammas	1.0
		costs	1.0
		epsilons	0.1
	Random Forests	min leaf sizes	5
		vars to split	nvars // 3.0, nvars // 5.0, nvars // 7.0
		splits to perform	1.0

Algorithm Type	Algorithm	Hyper-parameter	Set of Values
		ntrees	100
	Decision Tree	min leaf sizes	5
		vars to split	nvars // 1.0
		splits to perform	1.0
		alphas	0.05

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 with dropping.** Overall, 75 models were set out to train. Out of those, only 33 models were eventually trained, as JADBio stopped all configuration evaluations when it deemed that no sufficient progress was made. JADBio **used** the Early Dropping criterion (see [1]) to stop computations early on configurations that did not seem promising. Eventually, 33 had their estimation protocol completed.

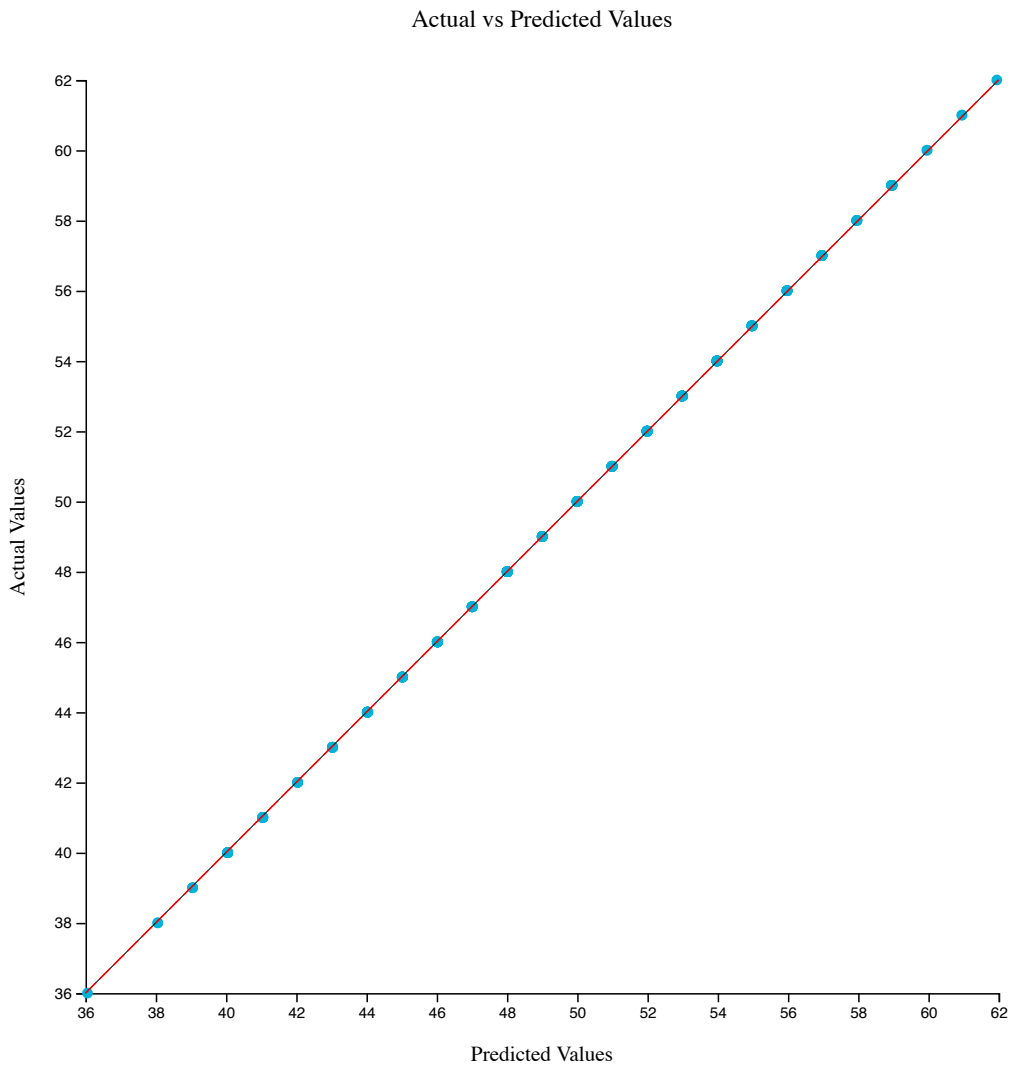
JADBio Results Summary

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	FullSelector	Ridge Linear Regression with penalty hyper-parameter lambda = 1.0

The R-squared is shown in the figure below:



Metric | Mean estimate | CI --- | --- | --- R-squared | 1.000 | [1.000, 1.000] Mean Absolute Error | 0.019 | [0.017, 0.021] Mean Squared Error | 0.001 | [0.000, 0.001] Relative Absolute Error | 0.005 | [0.004, 0.005] Relative Squared Error | 0.000 | [0.000, 0.000] Correlation Coefficient | 1.000 | [1.000, 1.000]

Feature Selection

Jadbio selected **all** features in the original dataset for the reference signature. Note that **55** features that were found constant are excluded.

