



ROBERT v 1.0.2 2023/07/15 12:10:37

Citation: ROBERT v 1.0.2, Dalmau, D.; Alegre-Requena, J. V., 2023. <https://github.com/jvalegre/robert>

Command line used in ROBERT: `robert --ignore [Name] --names Name --y Target_values --csv_name Robert_example.csv --csv_test Robert_example_test.csv`

Total execution time: 133.11 seconds



## CURATE

o Starting data curation with the CURATE module

o Database Robert\_example.csv loaded successfully, including:

- 37 datapoints
- 11 accepted descriptors
- 1 ignored descriptors
- 0 discarded descriptors

o Analyzing categorical variables

A total of 1 categorical variables were converted using the onehot mode in the categorical option

Initial descriptors:

- x4

Generated descriptors:

- Csub-Csub
- Csub-H
- Csub-O
- H-O

o Duplication filters activated

Excluded datapoints:

- No datapoints were removed

o Correlation filter activated with these thresholds:  $\text{thres\_x} = 0.9$ ,  $\text{thres\_y} = 0.001$

Excluded descriptors:

- x3:  $R^2 = 1.0$  with x1
- x1:  $R^2 = 0.96$  with x6

o 14 columns remaining after applying duplicate and correlation filters:

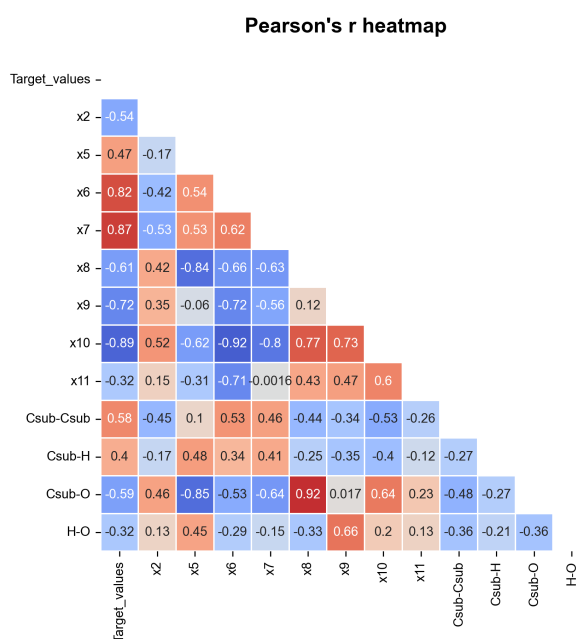
- Name
- Target\_values
- x2
- x5
- x6
- x7
- x8
- x9
- x10
- x11
- Csub-Csub

- Csub-H
- Csub-O
- H-O

- o The Pearson heatmap was stored in CURATE/Pearson\_heatmap.png.
- o The curated database was stored in CURATE/Robert\_example\_CURATE.csv.

