ROBERT v 1.0.2 2023/07/12 18:06:10
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

---

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: thres_x = 0.9, 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

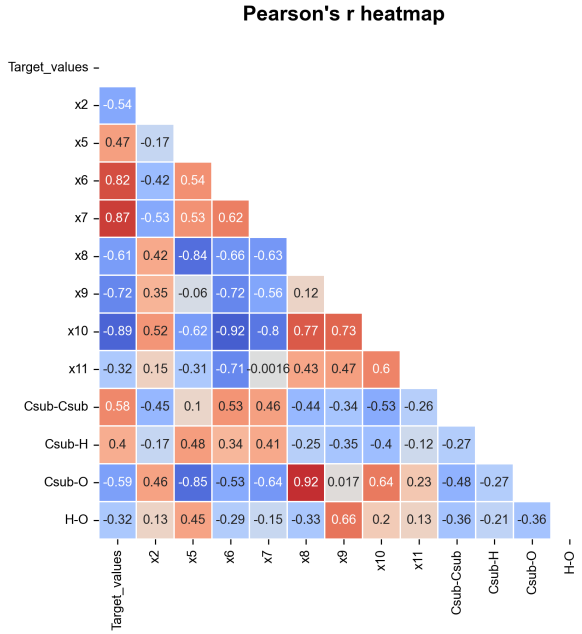
- H-O

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: 0.86 seconds

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



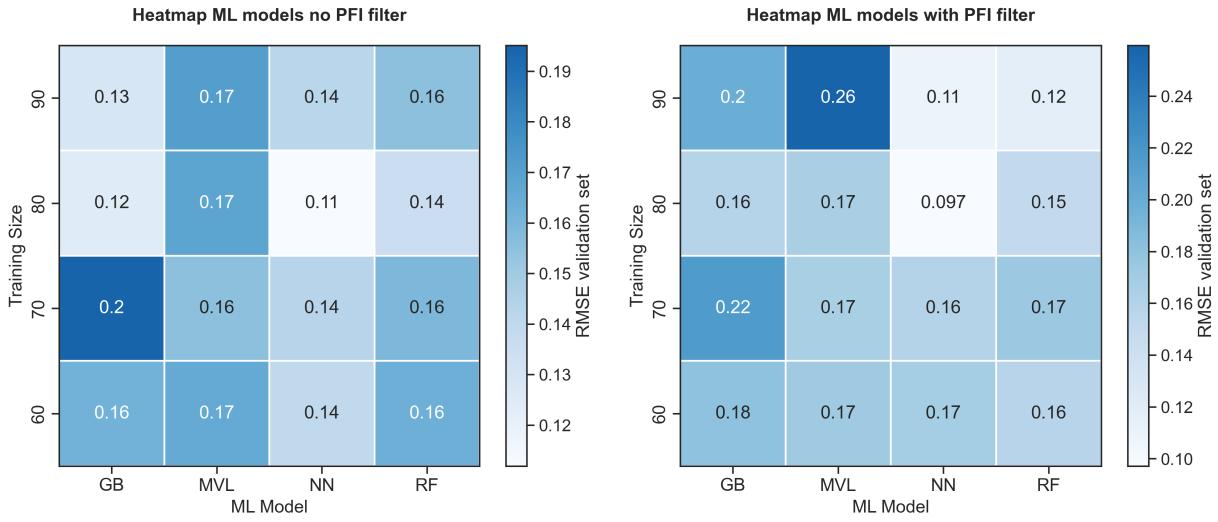
GENERATE

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

o Starting heatmap scan with 4 ML models (['RF', 'GB', 'NN', 'MVL']) and 4 training sizes ([60, 70, 80, 90]).
- 96 models were tested, for more information check the GENERATE_data.dat file in the GENERATE folder

