Forester report

version 1.1.4

2023-03-12 12:32:31

This report contains details about the best trained model, table with metrics for every trained model, scatter plot for chosen metric and info about used data.

The best models

This is the **regression** task.
The best model is: **xgboost_bayes**.

The names of the models were created by a pattern *Engine_TuningMethod_Id*, where:

- Engine describes the engine used for the training (random_forest, xgboost, decision_tree, lightgbm, catboost),

- TuningMethod describes how the model was tuned (basic for basic parameters, RS for random search, bayes for Bayesian optimization),

- Id for separating the random search parameters sets.

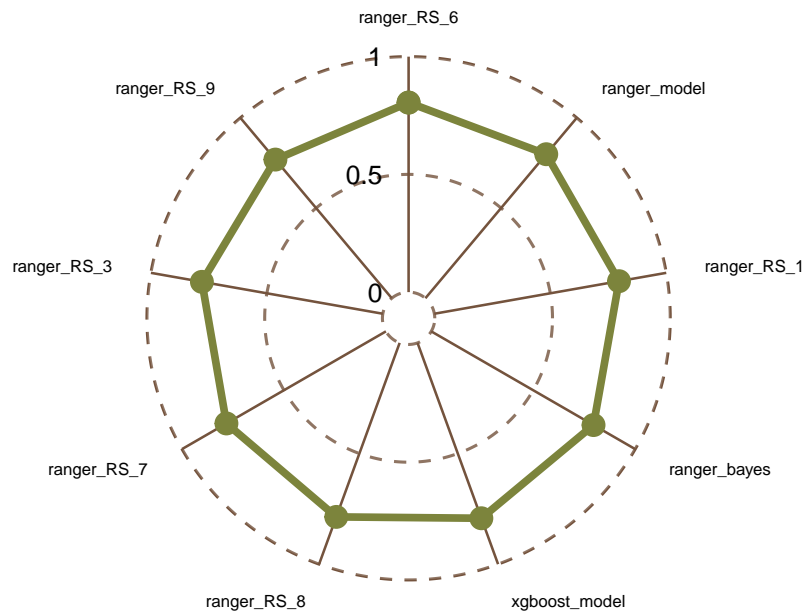*More details about the best model are present at the end of the report.*

no.	name	mse	r2	mae
46	xgboost_bayes	20019465477	0.8061	91233.10
10	ranger_RS_6	20136099053	0.8050	93352.71
1	ranger_model	20806262125	0.7985	94642.38
5	ranger_RS_1	21034721439	0.7963	91339.61
45	ranger_bayes	21088257110	0.7958	97276.05
2	xgboost_model	21417302625	0.7926	95079.18
12	ranger_RS_8	22047436990	0.7865	100513.53
11	ranger_RS_7	22499936628	0.7821	101934.84
7	ranger_RS_3	22705321932	0.7801	106266.35
13	ranger_RS_9	23875911122	0.7688	108421.42
36	lightgbm_RS_2	29894499479	0.7105	122674.08
39	lightgbm_RS_5	29894499479	0.7105	122674.08
48	lightgbm_bayes	31074970168	0.6990	125714.65
40	lightgbm_RS_6	31579607685	0.6942	125292.62
43	lightgbm_RS_9	31579607685	0.6942	125292.62
4	lightgbm_model	32891173014	0.6815	125920.39
44	lightgbm_RS_10	32891173014	0.6815	125920.39
41	lightgbm_RS_7	33164795309	0.6788	133063.34
9	ranger_RS_5	34126056622	0.6695	144816.67
47	decision_tree_bayes	34316091895	0.6677	119223.85
38	lightgbm_RS_4	36239460976	0.6490	130688.36