Time CURATE: 0.57 seconds

----- Images generated by the CURATE module -----

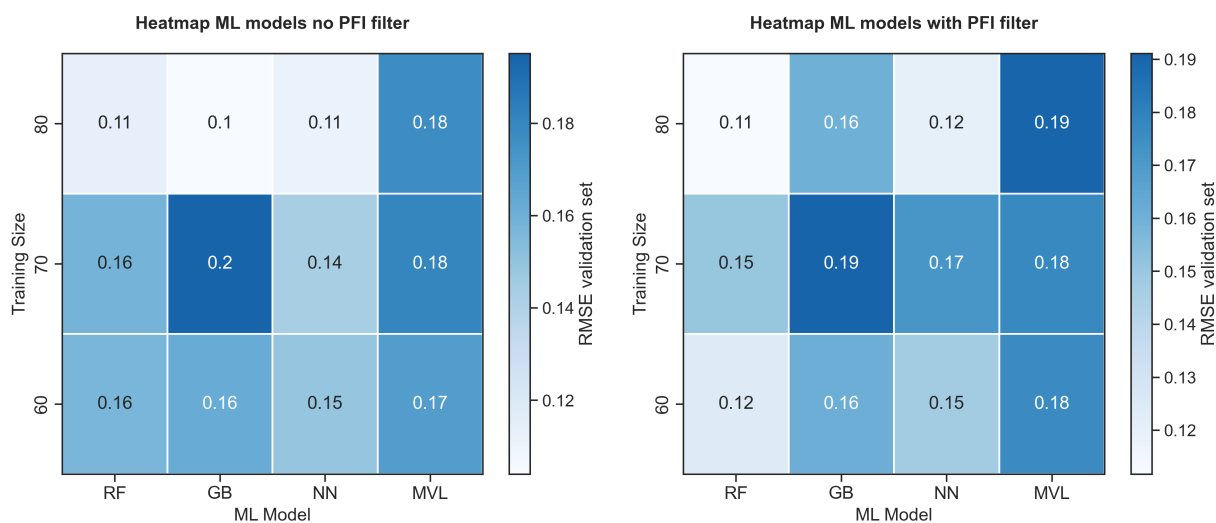


## GENERATE

- o Starting generation of ML models with the GENERATE module
- o Database Robert\_example\_CURATE.csv loaded successfully, including:
  - 37 datapoints
  - 12 accepted descriptors
  - 1 ignored descriptors
  - 0 discarded descriptors
- x WARNING! The database contains 37 datapoints, the 90% training size will be excluded (too few validation points to reach a reliable result). You can include this size using "--filter\_train False".
- o Starting heatmap scan with 4 ML models (['RF', 'GB', 'NN', 'MVL']) and 3 training sizes ([60, 70, 80]).
  - 72 models were tested, for more information check the GENERATE\_data.dat file in the GENERATE folder
- o Heatmap ML models no PFI filter succesfully created in GENERATE/Raw\_data
- o Heatmap ML models with PFI filter succesfully created in GENERATE/Raw\_data

Time GENERATE: 123.51 seconds

----- Images generated by the GENERATE module -----



- o Starting tests to verify the prediction ability of the ML models with the VERIFY module

----- Starting model with all variables (No PFI) -----

- o ML model GB\_80 (with no PFI filter) and Xy database were loaded, including:
    - Target value: Target\_values
    - Model: GB
    - Descriptors: ['x2', 'x5', 'x6', 'x7', 'x8', 'x9', 'x10', 'x11', 'Csub-Csub', 'Csub-H', 'Csub-O', 'H-O']
    - Training points: 29
    - Validation points: 8
  - o VERIFY donut plots saved in VERIFY/VERIFY\_tests\_GB\_80\_No\_PFI.png
  - o VERIFY test values saved in VERIFY/VERIFY\_tests\_GB\_80\_No\_PFI.dat
- Results of the VERIFY tests:
- Original score (train set in CV): RMSE = 0.031, +- 20% threshold (thres\_test option):
- 5-fold CV: NOT DETERMINED, data splitting was done with KN. CV result: RMSE = 0.22
- Original score (validation set): RMSE = 0.1, +- 20% threshold (thres\_test option):
- o y\_mean: PASSED, RMSE = 0.66 is higher than the threshold (0.12)
  - o y\_shuffle: PASSED, RMSE = 0.8 is higher than the threshold (0.12)
  - o onehot: PASSED, RMSE = 0.2 is higher than the threshold (0.12)

----- Starting model with PFI filter (only important descriptors used) -----

- o ML model RF\_80\_PFI (with PFI filter) and Xy database were loaded, including:
  - Target value: Target\_values
  - Model: RF
  - Descriptors: ['x6', 'x7', 'x9', 'x10']
  - Training points: 29
  - Validation points: 8

- o VERIFY donut plots saved in VERIFY/VERIFY\_tests\_RF\_80\_PFI.png
- o VERIFY test values saved in VERIFY/VERIFY\_tests\_RF\_80\_PFI.dat

Results of the VERIFY tests:

Original score (train set in CV): RMSE = 0.11, +- 20% threshold (thres\_test option):

