Improvement of regression model

Author: Yelu
Date: 2024/9/17

Feature selection

Dependent variables
- Presence of accident
- Severity of accident (Not finished)
  - Presence of person injury
  - Presence of property damage
  - Value of person injury
    - Severely injured person
    - Lightly injured person
  - Value of property damage

Feature selection methods:
- Sequential feature selection
  - SFFS: sequential forward floating selection
  - SBFS: sequential backward floating selection

Parameters setting for sequential feature selection:
- Forward/Backward
- Estimator/Model
  - Linear regression
  - Logistic regression model
- Metrics used in scoring performance in feature selection
  - For linear regression:
    - Neg_mean_squared_error
    - R2 score
    - Neg_median_absolute_error
    - Neg_mean_absolute_error
  - For logistic regression (Not finished)
    - Accuracy
    - F1
    - Recall
    - Roc_auc
- Cv
  - K value for stratified k-fold cross-validation: 5/10/15/20/25

Comparison
- Between forward and backward
- Between linear regression and logistic regression
- Between different scoring metrics
- Between different k values for cross validation

Regression model with all features

For comparison, regression result with all features without any feature selection: (split is with 0.3 test size)

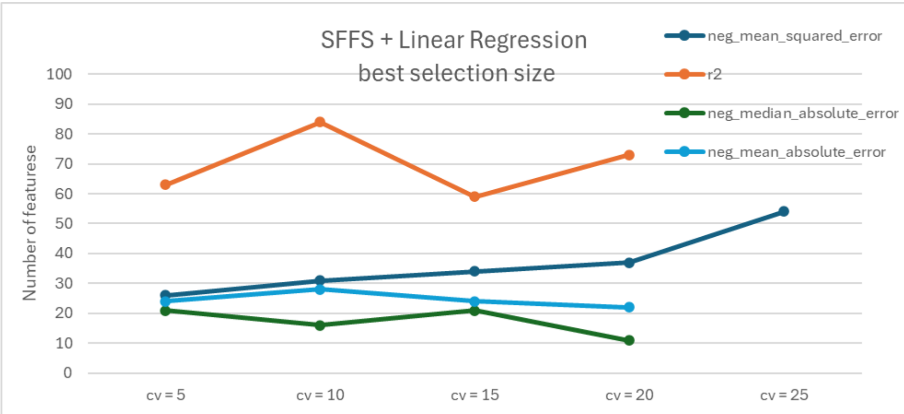
<table>
<tr><td rowspan="6">Original</td><td rowspan="3">OLS</td><td></td><td>R squared</td><td>Adj R</td><td>accuracy</td><td>precision</td><td>recall</td><td>f1 score</td></tr>
<tr><td>all</td><td>0.672</td><td>0.641</td><td>0.937</td><td>0.930</td><td>0.906</td><td>0.917</td></tr>
<tr><td>split</td><td>0.716</td><td>0.676</td><td>0.878</td><td>0.864</td><td>0.836</td><td>0.848</td></tr>
<tr><td rowspan="3">Logit</td><td></td><td colspan="2">Pseudo R squared</td><td>accuracy</td><td>precision</td><td>recall</td><td>f1 score</td></tr>
<tr><td>all</td><td colspan="2">0.780</td><td>0.949</td><td>0.936</td><td>0.935</td><td>0.936</td></tr>
<tr><td>split</td><td colspan="2">0.837</td><td>0.960</td><td>0.954</td><td>0.932</td><td>0.942</td></tr>
</table>

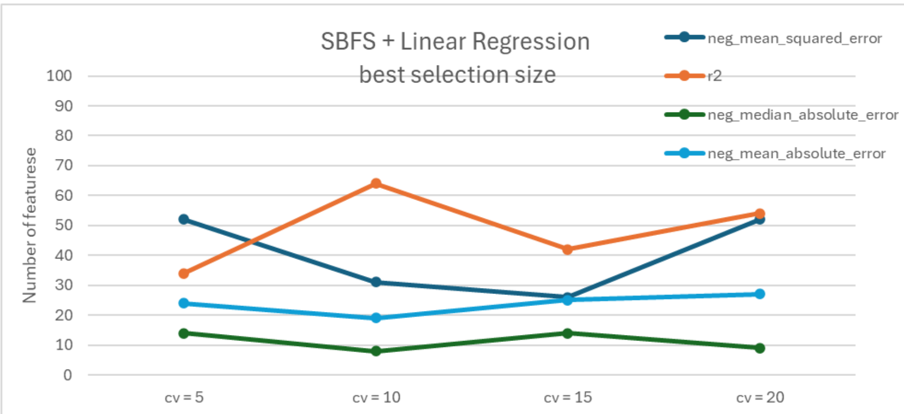
Find the best size of feature selection

Linear regression model

With four scoring metrics and four to five cv values setting, there are 17 selection results of forward and 16 results of backward with linear regression model in total.

