JADBio Description of Performed Analysis

Setup

JADBio version **1.4.69** ran on dataset **toeholds_binary** with **91534** samples and **236** 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 **02:39:18**.

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.

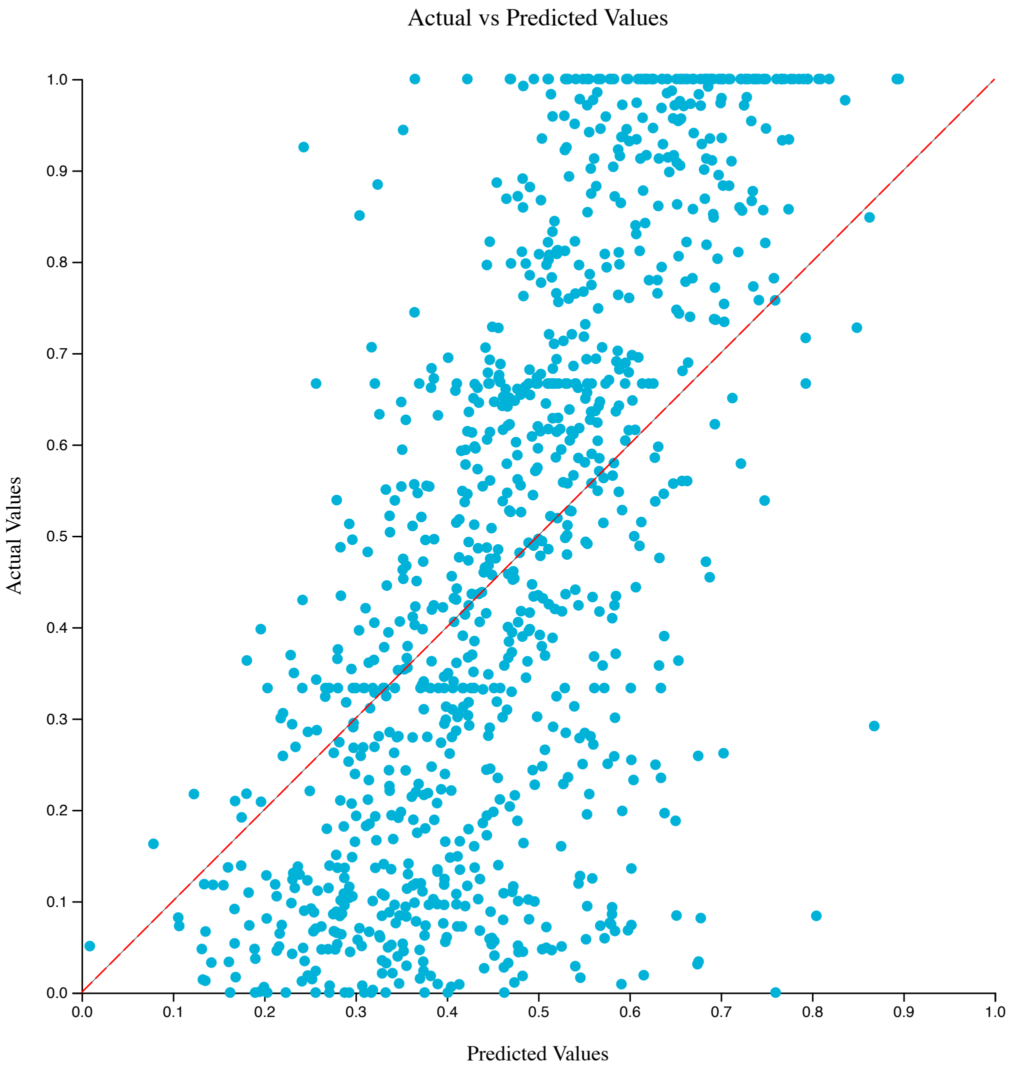
JADBio Results Summary

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 | 0.402 | [0.385, 0.419] Mean Absolute Error | 0.201 | [0.197, 0.206] Mean Squared Error | 0.059 | [0.057, 0.061] Relative Absolute Error | 0.739 | [0.727, 0.756] Relative Squared Error | 0.598 | [0.582, 0.615] Correlation Coefficient | 0.657 | [0.641, 0.672]

Feature Selection

Jadbio selected **all** features in the original dataset for the reference signature. Note that **213** 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	0.3973895837519589	00:00:22.22547	false

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	0.2646570827576885	00:00:03.3847	true
3	Mean Imputation, Mode Imputation, Constant Removal, Standardization	FullSelector	-	Ridge Linear Regression	lambda = 1.0	0.4044612063675912	00:00:09.9991	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	0.2665690813251754	00:00:04.4802	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	0.32766334495780547	00:02:34.154242	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	-0.5999379568948058	00:00:02.2611	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	0.34317294823798783	00:02:32.152641	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	-0.8105012526571775	00:00:04.4537	true

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
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	-0.5278349904017574	00:02:33.153598	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 Random Forests with Mean Squared Error splitting critetion	ntrees = 100, minimum leaf size = 5	0.3599120142979034	00:02:35.155792	true
11	IdentityFactory	NoSelector	-	Trivial model	-	4.3298697960381105e-15	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	0.29150156387962955	00:00:05.5602	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	0.3628689093941111	00:02:37.157305	false
14	Mean Imputation, Mode Imputation, Constant Removal, Standardization	FullSelector	-	Regression Random Forests with Mean Squared Error splitting critetion	ntrees = 100, minimum leaf size = 5	0.39257318384336815	00:00:48.48363	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	0.2959694162140162	00:00:06.6742	true

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
16	Mean Imputation, Mode Imputation, Constant Removal, Standardization	FullSelector	-	Regression Random Forests with Mean Squared Error splitting critetion	ntrees = 100, minimum leaf size = 5	0.39896325622164563	00:00:30.30749	false