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 **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 **01:50:09**.

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

Leading to **11** 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, 11 models were set out to train.

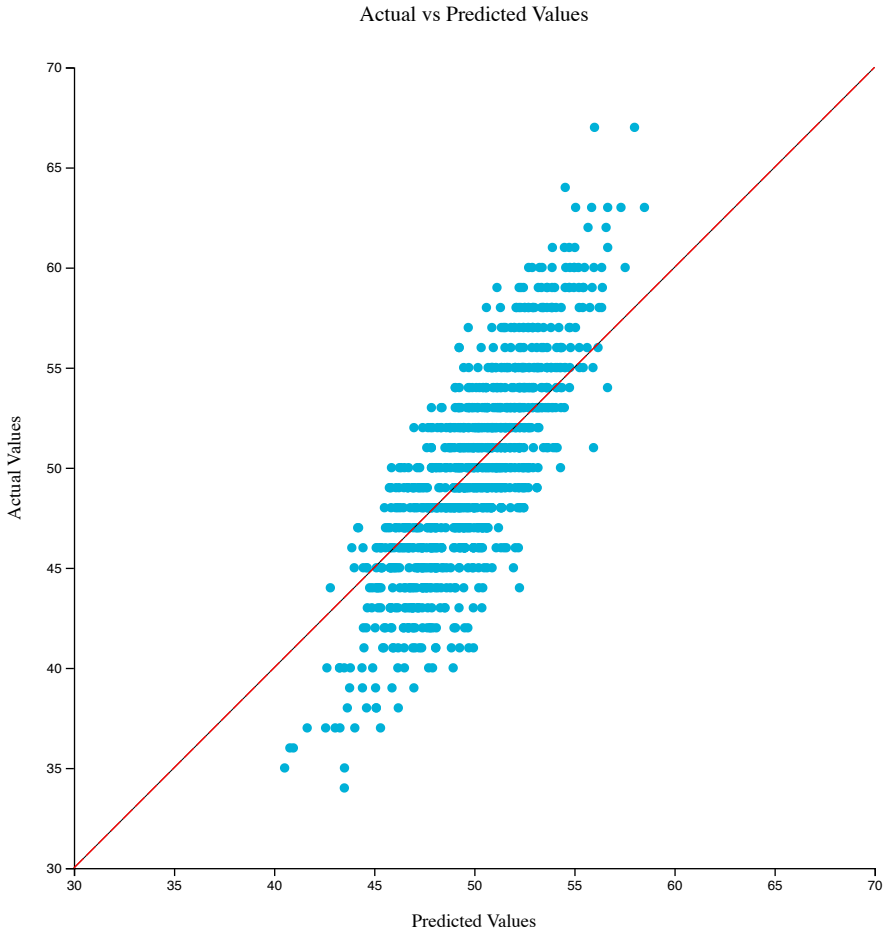
JADBio Results Summary

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	Test-Budgeted Statistically Equivalent Signature (SES) algorithm with hyper-parameters: maxK = 2, alpha = 0.05 and budget = 3 * nvars	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 | 0.587 | [0.576, 0.597] Mean Absolute Error | 2.567 | [2.517, 2.619] Mean Squared Error | 10.312 | [9.936, 10.707] Relative Absolute Error | 0.643 | [0.634, 0.653] Relative Squared Error | 0.414 | [0.403, 0.424] Correlation Coefficient | 0.805 | [0.796, 0.813]