Best size of feature selection (of both forward and backward)





Average size of feature selection with linear regression

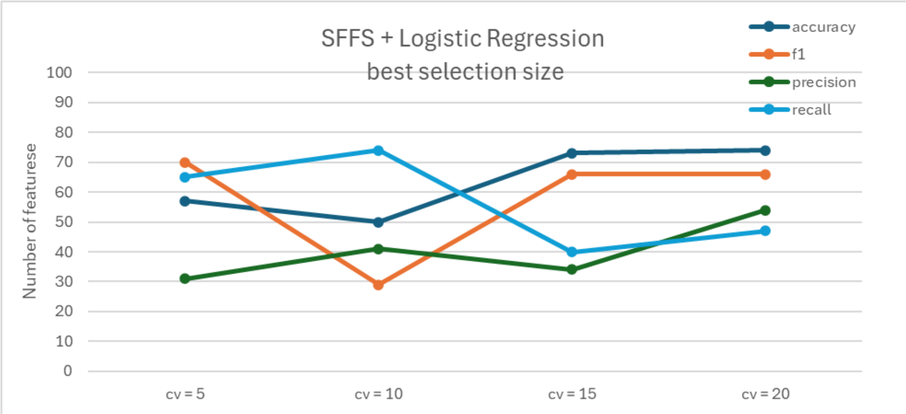
average selection size		
sffs	sbfs	average
37	31	34

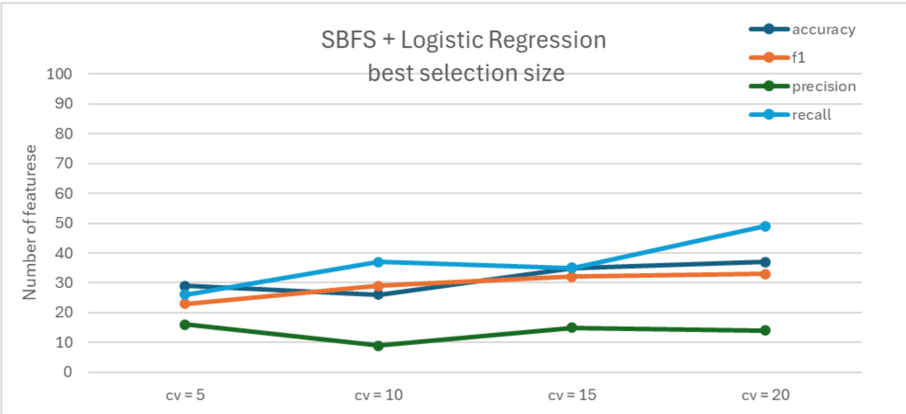
Best size of feature selection is around 34 for linear regression.

Logistic regression model

With four scoring metrics and four cv values setting, there are 16 selection results of forward and 16 results of backward with logistic regression model in total.

Best size of feature selection (of both forward and backward)





Average size of feature selection with linear regression

average selection size		
sffs	sbfs	average
54	28	41

Best size of feature selection is around 41 for logistic regression.

Comparison between feature selection using linear regression and logistic regression

Overall, best size of feature selection is around 37-38. And SFFS always has a larger 'best' number of features than SBFS. As for SFFS, feature selection with OLS has a smaller 'best' number than that with Logit, while feature selection with Logit has a smaller 'best' number than that with OLS for SBFS.

average selection size			
	sffs	sbfs	average
OLS	37	31	34
Logit	54	28	41

Summarize the selected features with best size of selection

Linear regression model

The following table shows features ordered by number of being selected. Curb-related variables including **cp, cmax, cmin, cmean** are always among the selection with the top number of being selected. Other numeric values such as **r_width_value, speedlimt_value, dtrafficarea** are also mostly frequently selected.

Summary table of feature selection results – linear regression

	SFFS		SBFS		SFS Total	
No.	Feature	Number	Feature	Number	Feature	Number
1	speedlimit_6	16	speedlimit_2	15	speedlimit_2	30
2	trafficarea_2	15	trafficarea_2	13	trafficarea_2	28
3	speedlimit_2	15	r_width_7	12	speedlimit_6	28
4	r_surface_2	14	speedlimit_6	12	**r_width_value**	23
5	r_width_6	13	**r_width_value**	11	r_width_2	23
6	speedlimit_4	13	**speedlimit_value**	11	r_width_6	23
7	speedlimit_5	13	r_width_2	11	**speedlimit_value**	22
8	**r_width_value**	12	speedlimit_1	11	r_width_7	22
9	r_width_2	12	speedlimit_3	11	speedlimit_1	22
10	r_surface_1	12	**cp**	10	speedlimit_4	22
11	**speedlimit_value**	11	**cmax**	10	speedlimit_5	22
12	curbtype_2	11	**dtrafficarea**	10	r_surface_1	21
13	z_qnr_30	11	r_width_5	10	speedlimit_3	21
14	z_qnr_31	11	r_width_6	10	**cp**	20
15	r_width_4	11	r_surface_1	9	**dtrafficarea**	20
16	speedlimit_1	11	speedlimit_4	9	**cmax**	19
17	**cp**	10	speedlimit_5	9	curbtype_2	19
18	**dtrafficarea**	10	curbtype_2	8	r_width_4	19
19	trafficarea_3	10	z_qnr_26	8	r_width_5	19
20	r_width_7	10	trafficarea_1	8	z_qnr_30	18
21	speedlimit_3	10	trafficarea_3	8	trafficarea_3	18
22	**cmax**	9	r_width_3	8	r_surface_2	18
23	curbtype_1	9	r_width_4	8	z_qnr_26	17
24	z_qnr_3	9	**cmean**	7	z_qnr_31	17
25	z_qnr_23	9	curbtype_1	7	curbtype_1	16
26	z_qnr_26	9	z_qnr_30	7	trafficarea_1	16
27	r_width_5	9	z_knr_9	7	r_width_3	16
28	**cmin**	8	**cmin**	6	z_qnr_3	15
29	z_qnr_7	8	z_qnr_3	6	**cmin**	14
30	z_qnr_25	8	z_qnr_7	6	z_qnr_7	14
31	trafficarea_1	8	z_qnr_10	6	z_qnr_23	14
32	r_width_1	8	z_qnr_16	6	z_knr_9	14
33	r_width_3	8	z_qnr_31	6	**cmean**	13
34	z_qnr_8	7	**dtraml**	5	z_qnr_25	13
35	z_knr_9	7	**drainl**	5	r_width_1	13
36	**iemin**	6	**gvm_dwv**	5	z_qnr_8	12
37	**cmean**	6	z_qnr_8	5	z_qnr_10	11

Logistic regression model

The following table shows features ordered by number of being selected. Curb-related variables including **cp, cmin, cmean** are always among the selection with the top number of being selected. Other numeric values such as **ie, mew, meg, r_width_value, speedlimt_value, dbusl, dtraml, droad, bicyclecount** are frequently selected.

