JADBio Description of Performed Analysis

Setup

JADBio version **1.4.69** ran on dataset **large_synthetic_binary** with **100000** 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 **6**.
The execution time was **00:37:07**.

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
	PolynomialSVR	gammas	], costs=[
		costs	], epsilons=[
		epsilons	], degrees=[
		degrees	
	RBFSVR	gammas	], costs=[
		costs	], epsilons=[
		epsilons	
	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

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

Leading to **16** 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 **90.00 % - % 10.00 hold-out.** Overall, 16 models were set out to train.

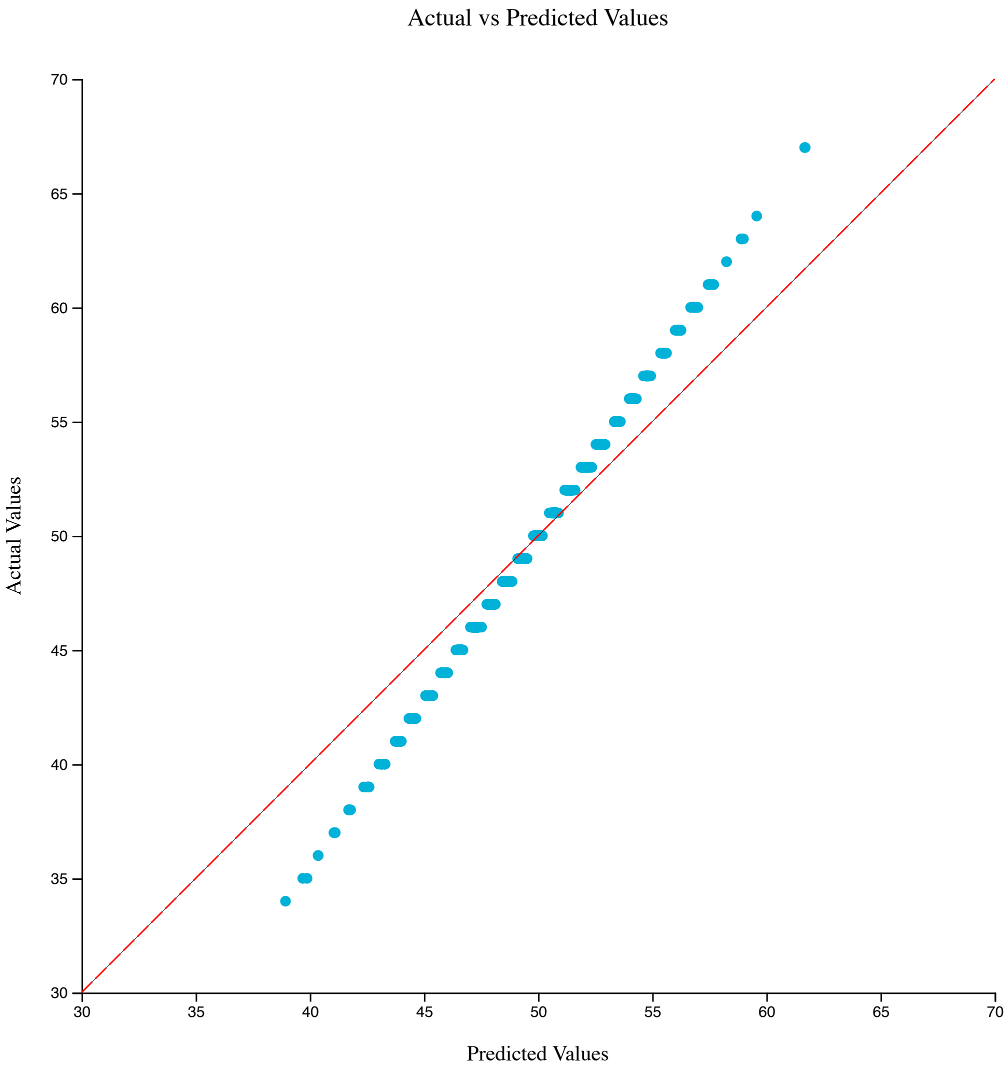
JADBio Results Summary

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	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 | 0.901 | [0.900, 0.901] Mean Absolute Error | 1.260 | [1.237, 1.283] Mean Squared Error | 2.478 | [2.397, 2.562] Relative Absolute Error | 0.316 | [0.315, 0.316] Relative Squared Error | 0.099 | [0.099, 0.100] 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	0.5750032685301922	00:00:14.14683	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	0.5069873499447379	00:00:05.5109	true
3	Mean Imputation, Mode Imputation, Constant Removal, Standardization	FullSelector	-	Ridge Linear Regression	lambda = 1.0	0.9006474222978912	00:00:02.2455	false
4	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	0.41236408960544657	00:00:07.7222	true
5	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	0.4361316703437021	00:00:13.13837	true
6	Mean Imputation, Mode Imputation, Constant Removal, Standardization	FullSelector	-	Regression Decision Tree with Mean Squared Error splitting critetion	minimum leaf size = 5, alpha = 0.05	-1.353860468197353	00:00:03.3441	true
7	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	0.5867242704434097	00:00:11.11571	false
8	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	-1.3913350469950294	00:00:05.5798	true
9	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	-1.3731097109537496	00:00:12.12240	true

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
10	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	0.49365970838836226	00:00:15.15433	true
11	IdentityFactory	NoSelector	-	Trivial model	-	1.695310558602614e-13	00:00:00.000	false
12	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	0.45006056574782044	00:00:08.8265	true
13	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	0.5140825150891726	00:00:17.17282	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	0.6088445186000755	00:00:27.27581	false
15	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	0.48812920855286257	00:00:11.11613	true
16	Mean Imputation, Mode Imputation, Constant Removal, Standardization	FullSelector	-	Regression Random Forests with Mean Squared Error splitting critetion	ntrees = 100, minimum leaf size = 5	0.5884311071948496	00:00:19.19554	true