Improvement of regression model

Between different k values for cross validation

Author: Yelu Date: 2024/9/6

Content: 1. Feature selection; 2. Correlation analysis + (3. Issues; 4. Appendix; 5. References.)

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 (Not finished)
    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
```

Regression model with all features

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

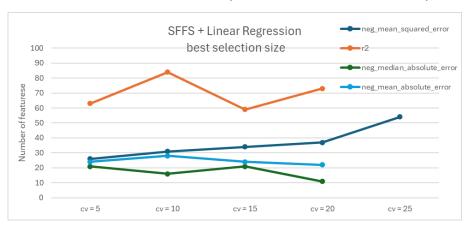
			R squared	Adj R	accuracy	precision	recall	f1 score	
	OLS	all	0.672	0.641	0.937	0.930	0.906	0.917	
Original		split	0.716	0.676	0.878	0.864	0.836	0.848	
Original	Juginai		Pseudo R so	quared	accuracy	precision	recall	f1 score	
	Logit	all	0.780)	0.949	0.936	0.935	0.936	
		split	0.837	,	0.960	0.954	0.932	0.942	

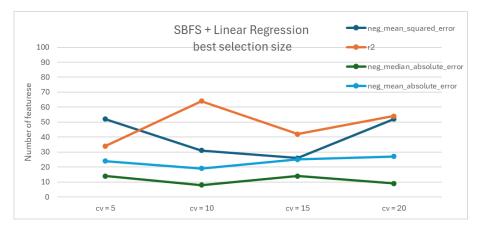
Find the best size of feature selection

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.

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.

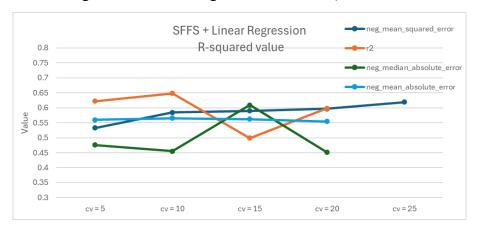
Summary table of feature selection results

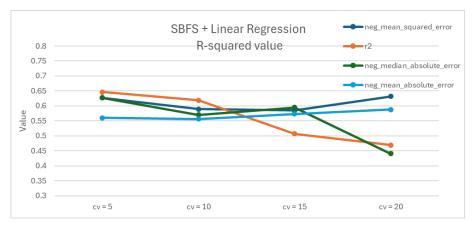
	SFFS		SBFS		SFS Tota	I	
No.	Feature	Number	Feature	Number	Feature	Number	
1	speedlimit_6	16	speedlimit_2	15	speedlimit_2	30	
2	trafficarea_2	15	trafficarea_2	13	trafficarea_2	28	
3	speedlimit_2	15	r_width_7	12	speedlimit_6	28	
4	r_surface_2	14	speedlimit_6	12	r_width_value	23	
5	r_width_6	_6				23	
6	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	r_width_7	22			
9	r_width_2	12	speedlimit_3	11	speedlimit_1	22	
10	r_surface_1	12	ср	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	ср	20	
15	r_width_4	11	r_surface_1	9	dtrafficarea	20	
16	speedlimit_1	11	speedlimit_4	9	cmax	19	
17	ср	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	
	trafficarea_1		z_qnr_10		z_qnr_23	14	
32	r_width_1	8	z_qnr_16	6	z_knr_9	14	
33	r_width_3		z_qnr_31		cmean	13	
	z_qnr_8	7	dtraml		z_qnr_25	13	
35	z_knr_9	7	dtrainl		r_width_1	13	
36	iemin	6	gvm_dwv		z_qnr_8	12	
37	cmean	6	z_qnr_8	5	z_qnr_10	11	

Apply selected features to 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)



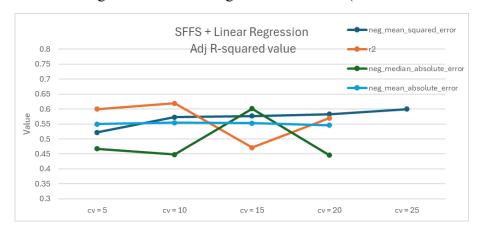


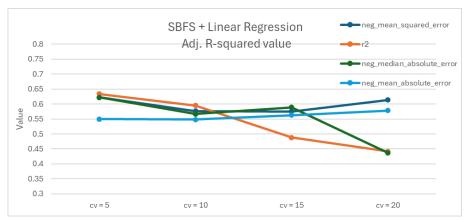
Average R-squared value of OLS regression model with feature selection

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 ac	lj 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.