Summary table of feature selection results – logistic regression

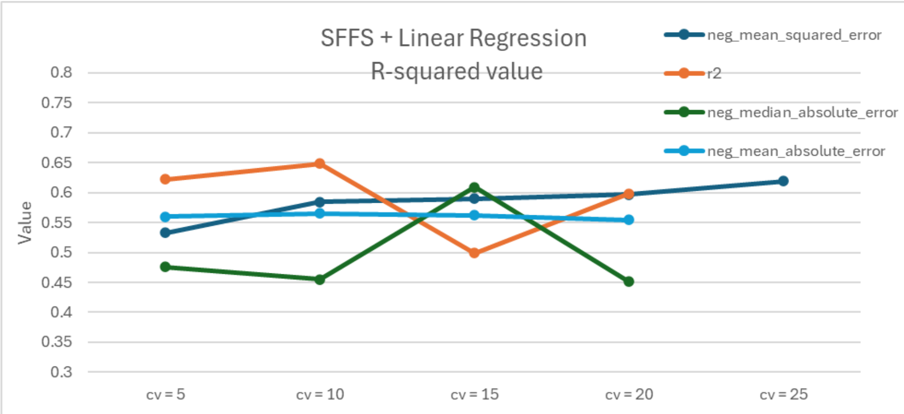
	SFFS		SBFS		SFS Total	
No.	Feature	Number	Feature	Number	Feature	Number
1	curbtype_1	16	trafficarea_2	16	speedlimit_4	32
2	z_qnr_4	16	speedlimit_4	16	curbtype_1	30
3	z_qnr_16	16	curbtype_1	14	trafficarea_2	28
4	speedlimit_4	16	**cp**	13	**cp**	26
5	**iemin**	15	speedlimit_2	13	r_surface_1	26
6	**iemax**	15	speedlimit_6	13	z_qnr_16	25
7	curbtype_2	15	**speedlimit_value**	12	speedlimit_2	25
8	z_qnr_23	15	z_knr_4	12	z_qnr_23	24
9	z_qnr_30	15	z_knr_9	12	r_width_2	24
10	r_width_2	15	r_surface_1	12	**iemin**	23
11	z_qnr_2	14	**dtraml**	11	**speedlimit_value**	23
12	z_qnr_6	14	z_qnr_26	11	z_qnr_26	23
13	z_qnr_25	14	z_knr_5	11	speedlimit_6	23
14	z_qnr_31	14	**r_width_value**	10	**cmean**	22
15	r_surface_1	14	**cmean**	9	curbtype_2	22
16	speedlimit_5	14	z_qnr_10	9	z_knr_9	22
17	**cp**	13	z_qnr_16	9	**ie**	20
18	**ie**	13	z_qnr_23	9	z_qnr_13	19
19	**mew**	13	r_width_2	9	z_qnr_25	19
20	**meg**	13	r_width_7	9	z_knr_4	19
21	**cmean**	13	**iemin**	8	**mew**	18
22	z_qnr_14	13	**dbusl**	8	**mewmax**	18
23	z_qnr_29	13	**droad**	8	z_qnr_10	18
24	**mewmin**	12	z_qnr_8	8	z_qnr_29	18
25	**mewmax**	12	r_width_5	8	r_width_7	18
26	**cmin**	12	**ie**	7	**cmin**	17
27	z_qnr_3	12	**dtrainl**	7	z_qnr_4	17
28	z_qnr_13	12	curbtype_2	7	z_qnr_7	17
29	z_qnr_26	12	z_qnr_13	7	z_qnr_33	17
30	trafficarea_2	12	z_qnr_33	7	**iemax**	16
31	speedlimit_2	12	r_width_1	7	**dbusl**	16
32	**megmin**	11	**mewmax**	6	**dtraml**	16
33	**megmax**	11	**gvm_dwv**	6	**droad**	16
34	**dcurb**	11	z_qnr_7	6	z_qnr_30	16
35	**dvfpath**	11	z_knr_8	6	z_qnr_31	16
36	**speedlimit_value**	11	r_width_3	6	z_knr_5	16
37	z_qnr_7	11	**mew**	5	z_knr_8	16
38	r_surface_2	11	**cmin**	5	r_width_1	16
39	z_qnr_9	10	**cmax**	5	**meg**	15
40	z_qnr_12	10	**bicyclecount**	5	**r_width_value**	15
41	z_qnr_33	10	z_qnr_25	5	z_qnr_2	15
42	z_knr_2	10	z_qnr_29	5	z_qnr_6	15
43	z_knr_8	10	z_knr_11	5	z_qnr_8	15
44	z_knr_9	10	r_width_4	5	r_width_5	15
45	speedlimit_6	10	**dparktw**	4	**bicyclecount**	14
46	**gvm_msp**	9	z_qnr_28	4	z_knr_11	14
47	**bicyclecount**	9	**dstopsign**	3	speedlimit_5	14
48	z_qnr_5	9	z_qnr_5	3	**mewmin**	13
49	z_qnr_10	9	z_qnr_11	3	**megmin**	13
50	z_qnr_28	9	z_knr_1	3	**dcurb**	13
51	z_qnr_32	9	z_knr_10	3	**dvfpath**	13
52	z_knr_11	9	**meg**	2	z_qnr_3	13
53	r_width_1	9	**megmin**	2	z_qnr_14	13
54	r_width_6	9	**dcurb**	2	z_qnr_28	13