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



VERIFY

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 NN_80 (with no PFI filter) and Xy database were loaded, including:
- Target value: Target_values
- Model: NN
- 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_NN_80_No_PFI.png
o VERIFY test values saved in VERIFY/VERIFY_tests_NN_80_No_PFI.dat
Results of the VERIFY tests:
Original score (train set in CV): RMSE = 0.12, +- 20% threshold (thres_test option):
- 5-fold CV: NOT DETERMINED, data splitting was done with KN. CV result: RMSE = 0.4
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 = 1.1 is higher than the threshold (0.13)
o onehot: PASSED, RMSE = 0.2 is higher than the threshold (0.13)

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

o ML model NN_80_PFI (with PFI filter) and Xy database were loaded, including:
- Target value: Target_values
- Model: NN
- Descriptors: ['x5', 'x7', 'x8', 'x9', 'x11', 'Csub-Csub', 'Csub-H']
- Training points: 29
- Validation points: 8

o VERIFY donut plots saved in VERIFY/VERIFY_tests_NN_80_PFI.png
o VERIFY test values saved in VERIFY/VERIFY_tests_NN_80_PFI.dat
Results of the VERIFY tests:
Original score (train set in CV): RMSE = 0.18, +- 20% threshold (thres_test option):

- 5-fold CV: NOT DETERMINED, data splitting was done with KN. CV result: RMSE = 0.27

Original score (validation set): RMSE = 0.097, +- 20% threshold (thres_test option):

o y_mean: PASSED, RMSE = 0.49 is higher than the threshold (0.12)

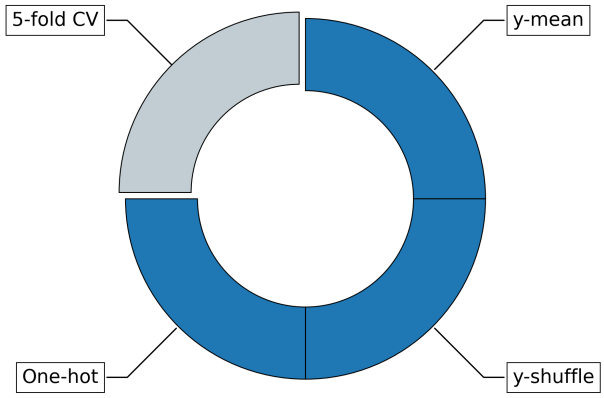
o y_shuffle: PASSED, RMSE = 0.9 is higher than the threshold (0.12)

o onehot: PASSED, RMSE = 0.27 is higher than the threshold (0.12)

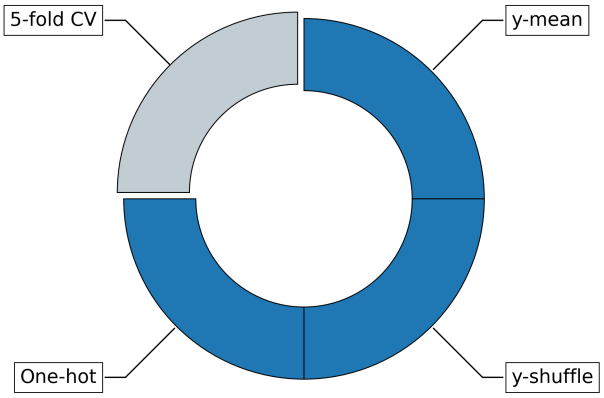
Time VERIFY: 1.12 seconds

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

VERIFY tests of NN_80_No_PFI



VERIFY tests of NN_80_PFI



PREDICT

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

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

o ML model NN_80 (with no PFI filter) and Xy database were loaded, including:

- Target value: Target_values
- Model: NN
- 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: NN_80_train_No_PFI.csv
- Validation set with predicted results: NN_80_valid_No_PFI.csv
- Test set with predicted results: NN_80_test_No_PFI.csv

o Saving graphs and CSV databases in:

- Graph in: PREDICT/Results_NN_80_No_PFI.png

o Results saved in PREDICT/Results_NN_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 = 0.97, MAE = 0.071, RMSE = 0.12
- Validation : R2 = 0.98, MAE = 0.089, RMSE = 0.11
- Test : R2 = 0.99, MAE = 0.056, RMSE = 0.069

o SHAP plot saved in PREDICT/SHAP_NN_80_No_PFI.png
o SHAP values saved in PREDICT/SHAP_NN_80_No_PFI.dat:
- x7 = min: -0.29, max: 0.22
- x6 = min: -0.35, max: 0.21
- Csub-Csub = min: -0.081, max: 0.13
- x9 = min: -0.3, max: 0.13
- x10 = min: -0.25, max: 0.092
- Csub-O = min: -0.24, max: 0.084
- x5 = min: -0.058, max: 0.071
- x8 = min: -0.13, max: 0.065
- x11 = min: -0.11, max: 0.056
- Csub-H = min: -0.17, max: 0.038
- x2 = min: -0.03, max: 0.037
- H-O = min: 0.0, max: 0.0

o PFI plot saved in PREDICT/PFI_NN_80_No_PFI.png
o PFI values saved in PREDICT/PFI_NN_80_No_PFI.dat:
Original score (from model.score, R2) = 0.97
- x7 = 0.26 +- 0.12
- x6 = 0.2 +- 0.16
- x9 = 0.12 +- 0.058
- Csub-Csub = 0.1 +- 0.06
- x10 = 0.094 +- 0.094
- x8 = 0.045 +- 0.049
- x11 = 0.041 +- 0.027
- Csub-H = 0.04 +- 0.033
- x5 = 0.027 +- 0.028
- x2 = -0.0029 +- 0.0083

o Outliers plot saved in PREDICT/Outliers_NN_80_No_PFI.png
o Outlier values saved in PREDICT/Outliers_NN_80_No_PFI.dat:
Train: 2 outliers out of 29 datapoints (6.9%)
- 21 (4.3 SDs)
- 23 (2.4 SDs)

Validation: 0 outliers out of 8 datapoints (0.0%)
Test: 0 outliers out of 9 datapoints (0.0%)

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

o ML model NN_80_PFI (with PFI filter) and Xy database were loaded, including:
- Target value: Target_values
- Model: NN
- Descriptors: ['x5', 'x7', 'x8', 'x9', 'x11', 'Csub-Csub', 'Csub-H']
- 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: NN_80_train_PFI.csv
- Validation set with predicted results: NN_80_valid_PFI.csv
- Test set with predicted results: NN_80_test_PFI.csv

o Saving graphs and CSV databases in:
- Graph in: PREDICT/Results_NN_80_PFI.png

o Results saved in PREDICT/Results_NN_80_PFI.dat:
- Points Train:Validation:Test = 29:8:9
- Proportion Train:Validation:Test = 63:17:20
- Number of descriptors = 7
- Proportion points:descriptors = 37:7
- Train : R2 = 0.93, MAE = 0.11, RMSE = 0.18
- Validation : R2 = 0.97, MAE = 0.081, RMSE = 0.097
- Test : R2 = 0.96, MAE = 0.088, RMSE = 0.13

o SHAP plot saved in PREDICT/SHAP_NN_80_PFI.png
o SHAP values saved in PREDICT/SHAP_NN_80_PFI.dat:
- x7 = min: -0.71, max: 0.32
- x11 = min: -0.086, max: 0.25
- Csub-Csub = min: -0.28, max: 0.19
- Csub-H = min: -0.084, max: 0.12
- x5 = min: -0.064, max: 0.092
- x9 = min: -0.4, max: 0.076
- x8 = min: -0.18, max: 0.034

o PFI plot saved in PREDICT/PFI_NN_80_PFI.png
o PFI values saved in PREDICT/PFI_NN_80_PFI.dat:
Original score (from model.score, R2) = 0.96
- x7 = 1.3 +- 0.51
- Csub-Csub = 0.46 +- 0.26
- x9 = 0.21 +- 0.095
- x11 = 0.18 +- 0.083
- Csub-H = 0.13 +- 0.067
- x5 = 0.074 +- 0.028
- x8 = 0.053 +- 0.043

o Outliers plot saved in PREDICT/Outliers_NN_80_PFI.png
o Outlier values saved in PREDICT/Outliers_NN_80_PFI.dat:
Train: 2 outliers out of 29 datapoints (6.9%)
- 19 (3.2 SDs)
- 21 (2.7 SDs)
Validation: 0 outliers out of 8 datapoints (0.0%)
Test: 0 outliers out of 9 datapoints (0.0%)