no.	name	mse	r2	mae
6	ranger_RS_2	38146990720	0.6306	158780.45
42	lightgbm_RS_8	39835245775	0.6142	151987.37
14	ranger_RS_10	40073333196	0.6119	161930.48
35	lightgbm_RS_1	41400427683	0.5990	136431.52
8	ranger_RS_4	44127682866	0.5726	168106.95
21	xgboost_RS_7	44347693165	0.5705	153698.36
24	xgboost_RS_10	44347693165	0.5705	153698.36
3	decision_tree_model	51144007059	0.5047	139324.28
25	decision_tree_RS_1	51144007059	0.5047	139324.28
26	decision_tree_RS_2	51144007059	0.5047	139324.28
27	decision_tree_RS_3	51144007059	0.5047	139324.28
28	decision_tree_RS_4	51144007059	0.5047	139324.28
29	decision_tree_RS_5	51144007059	0.5047	139324.28
30	decision_tree_RS_6	51144007059	0.5047	139324.28
31	decision_tree_RS_7	51144007059	0.5047	139324.28
32	decision_tree_RS_8	51144007059	0.5047	139324.28
33	decision_tree_RS_9	51144007059	0.5047	139324.28
34	decision_tree_RS_10	51144007059	0.5047	139324.28
37	lightgbm_RS_3	55292995222	0.4645	151419.03
19	xgboost_RS_5	113027077705	-0.0946	260209.46
23	xgboost_RS_9	113027077705	-0.0946	260209.46
20	xgboost_RS_6	114481520396	-0.1087	268286.18
15	xgboost_RS_1	115523592767	-0.1188	271622.83
17	xgboost_RS_3	115523592767	-0.1188	271622.83
18	xgboost_RS_4	257551703920	-1.4943	427069.95
22	xgboost_RS_8	257551703920	-1.4943	427069.95
16	xgboost_RS_2	258582072126	-1.5043	424857.94