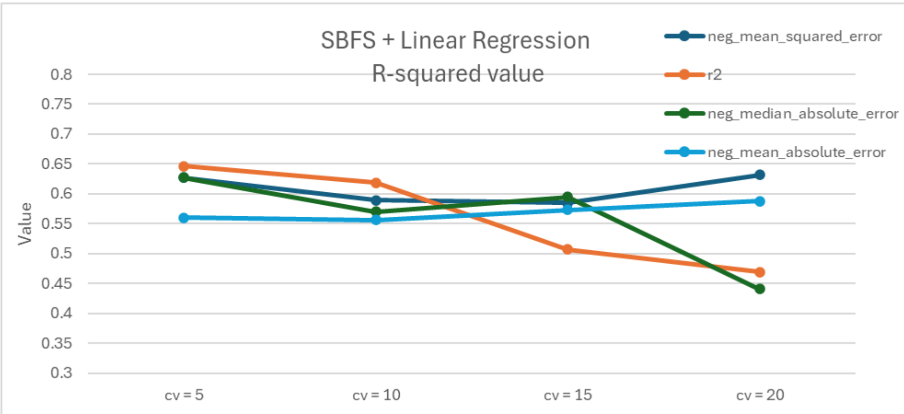
Apply selected features to regression model

Linear regression model

Compare R-squared value and Adj R-squared value of regression model with 17 forward selection results as well as 16 backward selection results.

R-squared value of regression model using selected features (of both forward and backward)



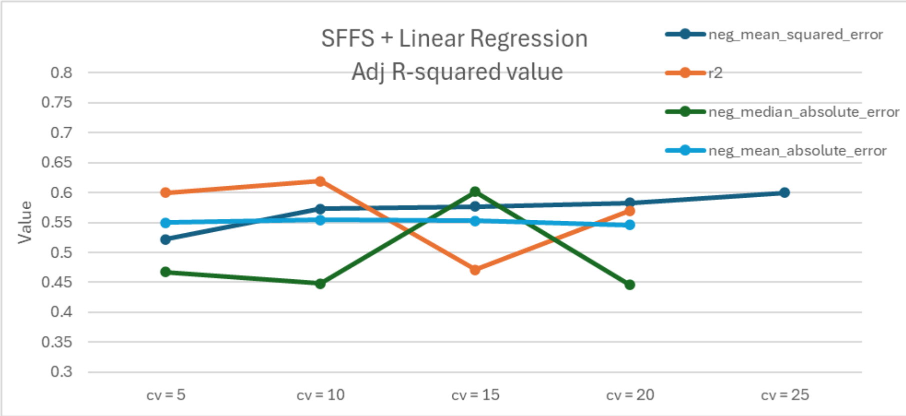


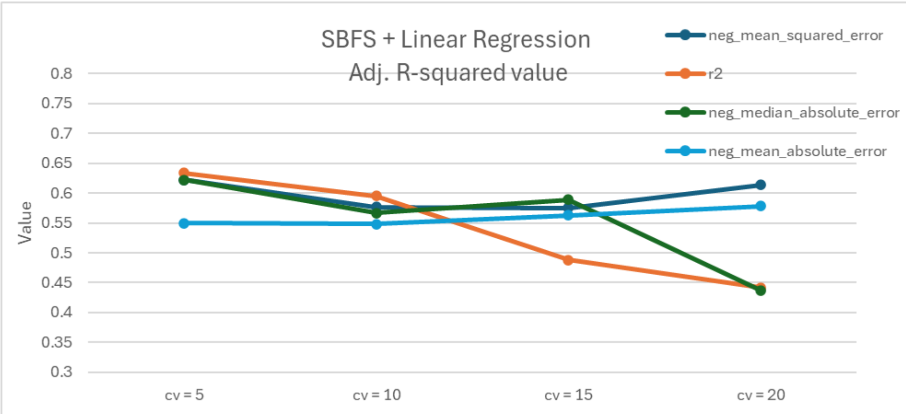
Average R-squared value of OLS regression model with feature selection

average r squared	
sffs	sbfs
0.559	0.574

R-squared value of regression model using only selected features from forward selection is around 0.559, using those from backward selection is around 0.574.

Adj R-squared value of regression model using selected features (of both forward and backward)