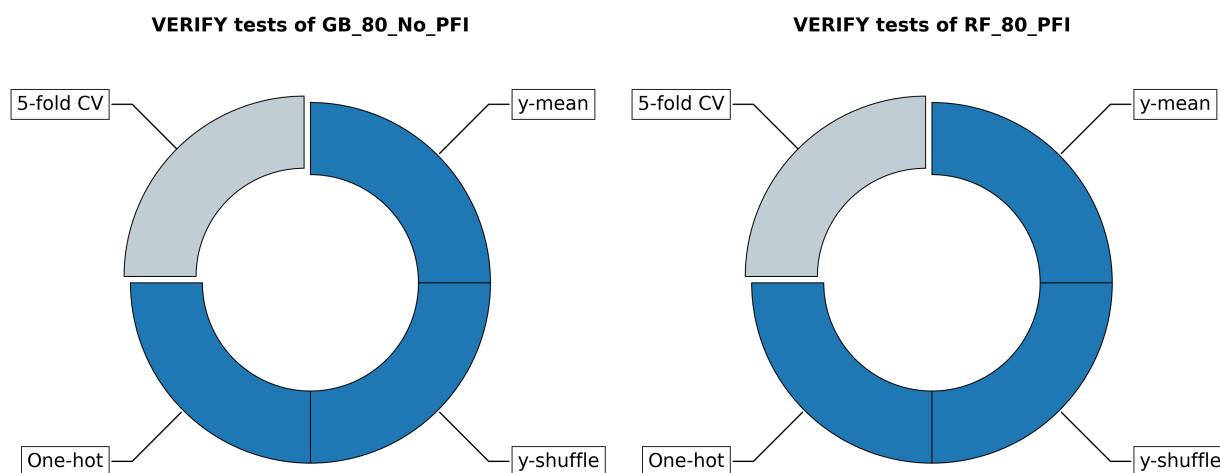
- 5-fold CV: NOT DETERMINED, data splitting was done with KN. CV result: RMSE = 0.22

Original score (validation set): RMSE = 0.11, +- 20% threshold (thres\_test option):

- o y\_mean: PASSED, RMSE = 0.66 is higher than the threshold (0.13)
- o y\_shuffle: PASSED, RMSE = 0.88 is higher than the threshold (0.13)
- o onehot: PASSED, RMSE = 0.21 is higher than the threshold (0.13)

Time VERIFY: 1.95 seconds

----- Images generated by the VERIFY module -----



## PREDICT

- o Representation of predictions and analysis of ML models with the PREDICT module

----- Starting model with all variables (No PFI) -----

- o ML model GB\_80 (with no PFI filter) and Xy database were loaded, including:
  - Target value: Target\_values
  - Model: GB
  - Descriptors: ['x2', 'x5', 'x6', 'x7', 'x8', 'x9', 'x10', 'x11', 'Csub-Csub', 'Csub-H', 'Csub-O', 'H-O']
  - Training points: 29
  - Validation points: 8
- o Test set Robert\_example\_test.csv loaded successfully, including:
  - 9 datapoints
- x There are missing descriptors in the test set! Looking for categorical variables converted from CURATE
- o The missing descriptors were successfully created
  - Train set with predicted results: GB\_80\_train\_No\_PFI.csv
  - Validation set with predicted results: GB\_80\_valid\_No\_PFI.csv
  - Test set with predicted results: GB\_80\_test\_No\_PFI.csv

- o Saving graphs and CSV databases in:
  - Graph in: PREDICT/Results\_GB\_80\_No\_PFI.png
- o Results saved in PREDICT/Results\_GB\_80\_No\_PFI.dat:
  - Points Train:Validation:Test = 29:8:9
  - Proportion Train:Validation:Test = 63:17:20
  - Number of descriptors = 12
  - Proportion points:descriptors = 37:12
  - Train : R2 = 1.0, MAE = 0.025, RMSE = 0.031
  - Validation : R2 = 0.98, MAE = 0.077, RMSE = 0.1
  - Test : R2 = 1.0, MAE = 0.033, RMSE = 0.038
- o SHAP plot saved in PREDICT/SHAP\_GB\_80\_No\_PFI.png
- o SHAP values saved in PREDICT/SHAP\_GB\_80\_No\_PFI.dat:
  - x6 = min: -0.33, max: 0.18
  - x10 = min: -0.39, max: 0.17
  - x5 = min: -0.1, max: 0.16
  - x2 = min: -0.067, max: 0.1
  - x7 = min: -0.094, max: 0.082
  - x9 = min: -0.13, max: 0.051
  - Csub-Csub = min: -0.056, max: 0.03
  - Csub-O = min: -0.06, max: 0.019
  - x11 = min: -0.043, max: 0.017
  - x8 = min: -0.072, max: 0.017
  - Csub-H = min: -0.019, max: 0.012
  - H-O = min: 0.0, max: 0.0
- o PFI plot saved in PREDICT/PFI\_GB\_80\_No\_PFI.png
- o PFI values saved in PREDICT/PFI\_GB\_80\_No\_PFI.dat:
  - Original score (from model.score, R2) = 0.98
  - x10 = 0.17 +- 0.061
  - x6 = 0.12 +- 0.053
  - x5 = 0.038 +- 0.029
  - x7 = 0.026 +- 0.019
  - x9 = 0.019 +- 0.01
  - x2 = 0.017 +- 0.0096
  - Csub-H = 0.0067 +- 0.0055
  - x11 = 0.0054 +- 0.0058
  - Csub-Csub = 0.0032 +- 0.0062
  - x8 = 0.0025 +- 0.0023
- o Outliers plot saved in PREDICT/Outliers\_GB\_80\_No\_PFI.png
- o Outlier values saved in PREDICT/Outliers\_GB\_80\_No\_PFI.dat:
  - Train: 1 outliers out of 29 datapoints (3.4%)
    - 21 (2.8 SDs)
  - Validation: 1 outliers out of 8 datapoints (12.5%)
    - 22 (2.4 SDs)
  - Test: 0 outliers out of 9 datapoints (0.0%)