Time PREDICT: 6.73 seconds

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

No PFI:

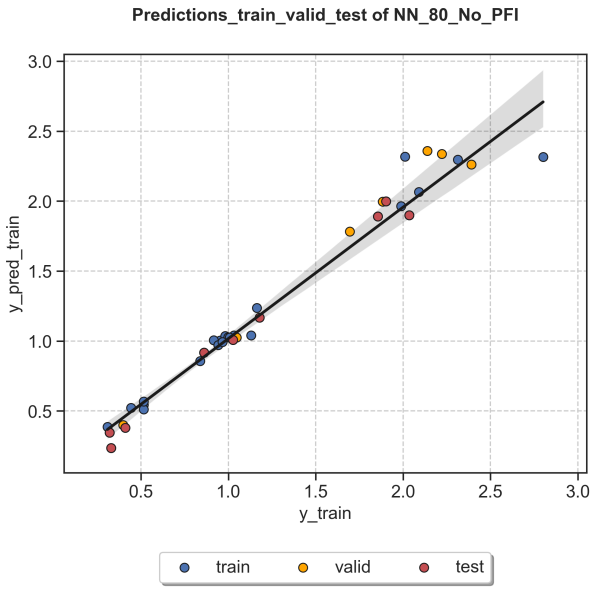
Results_NN_80_No_PFI.dat:

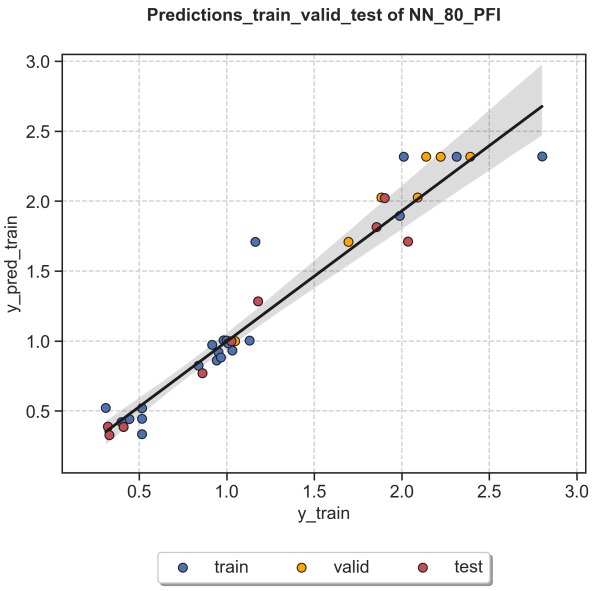
- Points Train:Validation:Test = 29:8:9
- Proportion Train:Validation:Test = 63:17:20
- Number of descriptors = 12
- Proportion points:descriptors = 37:12
- Train : R2 = 0.97, MAE = 0.071, RMSE = 0.12
- Validation : R2 = 0.98, MAE = 0.089, RMSE = 0.11
- Test : R2 = 0.99, MAE = 0.056, RMSE = 0.069

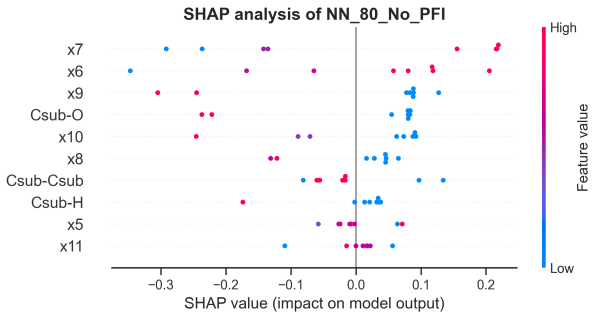
PFI:

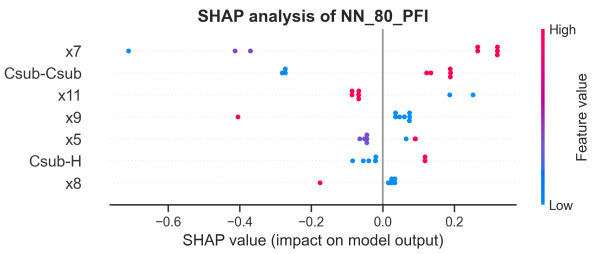
Results_NN_80_PFI.dat:

- Points Train:Validation:Test = 29:8:9
- Proportion Train:Validation:Test = 63:17:20
- Number of descriptors = 7
- Proportion points:descriptors = 37:7
- Train : R2 = 0.93, MAE = 0.11, RMSE = 0.18
- Validation : R2 = 0.97, MAE = 0.081, RMSE = 0.097
- Test : R2 = 0.96, MAE = 0.088, RMSE = 0.13

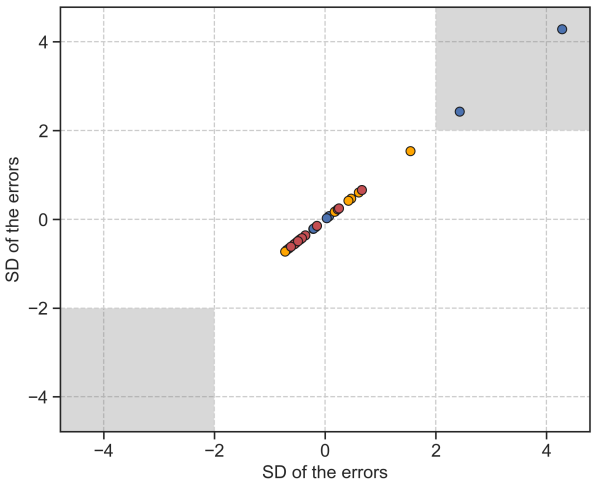




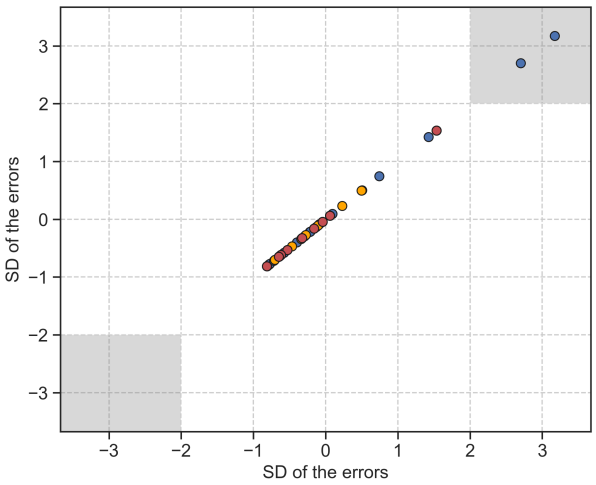




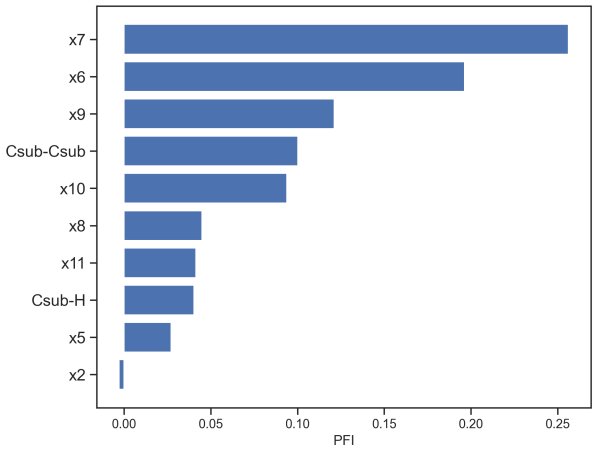
Outlier analysis of NN_80_No_PFI



Outlier analysis of NN_80_PFI



Permutation feature importances (PFIs) of NN_80_No_PFI



Permutation feature importances (PFIs) of NN_80_PFI