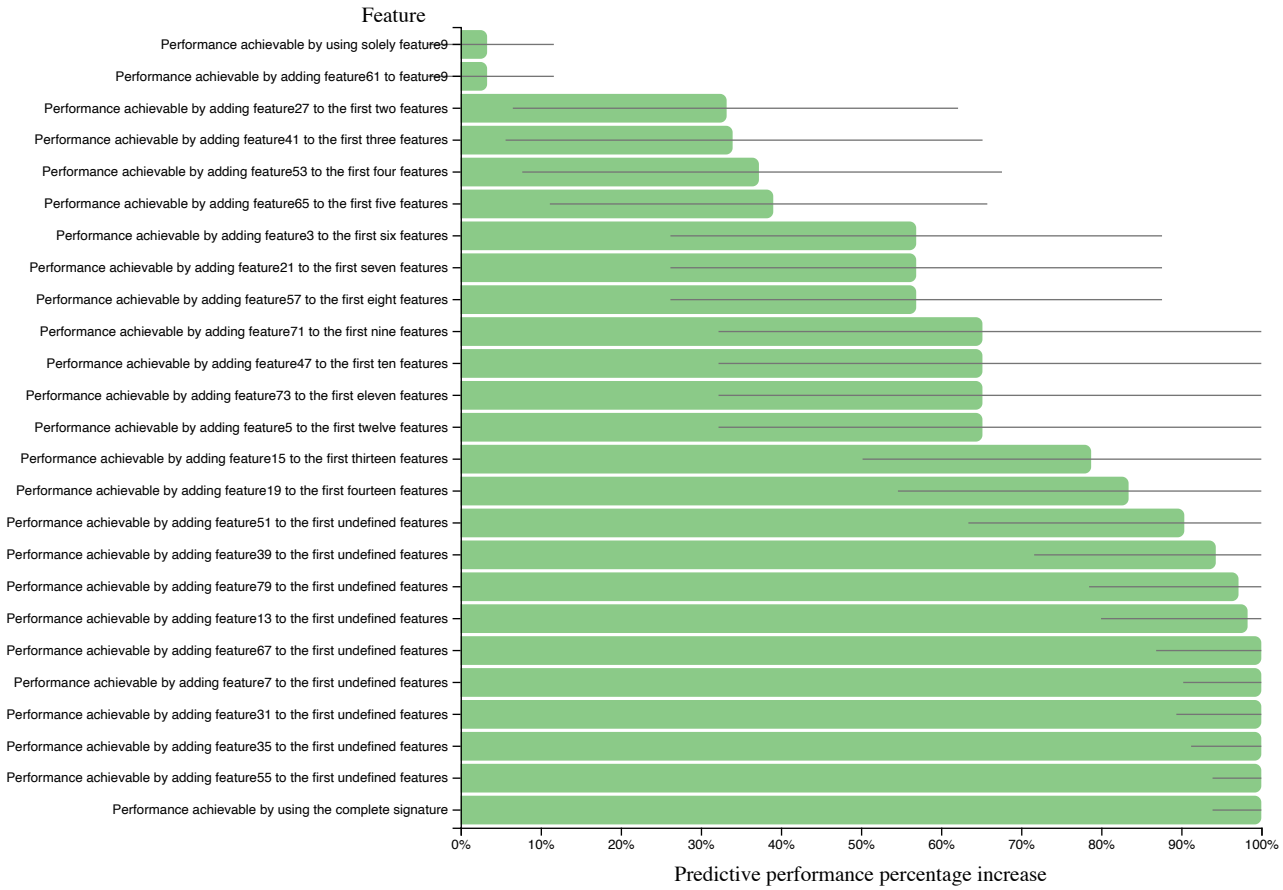
Feature Selection

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

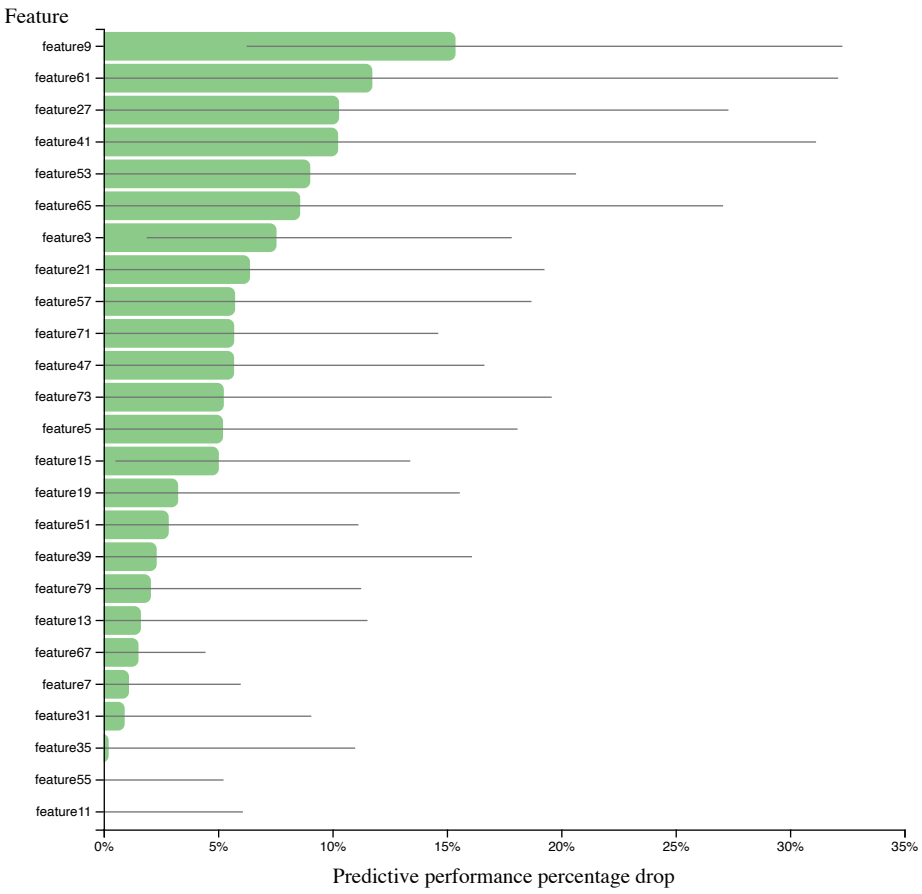
The selected features consist of the following subset called a signature. **There was a single signature identified.** The first signature identified by the system is the set: **feature9, feature13, feature55, feature27, feature79, feature73, feature71, feature7, feature65, feature39, feature35, feature21, feature47, feature51, feature57, feature61, feature3, feature31, feature19, feature41, feature11, feature5, feature15, feature53, feature67** in order

of importance. The following features cannot be substituted with others and still obtain an equal predictive performance: **feature9, feature13, feature55, feature27, feature79, feature73, feature71, feature7, feature65, feature39, feature35, feature21, feature47, feature51, feature57, feature61, feature3, feature31, feature19, feature41, feature11, feature5, feature15, feature53, feature67**.

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	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	00:00:17.17563	true
2	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	00:00:20.20725	true
3	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO Feature Selection	penalty = 1.0	Ridge Linear Regression	lambda = 1.0	0.5069873499447379	00:00:03.3115	false

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
4	IdentityFactory	NoSelector	-	Trivial model	-	1.695310558602614e-13	00:00:00.000	false
5	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	00:00:05.5177	true
6	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	00:00:19.19006	true
7	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	00:00:06.6051	true
8	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	00:00:16.16838	false
9	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	00:00:03.3747	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	0.5140825150891726	00:00:22.22842	false
11	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	00:00:09.9097	true