Correlation analysis

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

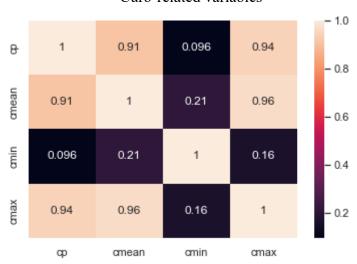
Summary of sorted correlation values of variable pairs is shown in the following table, only pairs with a correlation value higher than 0.5 or lower than -0.5 are included here. Among curb-related variables (cp, cmean, cmax) are correlated with each other.

Pairs o	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	ср	0.906			
iemax	ie	0.920			
cmax	ср	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

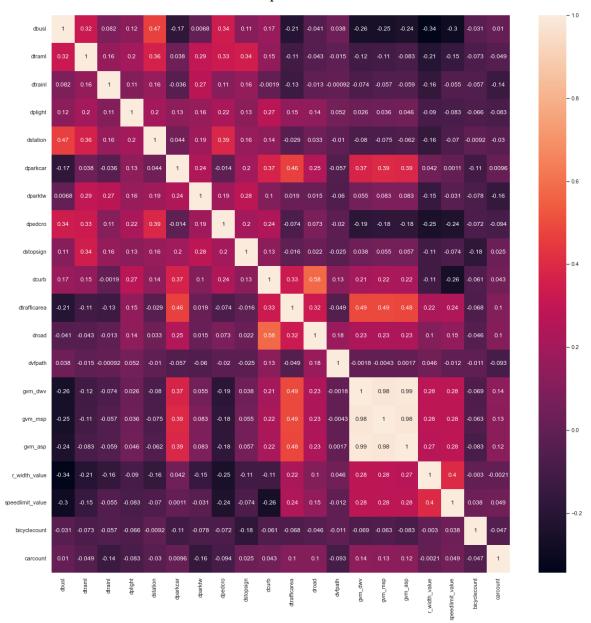
Curb-related variables



Entropy variables



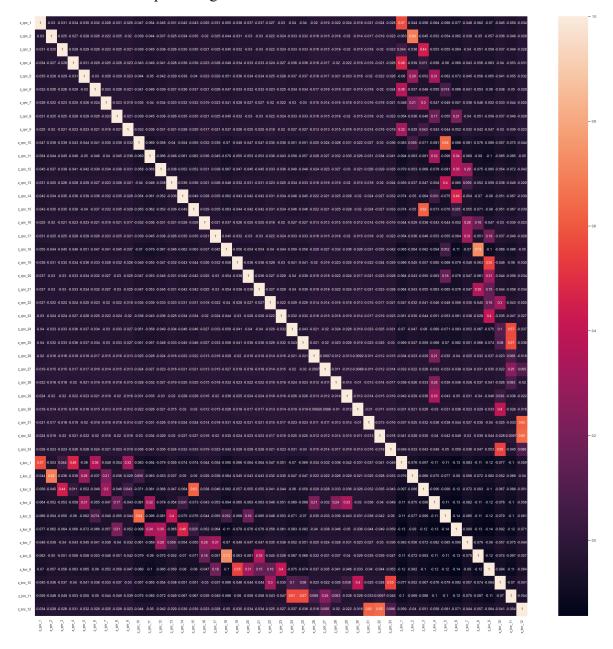
Traffic-transport numeric variables



Traffic-transport categorical variables - others

			_						_	_						_		_	_		
curbtype_1	1	-1	-0.091	-0.0095		-0.14	-0.13		-0.032		0.049	-0.065		-0.13	-0.12	-0.12			-0.091	-0.031	
curbtype_2	-1	1	0.091	0.0095	-0.16			-0.16	0.032	-0.11	-0.049	0.065	-0.13				-0.11	-0.19		0.031	
trafficarea_1	-0.091	0.091	1	-0.78	-0.58	-0.021	-0.0062		-0.026	0.0043	-0.00033		-0.042	0.042		-0.089	0.064	-0.21		-0.13	
trafficarea_2	-0.0095	0.0095	-0.78	1	-0.059			-0.055		-0.052		-0.052		-0.025	-0.17		-0.049		-0.0071		
trafficarea_3	0.16	-0.16	-0.58	-0.059	1	-0.034	-0.038	0.024	0.0026	0.06	-0.045		0.035	-0.035	-0.091		-0.037	0.058	-0.026	-0.022	
r_width_1	-0.14		-0.021	0.052	-0.034	1	-0.18	-0.31	-0.23	-0.37	-0.2	-0.13		-0.11		-0.0081	-0.17	-0.14	-0.022		
