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 **true** 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 **00:00:10**.

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
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
		ntrees	100
	Decision Tree	min leaf sizes	5
		vars to split	nvars // 1.0
		splits to perform	1.0
		alphas	0.05

Algorithm Type	Algorithm	Hyper-parameter	Set of Values

Leading to **17** 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, 51 models were set out to train. Out of those, only 27 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, 27 had their estimation protocol completed.

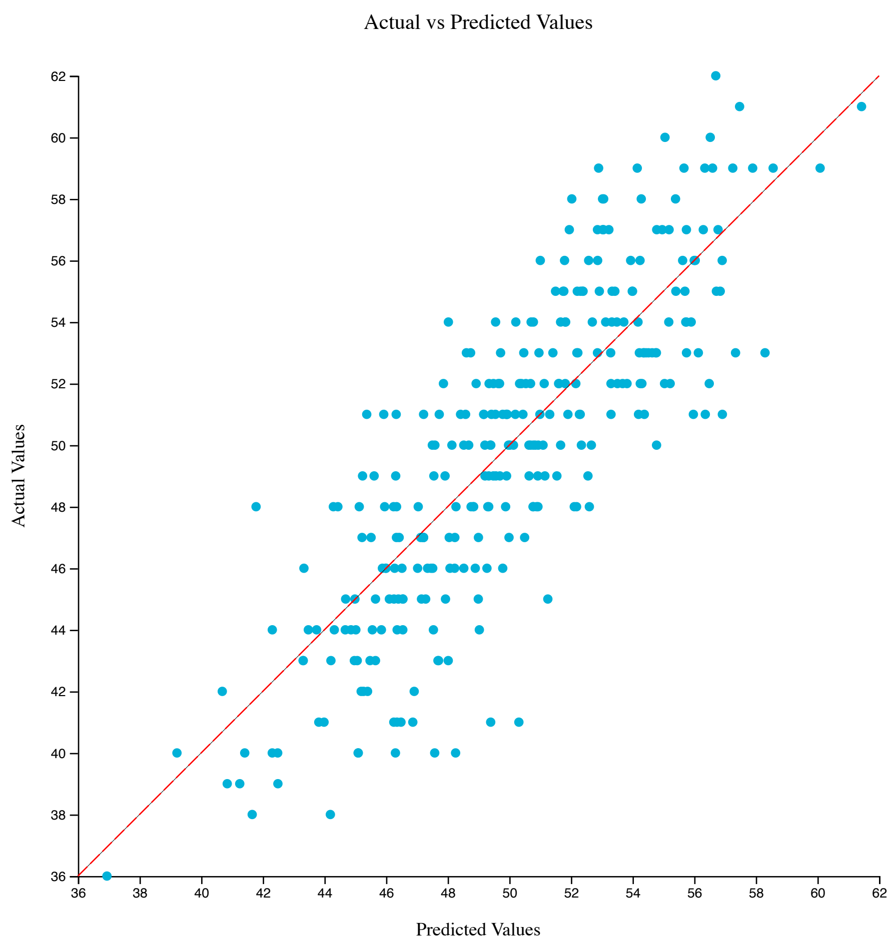
JADBio Results Summary

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	Test-Budgeted Statistically Equivalent Signature (SES) algorithm with hyper-parameters: maxK = 2, alpha = 0.05 and budget = 3 * nvars	Support Vector Regression Machines (SVR) of type epsilon-SVR with Linear Kernel and hyper-parameters: cost = 1.0, epsilon = 0.1

The R-squared is shown in the figure below:



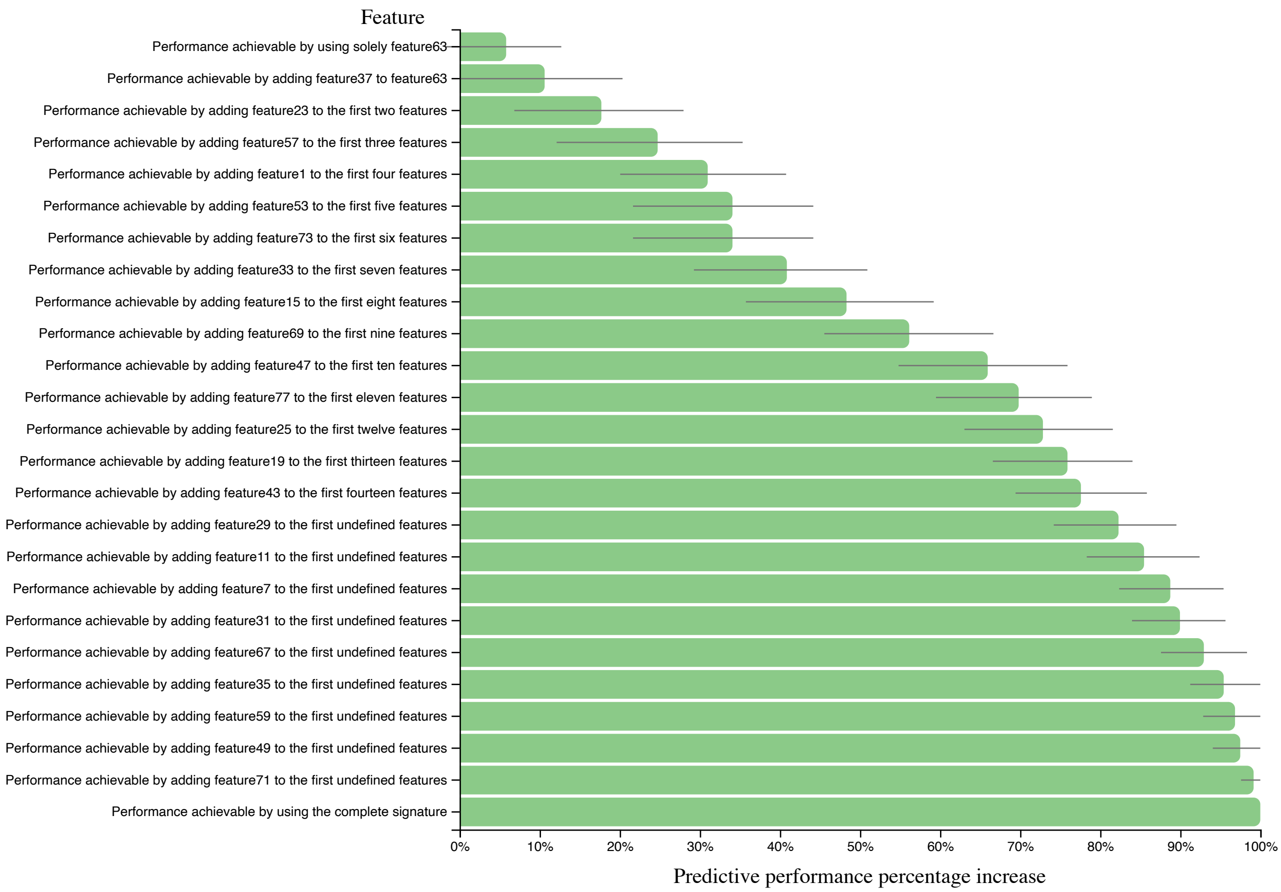
Metric | Mean estimate | CI --- | --- | --- R-squared | 0.682 | [0.599, 0.752] Mean Absolute Error | 2.319 | [2.049, 2.589] Mean Squared Error | 8.366 | [6.675, 10.232] Relative Absolute Error | 0.561 | [0.486, 0.639] Relative Squared Error | 0.326 | [0.253, 0.415] Correlation Coefficient | 0.829 | [0.778, 0.873]

