# JADBio Description of Performed Analysis

## Setup

JADBio version **1.4.69** ran on dataset **hollerer_rbs_medium_train_binary** with **50000** samples and **68** 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 **00:53:34**.

## 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.

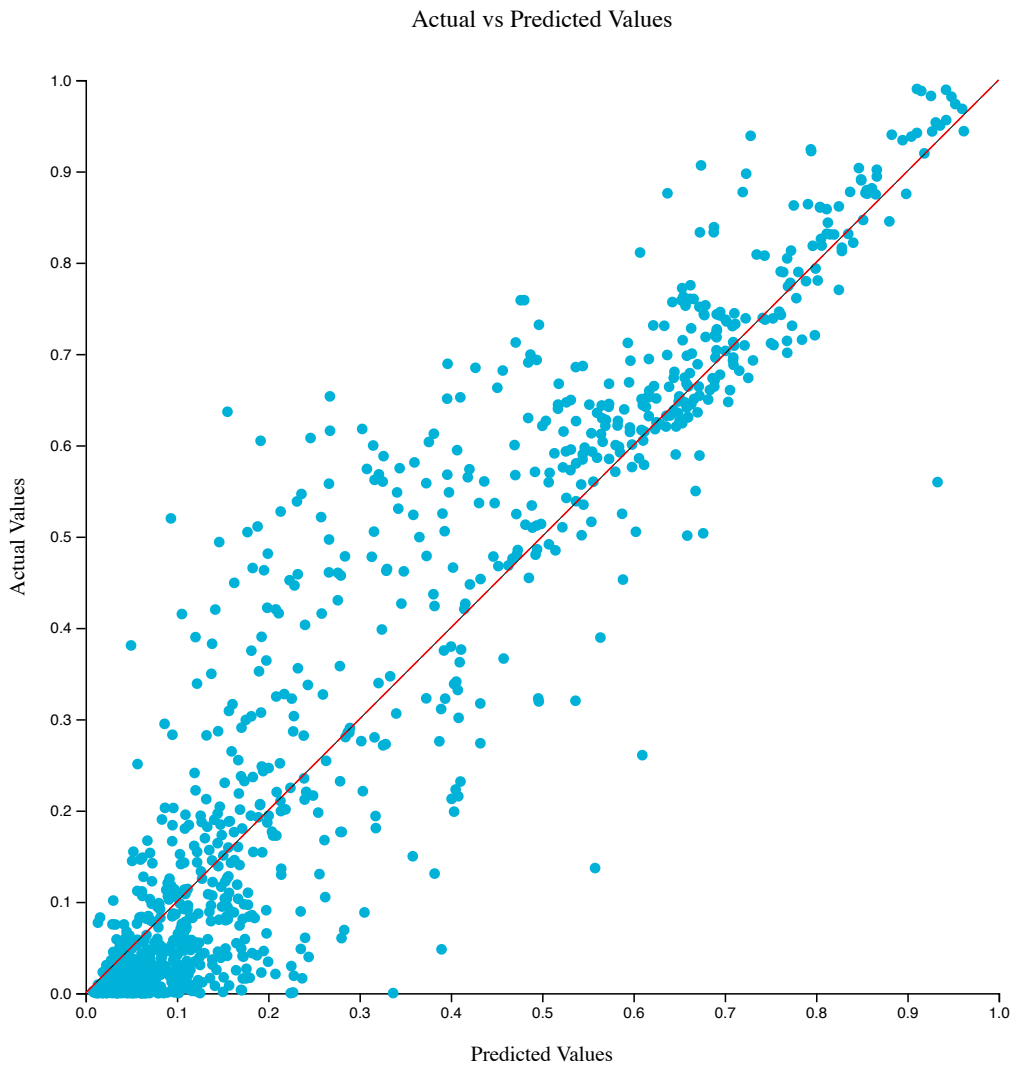
## JADBio Results Summary

### 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 | 0.874 | [0.861, 0.886] Mean Absolute Error | 0.073 | [0.070, 0.075] Mean Squared Error | 0.011 | [0.010, 0.012] Relative Absolute Error | 0.272 | [0.260, 0.284] Relative Squared Error | 0.126 | [0.114, 0.140] Correlation Coefficient | 0.936 | [0.928, 0.943]

## Feature Selection

Jadbio selected **all** features in the original dataset for the reference signature. Note that **43** 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 | 0.8521290427892383 | 00:00:04.4017 | 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 | 0.724157000185889 | 00:00:01.1634 | true |
| 3 | Mean Imputation, Mode Imputation, Constant Removal, Standardization | FullSelector | - | Ridge Linear Regression | lambda = 1.0 | 0.757359229592519 | 00:00:00.798 | 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 | 0.7872711057945774 | 00:00:02.2404 | 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 | 0.7582769327027066 | 00:00:31.31380 | 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 | -1.2767235545806943 | 00:00:01.1034 | 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 | 0.7045955612325286 | 00:00:30.30590 | 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 | -2.1496514818227093 | 00:00:02.2180 | 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 | -1.6689568614591859 | 00:00:31.31136 | 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 | 0.767192751090939 | 00:00:31.31994 | false |
| 11 | IdentityFactory | NoSelector | - | Trivial model | - | -8.881784197001252e-16 | 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 | 0.8071598919208639 | 00:00:03.3030 | 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 | 0.7668100626935437 | 00:00:32.32794 | 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 | 0.8736044160101505 | 00:00:08.8339 | 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 | 0.8116227502420829 | 00:00:03.3845 | false |
| 16 | Mean Imputation, Mode Imputation, Constant Removal, Standardization | FullSelector | - | Regression Random Forests with Mean Squared Error splitting critetion | ntrees = 100, minimum leaf size = 5 | 0.8635960622787634 | 00:00:05.5388 | false |