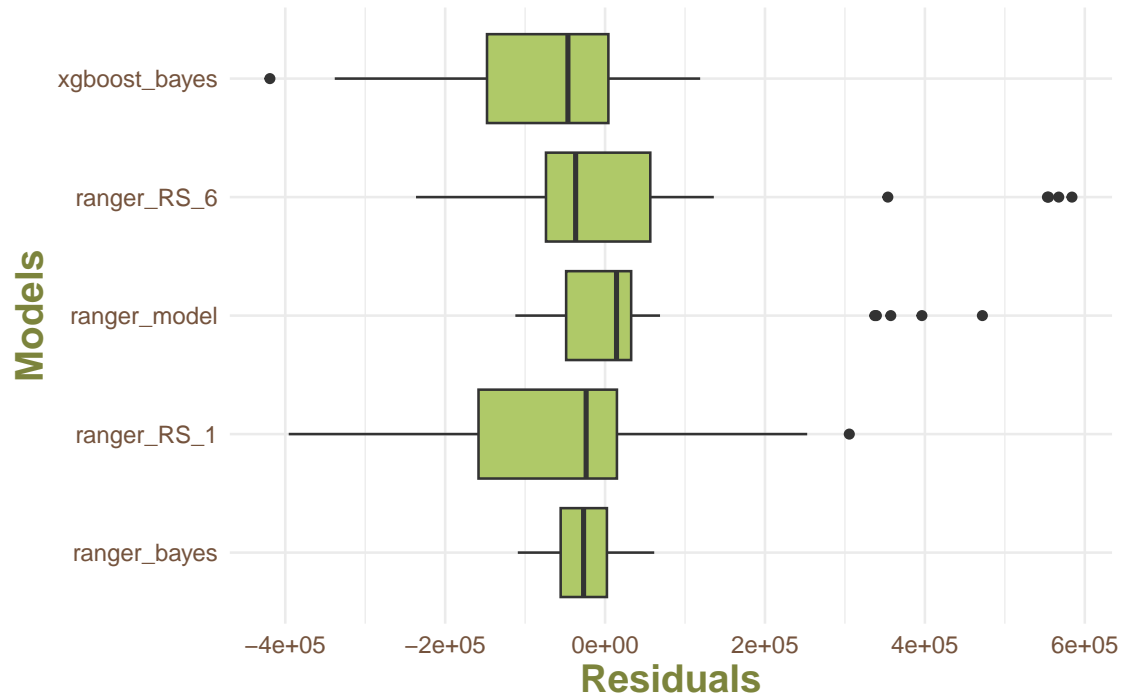## Plots for all models

```
[1] "no."     "name"    "engine" "tuning" "mse"    "r2"     "mae"
   no.           name engine          tuning         mse        r2        mae
1   46 xgboost_bayes xgboost     bayes_opt 20019465477 0.8061156   91233.10
2   10    ranger_RS_6  ranger random_search 20136099053 0.8049861   93352.71
3    1   ranger_model  ranger          basic 20806262125 0.7984957   94642.38
4    5    ranger_RS_1  ranger random_search 21034721439 0.7962831   91339.61
5   45   ranger_bayes  ranger     bayes_opt 21088257110 0.7957646   97276.05
6    2 xgboost_model xgboost          basic 21417302625 0.7925779   95079.18