Feature Selection

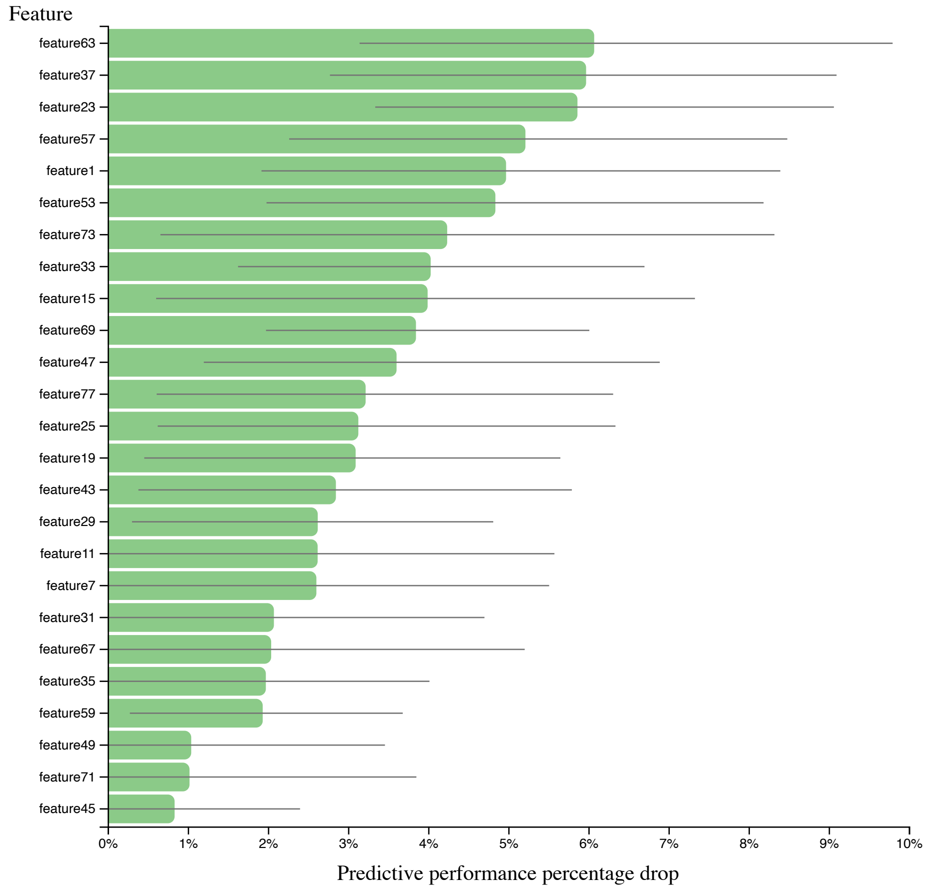
There were **25** features selected out of the **80** available.

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: **feature47, feature15, feature31, feature71, feature57, feature63, feature37, feature49, feature19, feature11, feature33, feature77, feature1, feature25, feature67, feature23, feature73, feature53, feature43, feature7, feature59, feature29, feature45, feature69, feature35** in order of importance. The following features cannot be substituted with others and still obtain an equal predictive performance: **feature47, feature15, feature31, feature71, feature57, feature63, feature37, feature49, feature19, feature11, feature33, feature77, feature1, feature25, feature67, feature23, feature73, feature53, feature43, feature7, feature59, feature29, feature45, feature69, feature35**.

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.

Appendix

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
1	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	0.6125799801505796	00:00:00.230	true
2	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO Feature Selection	penalty = 1.0	Ridge Linear Regression	lambda = 1.0	0.6204000166918998	00:00:00.125	true
3	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	00:00:00.179	true

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
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	00:00:01.1182	true
5	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	00:00:00.752	false
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 = 'Radial Basis Function Kernel', cost = 1.0, gamma = 1.0, epsilon = 0.1	0.20570296702965263	00:00:01.1199	true
7	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	00:00:00.106	true
8	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	0.6920239497493902	00:00:00.843	false
9	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	00:00:00.127	true
10	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	00:00:01.1174	true
11	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	00:00:01.1189	true
12	IdentityFactory	NoSelector	-	Trivial model	-	9.251858538542972e-16	00:00:00.000	false

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
13	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	00:00:00.143	true
14	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	00:00:00.829	true
15	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	00:00:01.1198	true
16	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	00:00:00.128	true
17	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	00:00:01.1575	true