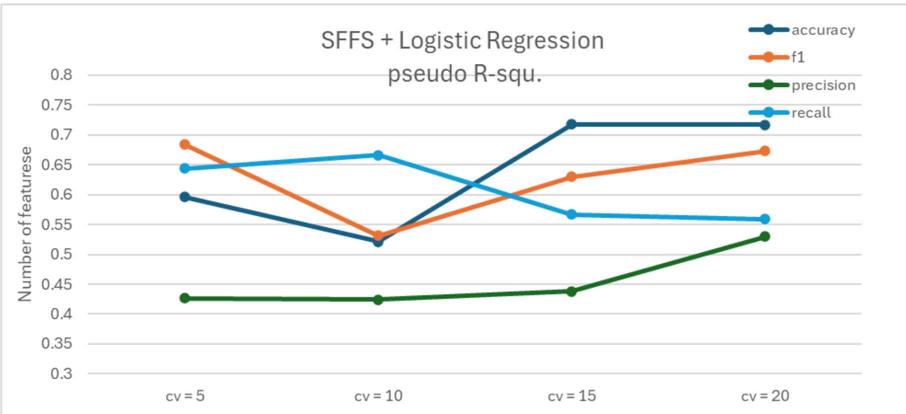


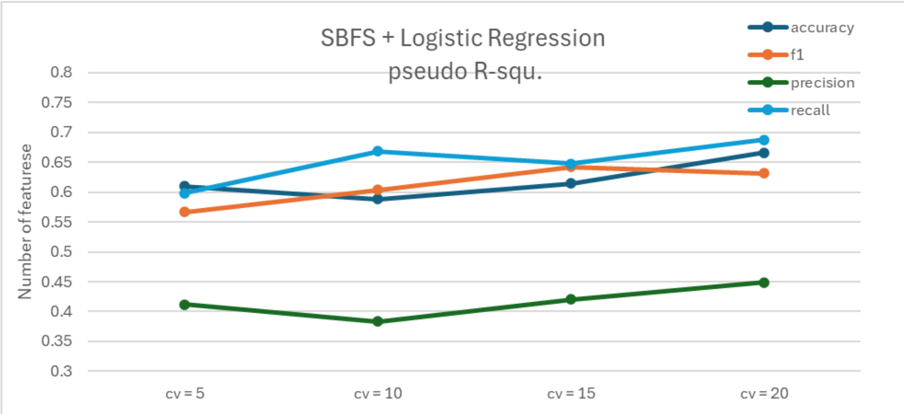
Average Adj. R-squared value of OLS regression model with feature selection

average adj r squared	
sffs	sbfs
0.544	0.563

Adj. R-squared value of regression model using only selected features from forward selection is around 0.544, using those from backward selection is around 0.563.

Logistic regression model





average pseudo R-squ.	
sffs	sbfs
0.583	0.574

Pseudo R-squared value of regression model using only selected features from forward selection is around 0.583, using those from backward selection is around 0.574.

Comparison between feature selection using linear regression and logistic regression

Overall, R-squared value of Logit is slightly larger than that of OLS after feature selection.

Correlation analysis

Pairwise correlation is analysed for all features, of which the visualization result is in appendix.

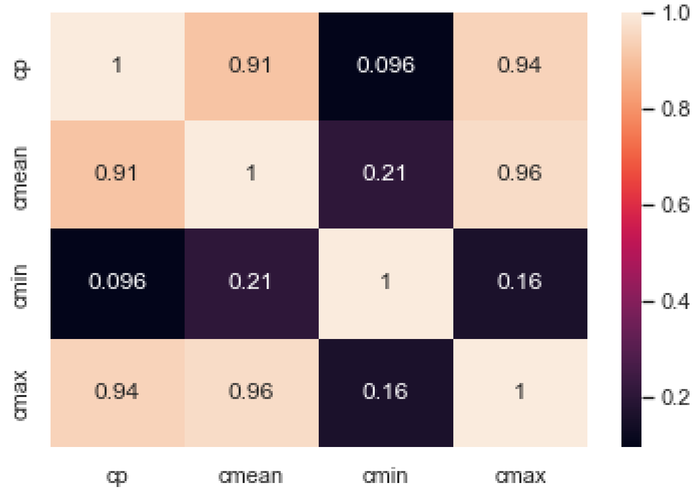
Summary of sorted correlation values of variable pairs is shown in the following table, only pairs with a correlation value higher than 0.5 or lower than -0.5 are included here. Among curb-related variables (**cp, cmean, cmax**) are correlated with each other.

Pairs of variables		Correlation
curbtype_2	curbtype_1	-1.000
r_surface_2	r_surface_1	-1.000
speedlimit_2	**speedlimit_value**	-0.844
trafficarea_2	trafficarea_1	-0.778
speedlimit_1	**dtrafficarea**	-0.621
trafficarea_3	trafficarea_1	-0.581
speedlimit_3	**speedlimit_value**	0.508
z_knr_9	z_qnr_19	0.547
z_knr_10	z_qnr_33	0.549
mewmax	**mewmin**	0.560
z_knr_11	z_qnr_24	0.567
z_knr_11	z_qnr_25	0.575
z_knr_1	z_qnr_1	0.575
droad	**dcurb**	0.580
r_width_3	**r_width_value**	0.596
z_knr_3	z_qnr_15	0.620
z_knr_12	z_qnr_31	0.622
z_knr_5	z_qnr_10	0.641
z_knr_12	z_qnr_32	0.646
z_knr_2	z_qnr_2	0.691
z_knr_8	z_qnr_18	0.730
megmax	**meg**	0.749
iemax	**iemin**	0.788
mewmax	**mew**	0.844
megmin	**meg**	0.859
mewmin	**mew**	0.862
cmean	**cp**	0.906
iemax	**ie**	0.920
cmax	**cp**	0.940
iemin	**ie**	0.957
cmax	**cmean**	0.964
gvm_msp	**gvm_dwv**	0.982
gvm_asp	**gvm_msp**	0.982
gvm_asp	**gvm_dwv**	0.986

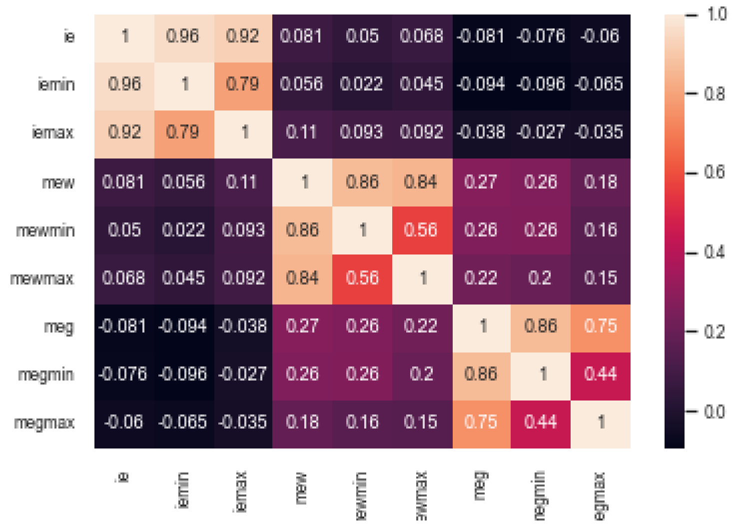
For visualization of correlation matrix, features could be divided into several groups:

1.curb-related features
2.entropy features
3.traffic-transport numeric features
4.traffic-transport categorical features – urban zone and city districts
5.traffic-transport categorical features – others

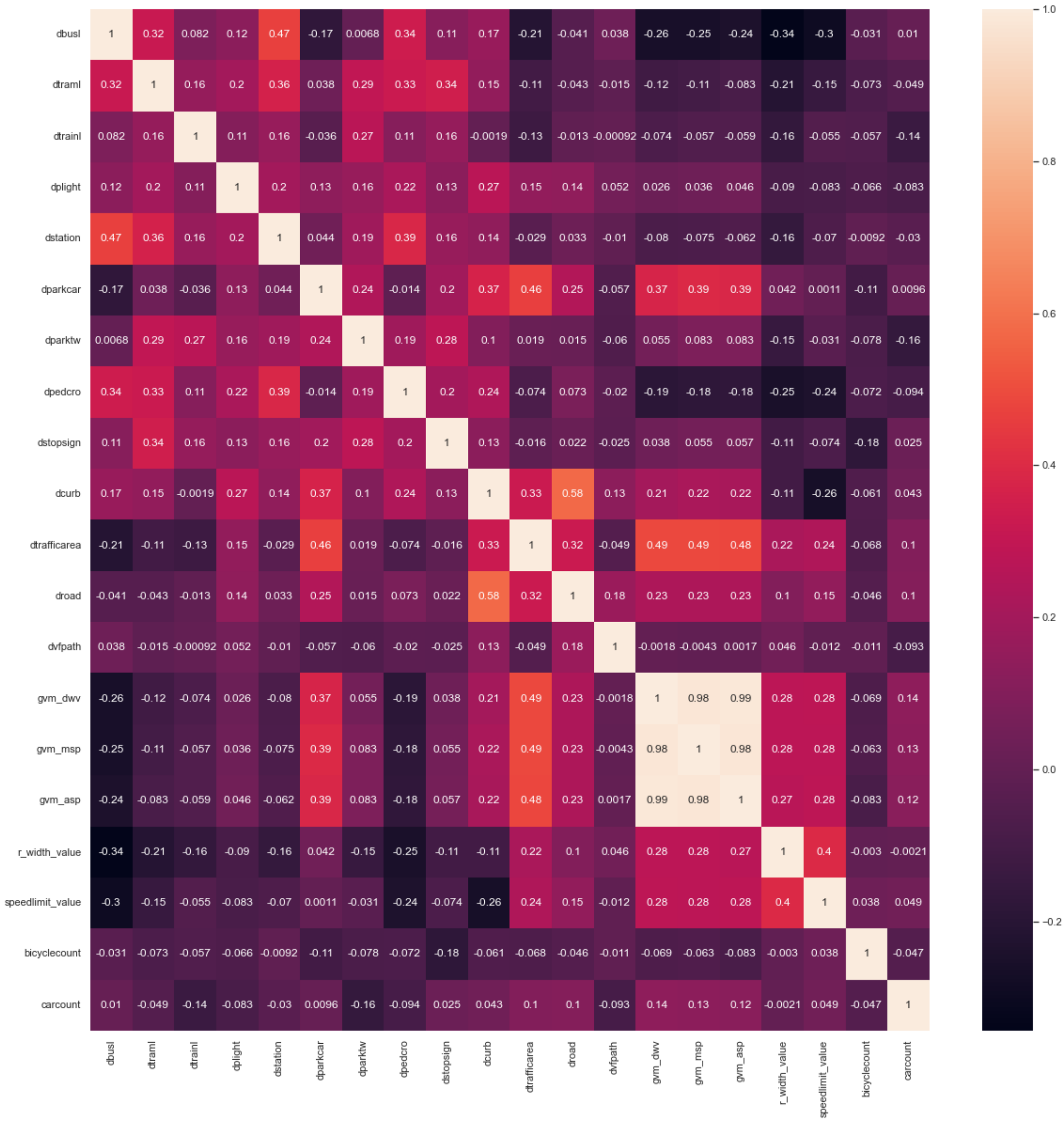
Curb-related variables



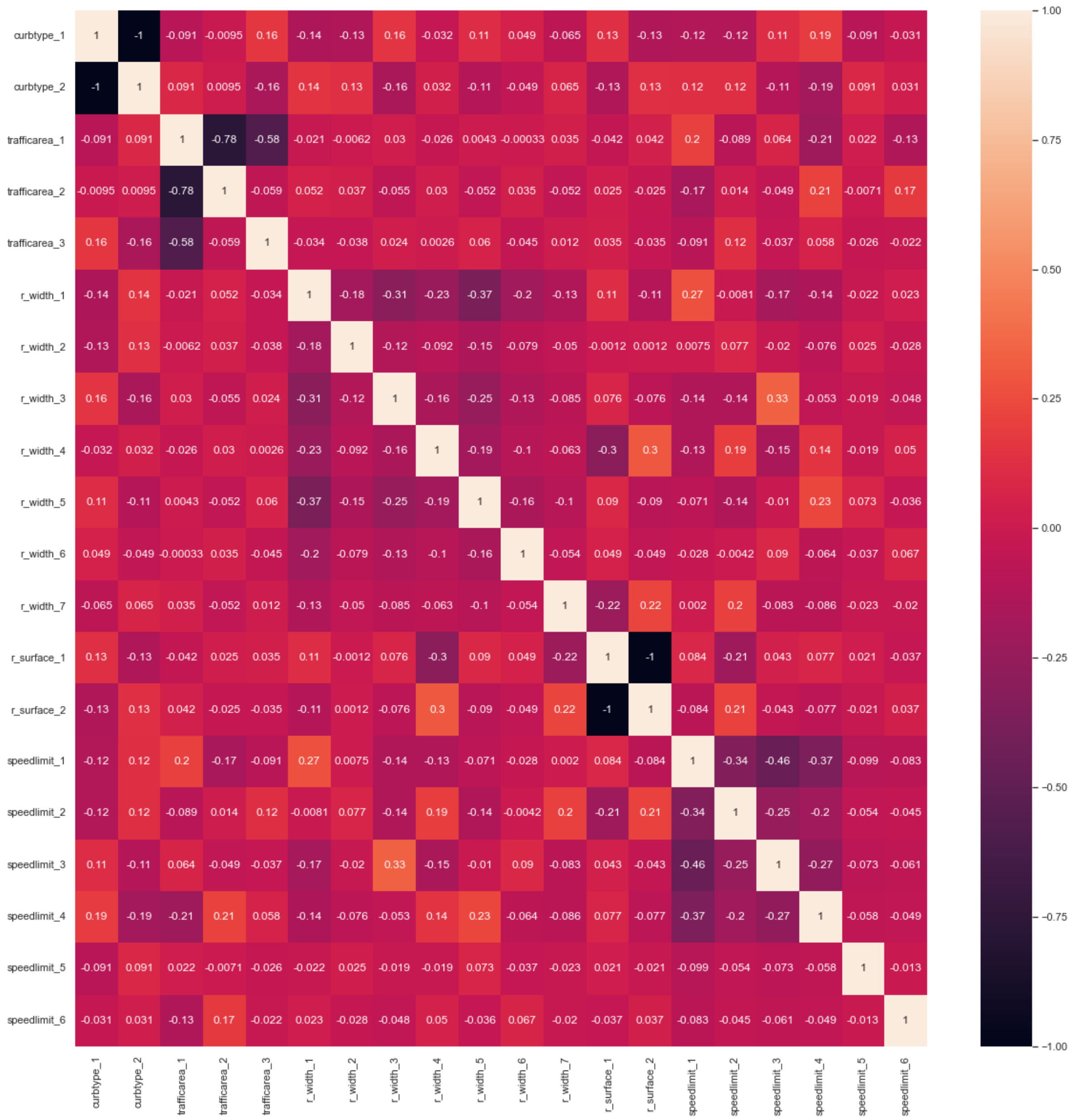
Entropy variables



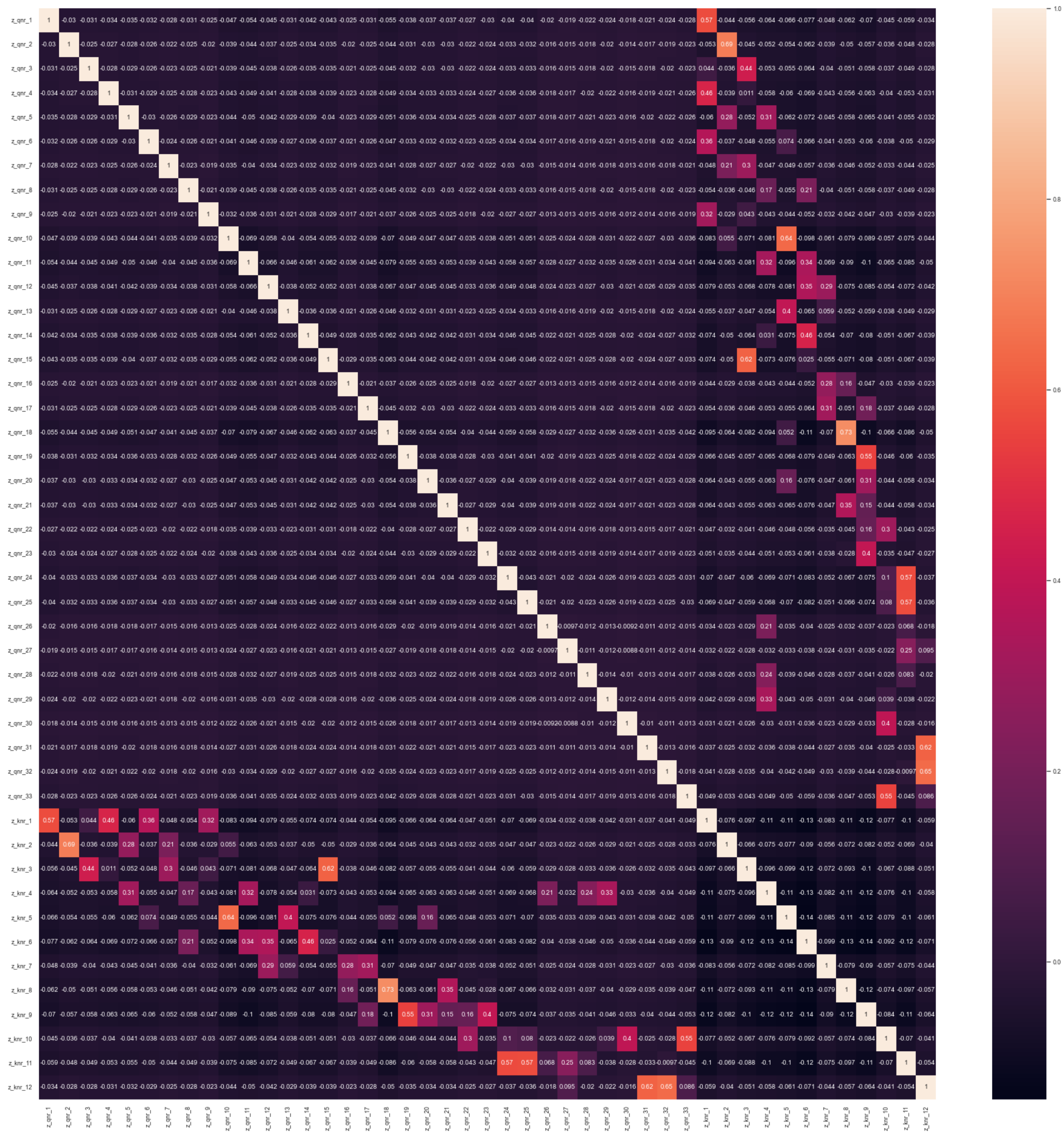
Traffic-transport numeric variables



Traffic-transport categorical variables - others



Traffic-transport categorical variables – Urban districts & statistical zones



Issues:

Sequential feature selection with floating and cross validation is time consuming, especially in local environment without parallel jobs.