7   12    ranger_RS_8  ranger random_search 22047436990 0.7864752  100513.53
8   11    ranger_RS_7  ranger random_search 22499936628 0.7820928  101934.84
9    7    ranger_RS_3  ranger random_search 22705321932 0.7801037  106266.35
10  13    ranger_RS_9  ranger random_search 23875911122 0.7687668  108421.42
  xgboost_bayes ranger_RS_6 ranger_model ranger_RS_1 ranger_bayes xgboost_model
1     0.8061156   0.8049861    0.7984957   0.7962831    0.7957646     0.7925779
  ranger_RS_8 ranger_RS_7 ranger_RS_3 ranger_RS_9
1   0.7864752   0.7820928   0.7801037   0.7687668
```
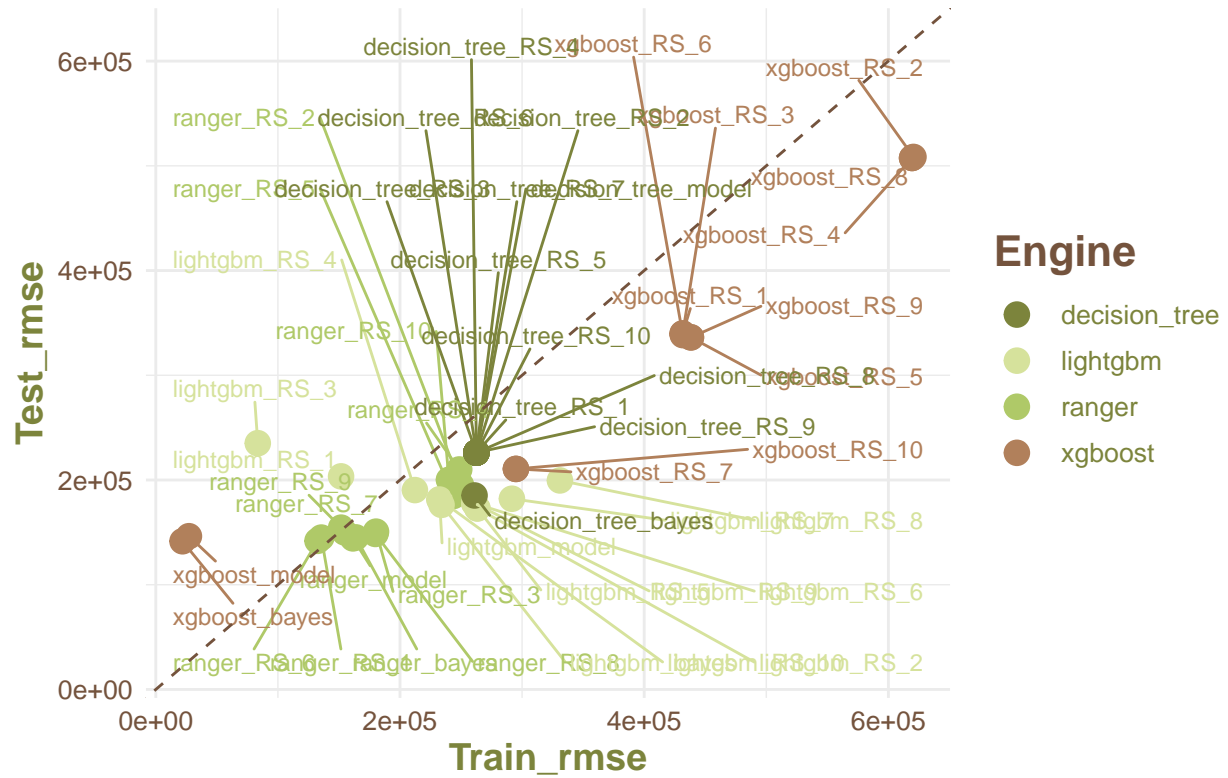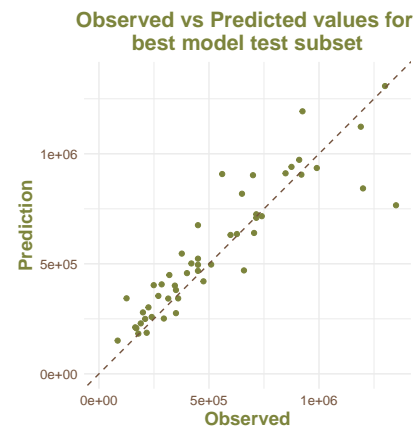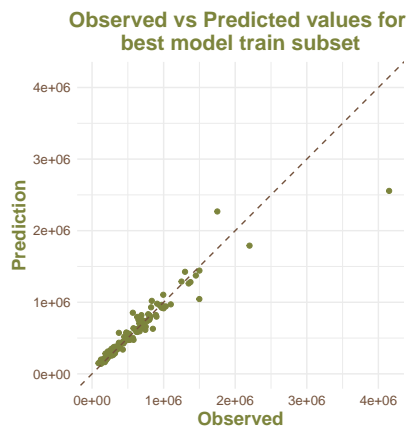
R2 comparison
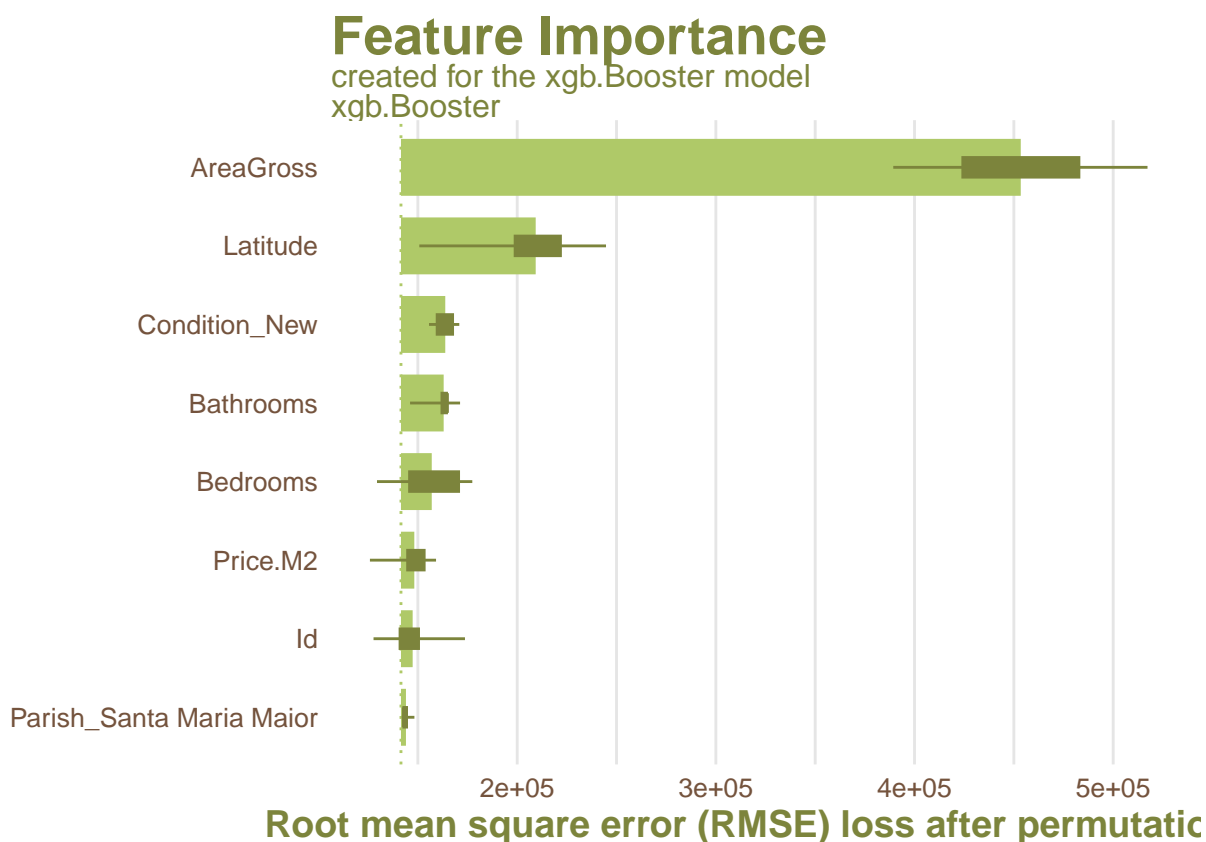