----- Starting model with PFI filter (only important descriptors used) -----

- o ML model RF\_80\_PFI (with PFI filter) and Xy database were loaded, including:

- Target value: Target\_values
- Model: RF
- Descriptors: ['x6', 'x7', 'x9', 'x10']
- Training points: 29
- Validation points: 8
  
- o Test set Robert\_example\_test.csv loaded successfully, including:
  - 9 datapoints
  - Train set with predicted results: RF\_80\_train\_PFI.csv
  - Validation set with predicted results: RF\_80\_valid\_PFI.csv
  - Test set with predicted results: RF\_80\_test\_PFI.csv
  
- o Saving graphs and CSV databases in:
  - Graph in: PREDICT/Results\_RF\_80\_PFI.png
  
- o Results saved in PREDICT/Results\_RF\_80\_PFI.dat:
  - Points Train:Validation:Test = 29:8:9
  - Proportion Train:Validation:Test = 63:17:20
  - Number of descriptors = 4
  - Proportion points:descriptors = 37:4
  - Train : R2 = 0.98, MAE = 0.076, RMSE = 0.11
  - Validation : R2 = 0.97, MAE = 0.084, RMSE = 0.11
  - Test : R2 = 0.99, MAE = 0.056, RMSE = 0.072
  
- o SHAP plot saved in PREDICT/SHAP\_RF\_80\_PFI.png
- o SHAP values saved in PREDICT/SHAP\_RF\_80\_PFI.dat:
  - x6 = min: -0.38, max: 0.21
  - x7 = min: -0.26, max: 0.18
  - x10 = min: -0.35, max: 0.13
  - x9 = min: -0.2, max: 0.062
  
- o PFI plot saved in PREDICT/PFI\_RF\_80\_PFI.png
- o PFI values saved in PREDICT/PFI\_RF\_80\_PFI.dat:
  - Original score (from model.score, R2) = 0.97
  - x6 = 0.24 +- 0.11
  - x7 = 0.18 +- 0.064
  - x10 = 0.15 +- 0.073
  - x9 = 0.047 +- 0.04
  
- o Outliers plot saved in PREDICT/Outliers\_RF\_80\_PFI.png
- o Outlier values saved in PREDICT/Outliers\_RF\_80\_PFI.dat:
  - Train: 1 outliers out of 29 datapoints (3.4%)
    - 21 (4.1 SDs)
  - Validation: 0 outliers out of 8 datapoints (0.0%)
  - Test: 0 outliers out of 9 datapoints (0.0%)

Time PREDICT: 7.08 seconds

----- Images and summary generated by the PREDICT module -----

**No PFI (all descriptors):**

Results\_GB\_80\_No\_PFI.dat:

- Points Train:Validation:Test = 29:8:9
- Proportion Train:Validation:Test = 63:17:20
- Number of descriptors = 12
- Proportion points:descriptors = 37:12
- Train : R2 = 1.0, MAE = 0.025, RMSE = 0.031
- Validation : R2 = 0.98, MAE = 0.077, RMSE = 0.1
- Test : R2 = 1.0, MAE = 0.033, RMSE = 0.038

**PFI (only important descriptors):**

Results\_RF\_80\_PFI.dat:

- Points Train:Validation:Test = 29:8:9
- Proportion Train:Validation:Test = 63:17:20
- Number of descriptors = 4
- Proportion points:descriptors = 37:4
- Train : R2 = 0.98, MAE = 0.076, RMSE = 0.11
- Validation : R2 = 0.97, MAE = 0.084, RMSE = 0.11
- Test : R2 = 0.99, MAE = 0.056, RMSE = 0.072

