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 $\mbox{\bf Preliminary}.$

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:

Preprocessing Mode imputation Gontant Removal Feature Selection Standardization Feature Selection Test-Budgeted Statistically Equivalent Signature (SES) alpha 0.05 Modeling LASSO penalties 1.0 Modeling Linear Regression lambdas 1.0 Linear SWR costs 1.0 PolynomialsVR gammas 1.0 Linear SWR degrees 1.0 Embedded Linear Regression 1.0 Embody polynomialsVR gammas 1.0 Linear SWR genitons 1.0 Embody 1.0 1.0 Embody 1.0 1.0 Embody 1.0 1.0 Embody 2.0 1.0 Embody 2.0 1.0 Embody 3.0 1.0 Embody 3.0 1.0 Embody 4.0 1.0 Embody 4.0 1.0 Embody 4.0 1.0<	Algorithm Type	Algorithm	Hyper-parameter	Set of Values
Contant Removal Standardization Feature Selection Test-Budgeted Statistically Equivalent Signature (SES) alpha 0.05 LASSO maxk 2 Modeling Linear Regression lambdas 1.0 LinearSVR costs 1.0 PolynomialSVR gammas 1.0 Linear Regression gammas 1.0 Linear Regression gammas 1.0 Linear Regression gammas 1.0 Linear Regression gammas 1.0 REFSVR gammas 1.0 REFSVR gammas 1.0 Regress 1.0 Regress 1.0 Regress 5 Regress 5 Regress 1.0 Regress 1.0 Regress 5 Regress 1.0 Regress 1.0 Regress 1.0 Regress 1.0 Regress 1.0 <	Preprocessing	Mode imputation		
Feature Selection Test-Budgeted Statistically Equivalent Signature (SES) alpha 0.05 LASSO penalties 1.0 Modeling Linear Regression ambdas 1.0 Linear SVR costs 1.0 PolynomialSVR gammas 1.0 REFSVR pesilons 1.0 REFSVR gammas 1.0 LINEAR SVR gammas 1.0 REFSVR gammas 1.0 REFSVR gammas 1.0 LINEAR SVR gammas 1.0		Mean imputation		
Feature Selection Test-Budgeted Statistically Equivalent Signature (SES) alpha 0.05 Rak 2 Local Local 1.0 Modeling Linear Regression lambdas 1.0 LinearSVR costs 1.0 PolynomialSVR gammas 1.0 LinearSVR epsilons 0.1 REFSVR gammas 1.0 LinearSVR gammas 1.0 LinearSVR gammas 1.0 Reference 3 1.0 LinearSVR gammas 1.0 LinearSVR gammas 1.0 LinearSVR gammas 1.0 LinearSVR gammas 1.0 LinearSVR papilons 0.1 LinearSVR papilons 1.0 LinearSVR papilons 0.1 LinearSVR papilons 0.1 LinearSVR papilons 0.1 LinearSvar papilons 0.1 LinearSvar		Contant Removal		
LASSO penalties 1.0 Modeling Linear Regression lambdas 1.0 LinearSVR costs 1.0 PolynomialSVR gammas 1.0 ELINEAT costs 1.0 REFSVR degrees 3 REFSVR gammas 1.0 Costs 1.0 Costs 1.0 RAMINITARY psilons 1.0 Costs 1.0 Costs<		Standardization		
Modeling Linear Regression lambdas 1.0 Linear SVR costs 1.0 Pollynomial SVR gammas 1.0 Costs 1.0 Linear SVR epsilons 0.1 REFSVR gammas 1.0 Costs 1.0 1.0 Random Forests costs 1.0 Random Forests min leaf sizes 5 vars to split nvars // 3.0, nvars // 5.0, nvars // 7.0 Decision Tree min leaf sizes 5 vars to split nvars // 1.0 vars to split nvars // 1.0 polits to perform 1.0	Feature Selection	Test-Budgeted Statistically Equivalent Signature (SES)	alpha	0.05
Modeling Linear Regression lambdas 1.0 LinearSVR costs 1.0 PolynomialSVR gammas 1.0 costs 1.0 pepilons 0.1 degrees 3 RBFSVR gammas 1.0 costs 1.0 pesilons 0.1 min leaf sizes 5 vars to split nvars // 3.0, nvars // 5.0, nvars // 7.0 policion Tree min leaf sizes 5 min leaf sizes 5 vars to split nvars // 3.0, nvars // 5.0, nvars // 7.0 policion Tree min leaf sizes 5 vars to split nvars // 3.0			maxk	2
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 Decision Tree min leaf sizes 5 vars to split nvars // 1.0 splits to perform 1.0		LASSO	penalties	1.0
PolynomialSVR gammas 1.0	Modeling	Linear Regression	lambdas	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 pecision Tree min leaf sizes 5 vars to split nvars // 3.0 ntrees 100 pecision Tree min leaf sizes 5 vars to split nvars // 1.0 splits to perform 1.0 pecision Tree nvars // 1.0 splits to perform 1.0		LinearSVR	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		PolynomialSVR	gammas	1.0
RBFSVR gammas 1.0			costs	1.0
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			epsilons	0.1
Costs 1.0			degrees	3
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		RBFSVR	gammas	1.0
Random Forests min leaf sizes vars to split nvars // 3.0, nvars // 5.0, nvars // 7.0 splits to perform ntrees 1.0 Decision Tree min leaf sizes 5 vars to split nvars // 1.0 splits to perform 1.0			costs	1.0
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			epsilons	0.1
splits to perform 1.0 ntrees 100 Decision Tree min leaf sizes 5 vars to split nvars // 1.0 splits to perform 1.0		Random Forests	min leaf sizes	5
ntrees 100 Decision Tree min leaf sizes 5 vars to split nvars // 1.0 splits to perform 1.0			vars to split	nvars // 3.0, nvars // 5.0, nvars // 7.0
Decision Tree min leaf sizes 5 vars to split nvars // 1.0 splits to perform 1.0			splits to perform	1.0
vars to split nvars // 1.0 splits to perform 1.0			ntrees	100
splits to perform 1.0		Decision Tree	min leaf sizes	5
			vars to split	nvars // 1.0
alabaa 0.05			splits to perform	1.0
aipnas U.U5			alphas	0.05