Combined Models Residuals Plot

RMSE Train vs Test plot

Plots for the best model - xgboost__bayes

Observed vs Predicted values for best model train subset

Observed vs Predicted values for best model test subset

Feature Importance for the best model - xgboost_bayes

Feature Importance
created for the xgb.Booster model
xgb.Booster

AreaGross	
Latitude	
Condition_New	
Bathrooms	
Bedrooms	
Price.M2	
Id	
Parish_Santa Maria Maior	

2e+05   3e+05   4e+05   5e+05

Root mean square error (RMSE) loss after permutatic

Details about data

———————— **CHECK DATA REPORT** ————————

The dataset has 246 observations and 17 columns which names are:

Id; Condition; PropertyType; PropertySubType; Bedrooms; Bathrooms; AreaNet; AreaGross; Parking; Latitude; Longitude; Country; District; Municipality; Parish; Price.M2; Price;

With the target value described by a column: Price.

** Static columns are: **Country; District; Municipality;

With dominating values: Portugal; Lisboa; Lisboa;

These column pairs are duplicate: District - Municipality;

No target values are missing.

No predictor values are missing.

No issues with dimensionality.

Strongly correlated, by Spearman rank, pairs of numerical values are:

Bedrooms - AreaNet: 0.77; Bedrooms - AreaGross: 0.77; Bathrooms - AreaNet: 0.78; Bathrooms - AreaGross: 0.78; AreaNet - AreaGross: 1;

** Strongly correlated, by Crammer's V rank, pairs of categorical values are: **

PropertyType - PropertySubType: 1;

These obserwation migth be outliers due to their numerical columns values:

145 146 196 44 5 51 57 58 59 60 61 62 63 64 69 75 76 77 78 ;

Target data is not evenly distributed with quantile bins: 0.25 0.35 0.14 0.26

Columns names suggest that some of them are IDs, removing them can improve the model. Suspicious columns are:

Id

Columns data suggest that some of them are IDs, removing them can improve the model. Suspicious columns are:

Id

———————— **CHECK DATA REPORT END** ————————

The best model details

```
------------ Xgboost model ------------

Parameters
  niter: 15
  evaluation_log:
    iter : train_rmse
      1 :   476244.897073358
      2 :   327432.980112583
      3 :   236386.934942405
      4 :   176361.318182452
      5 :   135291.939991913
      6 :   104979.966191163
      7 :   82787.6845203692
      8 :   66090.1209296145
      9 :   53088.237395824
     10 :   44430.7065202948
     11 :   37213.8807278742
     12 :   32829.8814129894
     13 :   27986.07969806
     14 :   25245.5741694385
     15 :   21583.0437507795
```