Appendix

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
1	Mean Imputation, Mode Imputation, Constant Removal, Standardization	FullSelector	-	Regression Random Forests with Mean Squared Error splitting critetion	ntrees = 100, minimum leaf size = 5	0.5091263947774449	00:00:00.166	true

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
2	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO Feature Selection	penalty = 1.0	Ridge Linear Regression	lambda = 1.0	0.6490218250774322	00:00:00.097	true
3	Mean Imputation, Mode Imputation, Constant Removal, Standardization	FullSelector	-	Support Vector Regression Machines (SVR) of type epsilon-SVR	kernel = 'Linear Kernel', cost = 1.0, epsilon = 0.1	0.9763142041333964	00:00:00.074	false
4	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES) algorithm	maxK = 2, alpha = 0.05, budget = 3 * nvars	Regression Random Forests with Mean Squared Error splitting critetion	ntrees = 100, minimum leaf size = 5	0.432374667358165	00:00:01.1649	true
5	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO Feature Selection	penalty = 1.0	Regression Decision Tree with Mean Squared Error splitting critetion	minimum leaf size = 5, alpha = 0.05	-1.1746022151186	00:00:00.101	true
6	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES) algorithm	maxK = 2, alpha = 0.05, budget = 3 * nvars	Support Vector Regression Machines (SVR) of type epsilon-SVR	kernel = 'Linear Kernel', cost = 1.0, epsilon = 0.1	0.6914083846872994	00:00:01.1704	true
7	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO Feature Selection	penalty = 1.0	Support Vector Regression Machines (SVR) of type epsilon-SVR	kernel = 'Radial Basis Function Kernel', cost = 1.0, gamma = 1.0, epsilon = 0.1	0.23512416107364076	00:00:00.149	true
8	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES) algorithm	maxK = 2, alpha = 0.05, budget = 3 * nvars	Regression Random Forests with Mean Squared Error splitting critetion	ntrees = 100, minimum leaf size = 5	0.47145921717329486	00:00:01.1656	true

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
9	IdentityFactory	NoSelector	-	Trivial model	-	9.251858538542972e-16	00:00:00.000	false
10	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO Feature Selection	penalty = 1.0	Regression Random Forests with Mean Squared Error splitting critetion	ntrees = 100, minimum leaf size = 5	0.44955473604865226	00:00:00.127	true
11	Mean Imputation, Mode Imputation, Constant Removal, Standardization	FullSelector	-	Support Vector Regression Machines (SVR) of type epsilon-SVR	kernel = 'Polynomial Kernel', cost = 1.0, gamma = 1.0, degree = 3, epsilon = 0.1	0.9560916612104304	00:00:00.045	true
12	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO Feature Selection	penalty = 1.0	Support Vector Regression Machines (SVR) of type epsilon-SVR	kernel = 'Polynomial Kernel', cost = 1.0, gamma = 1.0, degree = 3, epsilon = 0.1	0.43151221493540715	00:00:00.959	true
13	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES) algorithm	maxK = 2, alpha = 0.05, budget = 3 * nvars	Support Vector Regression Machines (SVR) of type epsilon-SVR	kernel = 'Polynomial Kernel', cost = 1.0, gamma = 1.0, degree = 3, epsilon = 0.1	0.3818752179823456	00:00:02.2050	true
14	Mean Imputation, Mode Imputation, Constant Removal, Standardization	FullSelector	-	Regression Random Forests with Mean Squared Error splitting critetion	ntrees = 100, minimum leaf size = 5	0.5204820137725311	00:00:00.059	true
15	Mean Imputation, Mode Imputation, Constant Removal, Standardization	FullSelector	-	Support Vector Regression Machines (SVR) of type epsilon-SVR	kernel = 'Radial Basis Function Kernel', cost = 1.0, gamma = 1.0, epsilon = 0.1	0.001367900762507701	00:00:00.069	true
16	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO Feature Selection	penalty = 1.0	Support Vector Regression Machines (SVR) of type epsilon-SVR	kernel = 'Linear Kernel', cost = 1.0, epsilon = 0.1	0.6387148769752511	00:00:00.162	true

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
17	Mean Imputation, Mode Imputation, Constant Removal, Standardization	FullSelector	-	Ridge Linear Regression	lambda = 1.0	0.9999783353690532	00:00:00.007	false
18	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO Feature Selection	penalty = 1.0	Regression Random Forests with Mean Squared Error splitting critetion	ntrees = 100, minimum leaf size = 5	0.4267025603514405	00:00:00.108	true
19	Mean Imputation, Mode Imputation, Constant Removal, Standardization	FullSelector	-	Regression Decision Tree with Mean Squared Error splitting critetion	minimum leaf size = 5, alpha = 0.05	-1.116293082131747	00:00:00.008	true
20	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES) algorithm	maxK = 2, alpha = 0.05, budget = 3 * nvars	Ridge Linear Regression	lambda = 1.0	0.691299269631292	00:00:00.977	false
21	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES) algorithm	maxK = 2, alpha = 0.05, budget = 3 * nvars	Support Vector Regression Machines (SVR) of type epsilon-SVR	kernel = 'Radial Basis Function Kernel', cost = 1.0, gamma = 1.0, epsilon = 0.1	0.20570296702965263	00:00:01.1668	true
22	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES) algorithm	maxK = 2, alpha = 0.05, budget = 3 * nvars	Regression Decision Tree with Mean Squared Error splitting critetion	minimum leaf size = 5, alpha = 0.05	-1.051857507675416	00:00:01.1643	true
23	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES) algorithm	maxK = 2, alpha = 0.05, budget = 3 * nvars	Regression Random Forests with Mean Squared Error splitting critetion	ntrees = 100, minimum leaf size = 5	0.48925122006577404	00:00:01.1666	true

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
24	Mean Imputation, Mode Imputation, Constant Removal, Standardization	FullSelector	-	Regression Random Forests with Mean Squared Error splitting critetion	ntrees = 100, minimum leaf size = 5	0.5327551379117574	00:00:00.109	true
25	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO Feature Selection	penalty = 1.0	Regression Random Forests with Mean Squared Error splitting critetion	ntrees = 100, minimum leaf size = 5	0.4835302108021339	00:00:00.124	true