https://app.jadbio.com/report/20328

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.

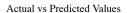
JADBio Results Summary

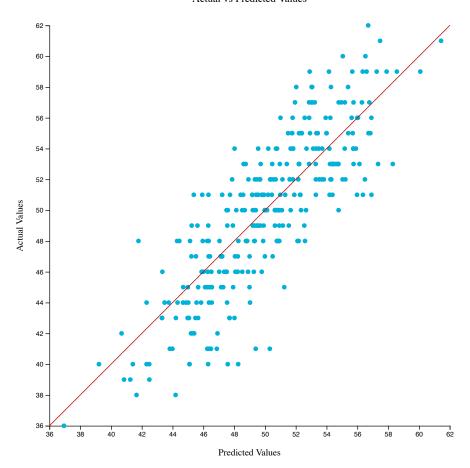
Overview

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: $\max K = 2$, 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:





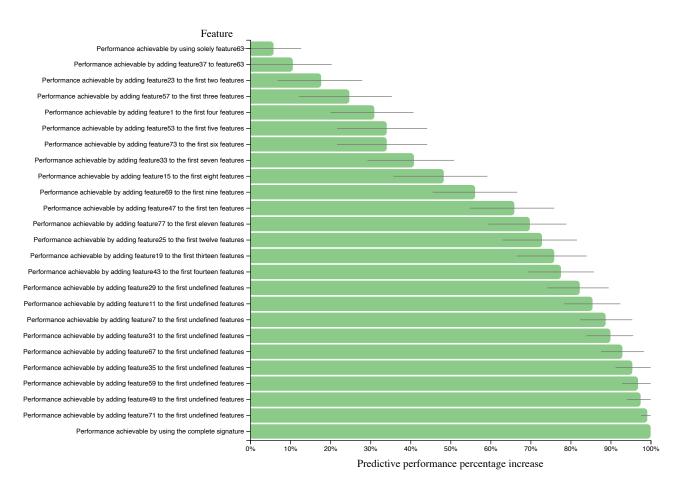
 $\label{lem:mean_stimate} $$ \text{Mean estimate } | CI -- | -- | -- R - squared | 0.682 | [0.599, 0.752] $$ Mean Absolute Error | 2.319 | [2.049, 2.589] $$ Mean Squared Error | 8.366 | [6.675, 10.232] $$ Relative Absolute Error | 0.561 | [0.486, 0.639] $$ Relative Squared Error | 0.326 | [0.253, 0.415] $$ Correlation Coefficient | 0.829 | [0.778, 0.873] $$$

Feature Selection

There were 25 features selected out of the 80 available.

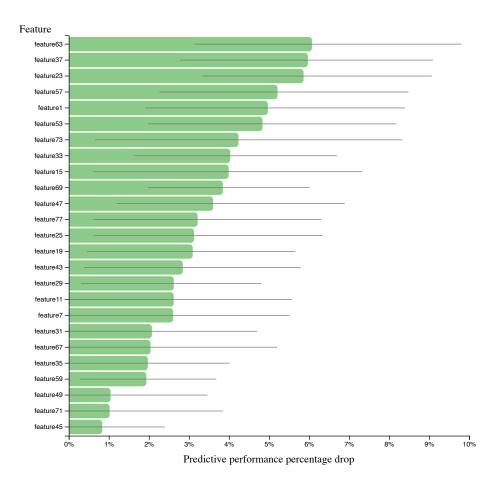
The selected features consist of the following subset called a signature. There was a single signature identified. The first signature identified by the system is the set: feature47, feature15, feature31, feature71, feature57, feature63, feature37, feature49, feature19, feature19, feature11, feature33, feature77, feature59, feature59, feature59, feature69, feature69, feature69, feature69, feature69, feature69, feature69, feature71, feature73, feat

