# JADBio Description of Performed Analysis

## Setup

JADBio version **1.4.69** ran on dataset **regression_peptides_binary** with **67769** samples and **400** 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 **01:51:28**.

## 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.

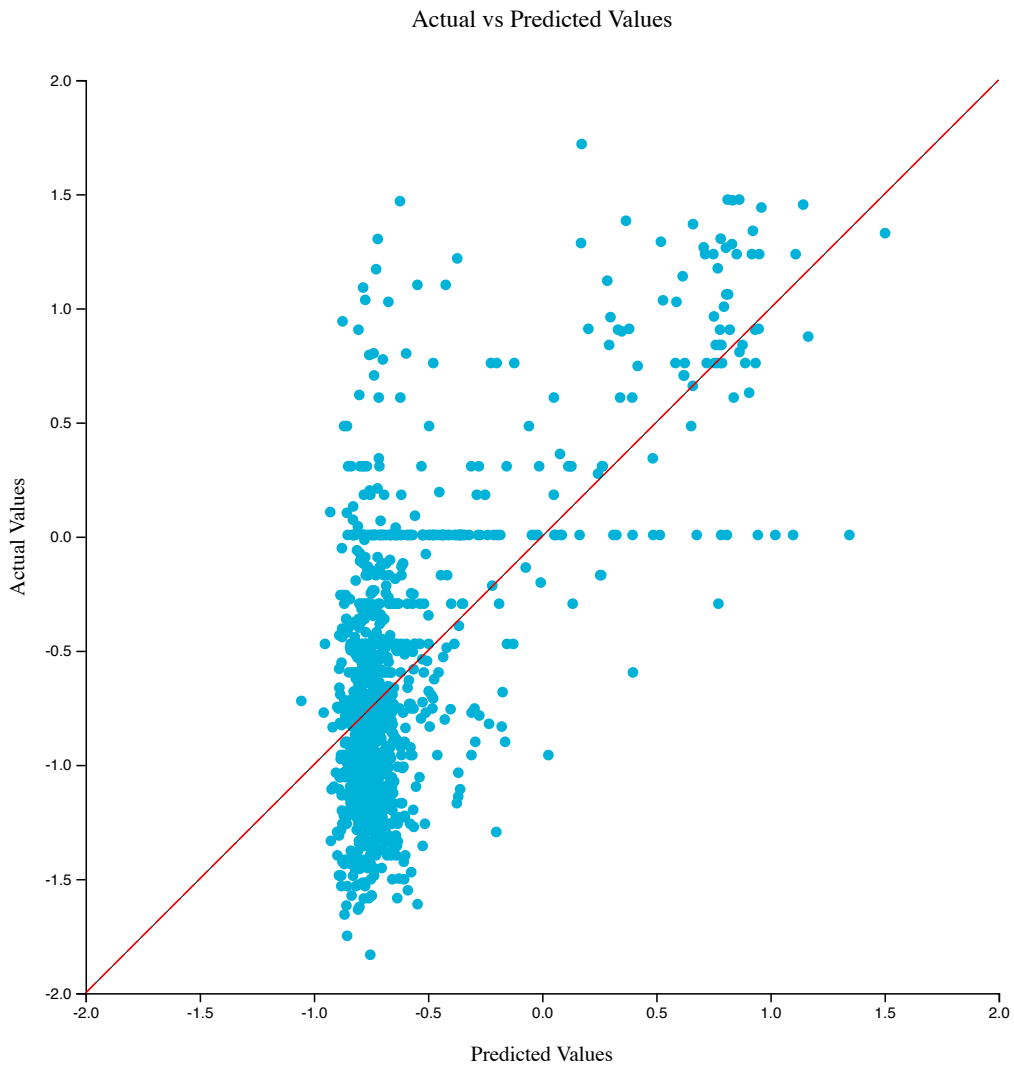
## JADBio Results Summary

### 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 | Regression Random Forests training 100 trees with Mean Squared Error splitting critetion, minimum leaf size = 5, and variables to split = nvars // 3.0 |

The R-squared is shown in the figure below:




Metric | Mean estimate | CI --- | --- | --- R-squared | 0.452 | [0.423, 0.485] Mean Absolute Error | 0.382 | [0.372, 0.392] Mean Squared Error | 0.233 | [0.221, 0.245] Relative Absolute Error | 0.743 | [0.724, 0.762] Relative Squared Error | 0.548 | [0.517, 0.577] Correlation Coefficient | 0.675 | [0.652, 0.700]

## Feature Selection

Jadbio selected **all** features in the original dataset for the reference signature. Note that **377** 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 | 0.44735605938379963 | 00:00:53.53376 | 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 | 0.12197628613293521 | 00:00:06.6402 | true |
| 3 | Mean Imputation, Mode Imputation, Constant Removal, Standardization | FullSelector | - | Ridge Linear Regression | lambda = 1.0 | 0.2247218354937398 | 00:00:25.25511 | 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 | 0.2354161526344407 | 00:00:07.7166 | 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 | 0.21701311598752182 | 00:05:23.323122 | 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 | -5.772042447098979 | 00:00:06.6874 | 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 | 0.11987874443301783 | 00:05:22.322346 | true |
| 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 | -4.945890781060981 | 00:00:07.7311 | 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 | -5.058838167365365 | 00:05:23.323288 | 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 | 0.23644479902228366 | 00:05:23.323761 | true |
| 11 | IdentityFactory | NoSelector | - | Trivial model | - | 1.0769163338864018e-14 | 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 | 0.2641053053769815 | 00:00:07.7914 | 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 | 0.24275069941841088 | 00:05:24.324694 | 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 | 0.45310886131672323 | 00:02:01.121327 | 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 | 0.27349547720819356 | 00:00:08.8916 | false |

| 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 | 0.4521954654689656 | 00:01:12.72895 | false |