Errors of setting scoring metric were found for feature selection with logistic regression and therefore that part needs to be corrected and not included in this report.

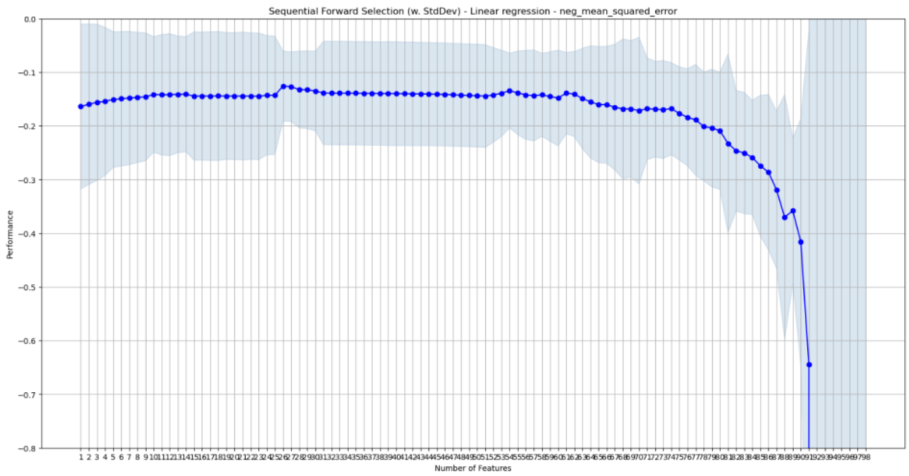
Working on correct feature selection with logistic regression has been paused since 09.04.

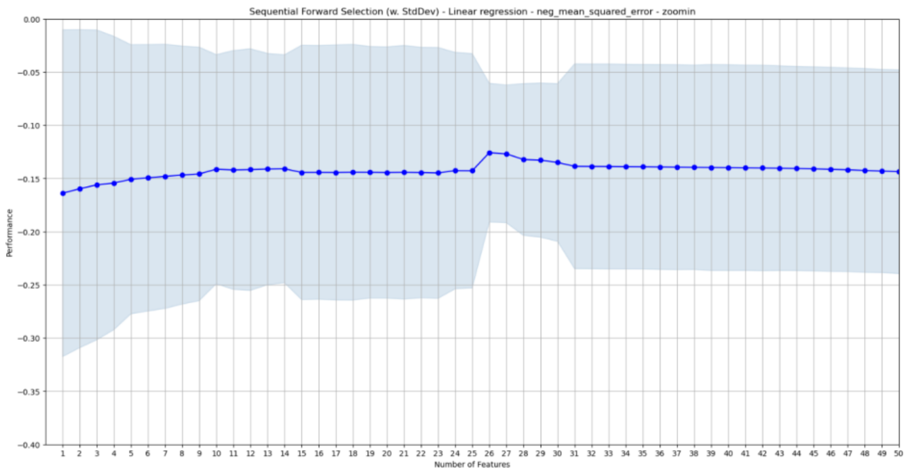
	Roman Briskine	[Science IT] ScienceCluster maintenance	04.09.2024 06:00	04.09.2024 22:15
	Roman Briskine	[Science IT] ScienceCluster service interruption	05.09.2024 08:00	09.09.2024 18:00

Appendix:

Process visualization of finding the best size of feature selection:

SFFS of linear regression, scoring = neg_mean_square_error, cv = 5. The best size is 26.





References:

https://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/
https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics
Joe Bemister-Buffington, Alex J. Wolf, Sebastian Raschka, and Leslie A. Kuhn (2020) Machine Learning to Identify Flexibility Signatures of Class A GPCR Inhibition Biomolecules 2020, 10, 454. https://www.mdpi.com/2218-273X/10/3/454#
Ferri, F. J., Pudil P., Hatef, M., Kittler, J. (1994). "Comparative study of techniques for large-scale feature selection." Pattern Recognition in Practice IV : 403-413.
Pudil, P., Novovičová, J., & Kittler, J. (1994). "Floating search methods in feature selection." Pattern recognition letters 15.11 (1994): 1119-1125.