# JADBio Description of Performed Analysis

## Visit analysis

## Setup

JADBio version **1.4.118** ran on dataset **smartphone_cleaned_v5** with **980** samples and **23** features to create a predictive model for outcome named **price**. 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 **Quick**.
The number of CPU cores to use for the analysis was set to **1**.
The execution time was **00:00:19**.

## 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 | Mean Imputation | | |
| | Mode Imputation | | |
| | Constant Removal | | |
| | Variable Normalization | | |
| Feature Selection | LASSO | penalty | 1.0 |
| | Test-Budgeted Statistically Equivalent Signature (SES) | alpha | 0.05 |
| | | maxK | 2.0 |
| Modeling | Ridge Linear Regression | lambda | 1.0 |
| | Regression Random Forest with Mean Squared Error splitting criterion | minLeafSize | 5.0 |
| | | nTrees | 100 |

Leading to **6** 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, 27 models were set out to train.

Eventually, 27 had their estimation protocol completed. A detailed report of the above is available at Visit analysis

# 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 | LASSO Feature Selection (penalty=1.0) | Regression Random Forest training 100 trees with Mean Squared Error splitting criterion, minimum leaf size = 5, splits = 1, alpha = 1, and variables to split = nvars // 7.0 |

The R-squared is shown in the figure below:

| Metric | Mean estimate | CI |
|---|---|---|
| R-squared | 0.837 | [0.629, 0.916] |
| Mean Absolute Error | 5847.145 | [4663.062, 7563.961] |
| Mean Squared Error | 102805411.775 | [53888465.369, 173855615.929] |
| Relative Absolute Error | 0.346 | [0.268, 0.475] |
| Relative Squared Error | 0.184 | [0.090, 0.478] |
| Correlation Coefficient | 0.927 | [0.861, 0.963] |

## Feature Selection

There were **18** features selected out of the **23** 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: **brand_name, has_nfc, processor_brand, num_cores, processor_speed, battery_capacity, fast_charging_available, fast_charging, ram_capacity, internal_memory, screen_size, refresh_rate, resolution, num_rear_cameras, num_front_cameras, os, primary_camera_rear, extended_memory_available** in order of importance. The following features cannot be substituted with others and still obtain an equal predictive performance: **brand_name, has_nfc, processor_brand, num_cores, processor_speed, battery_capacity, fast_charging_available, fast_charging, ram_capacity, internal_memory, screen_size, refresh_rate,**

**resolution, num_rear_cameras, num_front_cameras, os, primary_camera_rear, extended_memory_available**.

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 | LASSO | penalty = 1.0 | Ridge Linear Regression | lambda = 1.0 | 0.7103654268105509 | 00:00:00.973 | false |
| 2 | Mean Imputation, Mode Imputation, Constant Removal, Standardization | LASSO | penalty = 1.0 | Regression Random Forest with Mean Squared Error splitting criterion | ntrees = 100, minimum leaf size = 5 | 0.8693862591855136 | 00:00:00.988 | false |
| 3 | IdentityFactory | FullSelector | - | Trivial model | - | -5.181040781584064e-16 | 00:00:00.000 | false |
| 4 | Mean Imputation, Mode Imputation, Constant Removal, Standardization | LASSO | penalty = 1.0 | Regression Random Forest with Mean Squared Error splitting criterion | ntrees = 100, minimum leaf size = 5 | 0.8679272667667008 | 00:00:00.998 | false |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 5 | Mean Imputation, Mode Imputation, Constant Removal, Standardization | Test-Budgeted Statistically Equivalent Signature (SES) | maxK = 2, alpha = 0.05, budget = 3 * nvars | Regression Random Forest with Mean Squared Error splitting criterion | ntrees = 100, minimum leaf size = 5 | 0.7320232033232124 | 00:00:00.200 | false |
| 6 | Mean Imputation, Mode Imputation, Constant Removal, Standardization | LASSO | penalty = 1.0 | Regression Random Forest with Mean Squared Error splitting criterion | ntrees = 100, minimum leaf size = 5 | 0.8538581573817512 | 00:00:01.1135 | false |