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:

https://app.jadbio.com/report/20328



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

Mean Imputation, Equivalent Constant Signature Removal, Standardization algorithm Mean Test-Imputation, Equivalent Constant Signature Removal, Standardization algorithm Mean Test-Imputation, Equivalent Constant Signature Removal, (SES) Standardization algorithm Mean Test-Imputation, Equivalent Constant Signature Removal, (SES) Standardization algorithm Mean Test-Imputation, Equivalent Constant Signature Removal, (SES) Standardization algorithm Mean Test-Imputation, Equivalent Constant Signature Removal, (SES) Standardization algorithm Mean Imputation, Equivalent Constant Signature Removal, (SES) Standardization algorithm Mean Imputation, Equivalent Constant Signature Removal, (SES) Standardization algorithm Mean Imputation, Mode Constant Signature Removal, Standardization algorithm Mean Imputation, Mode Constant Signature Removal, Standardization Standardization algorithm Mean Imputation, Mode Constant Signature Removal, Standardization Standardizati
Imputation, Mode Statistically Imputation, Constant Signature Removal, (SES) Standardization Signature Removal, (SES) Standardization Signature Removal, (SES) Standardization Signature Removal, (SES) Standardization Signature Removal, (SES) Standardization Signature Removal, (SES) Standardization Selection Removal, Standardization Standardization Selection Removal, Selection Remov
Imputation, Budgeted Mode Statistically Imputation, Equivalent Constant Signature Removal, (SES) Standardization algorithm Mean Imputation, Constant Selection Regression Standardization Mean Imputation, Feature Constant Removal, Standardization Mean Imputation, Feature Constant Removal, Standardization Mean Imputation, Feature Selection Regression Decision Tree with Mean Squared Error splitting critetion Mean Imputation, Selection Regression Decision Tree with Mean Squared Error splitting critetion Mean Imputation, Test- Mean Test- Machines 1.0, gamma = 1.0, epsilon = 0.1 Negression Decision Tree with Mean Squared Error splitting critetion Negression Decision Tree with Mean Squared Error splitting critetion Mean Test-
Imputation, Mode LASSO Imputation, Feature Constant Removal, Standardization Regression Decision Tree with Mean Squared Error splitting critetion Regression Decision Tree with Mean Squared Error splitting critetion Mean Test-
Mean Test- Support
Imputation, Budgeted Mode Statistically Imputation, Equivalent Constant Removal, (SES) Standardization algorithm Statistically alpha = 0.05, Regression Standardization Regression (SVR) of type epsilon-SVR MaxK = 2, Vector kernel = 'Linear Kernel', cost = 1.0, epsilon = 0.6920239497493902 00:00:00.843 false 0.6920239497493902 00:00:00:00.843 false 0.6920239497493902 00:00:00:00:00.843 false 0.6920239497493902 00:00:00:00:00:00:00:00:00:00:00:00:00:
Mean Support Kernel = 'Radial Vector Basis Function Regression Removal, Standardization Standardization Support Support Kernel = 'Radial Vector Basis Function Regression Kernel', cost = Machines 1.0, gamma = (SVR) of type 1.0, epsilon = epsilon-SVR 0.1 Vector Constant Vector Basis Function Regression Kernel', cost = 0.23512416107364076 00:00:00.127 true Vector Vector Vector Basis Function Regression Kernel = 'Radial Vector Basis Function Regression Kernel = 'Radial Vector Basis Function Regression Kernel = 'Radial Vector Basis Function Vector Regression Kernel', cost = 0.23512416107364076 00:00:00.127 true Vector Vec
Mean Test- Imputation, Budgeted Mode Statistically Imputation, Equivalent Constant Signature Removal, (SES) Standardization algorithm Regression Decision Tree with Mean budget = 3 * Squared Error splitting critetion Regression maxK = 2, Decision Tree with Mean budget = 3 * Squared Error splitting critetion Regression minimum leaf size = 5, alpha = -1.051857507675416 00:00:01.1174 true 0.05
Mean Test- Imputation, Budgeted Mode Statistically Imputation, Equivalent Constant Signature Removal, (SES) Standardization algorithm Test- Regression Random Forests with ntrees = 100, Mean minimum leaf Squared Error size = 5 splitting critetion Regression Random Forests with ntrees = 100, Mean minimum leaf Squared Error size = 5 splitting critetion
12 IdentityFactory NoSelector - Trivial model - 9.251858538542972e- 16 00:00:00.000 false

https://app.jadbio.com/report/20328 5/6

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

https://app.jadbio.com/report/20328 6/6