r_width_2	-0.13		-0.0062		-0.038	-0.18	1	-0.12	-0.092	-0.15	-0.079	-0.05	-0.0012	0.0012	0.0075		-0.02	-0.076	0.025	-0.028	
r_width_3	0.16	-0.16		-0.055	0.024	-0.31	-0.12	1	-0.16	-0.25	-0.13	-0.085	0.076	-0.076	-0.14	-0.14		-0.053	-0.019	-0.048	
r_width_4	-0.032		-0.026	0.03	0.0026	-0.23	-0.092	-0.16	1	-0.19	-0.1	-0.063	-0.3	0.3	-0.13		-0.15		-0.019		
r_width_5	0.11	-0.11	0.0043	-0.052	0.06	-0.37	-0.15	-0.25	-0.19	1	-0.16	-0.1		-0.09	-0.071	-0.14	-0.01			-0.036	
r_width_6	0.049	-0.049	-0.00033	3 0.035	-0.045	-0.2	-0.079	-0.13	-0.1	-0.16	1	-0.054	0.049	-0.049	-0.028	-0.0042		-0.064	-0.037	0.067	
r_width_7	-0.065	0.065		-0.052		-0.13	-0.05	-0.085	-0.063	-0.1	-0.054	1	-0.22		0.002		-0.083	-0.086	-0.023	-0.02	
r_surface_1	0.13	-0.13	-0.042				-0.0012	0.076	-0.3		0.049	-0.22	1	-1	0.084	-0.21	0.043			-0.037	
r_surface_2	-0.13		0.042	-0.025	-0.035	-0.11	0.0012	-0.076	0.3	-0.09	-0.049		-1	1	-0.084		-0.043	-0.077	-0.021		
speedlimit_1	-0.12			-0.17	-0.091		0.0075	-0.14	-0.13	-0.071	-0.028	0.002	0.084	-0.084	1	-0.34	-0.46	-0.37	-0.099	-0.083	
speedlimit_2	-0.12	0.12	-0.089	0.014		-0.0081		-0.14		-0.14	-0.0042		-0.21		-0.34	1	-0.25	-0.2	-0.054	-0.045	
speedlimit_3	0.11	-0.11	0.064	-0.049	-0.037	-0.17	-0.02	0.33	-0.15	-0.01		-0.083	0.043	-0.043	-0.46	-0.25	1	-0.27	-0.073	-0.061	
speedlimit_4	0.19	-0.19	-0.21		0.058	-0.14	-0.076	-0.053			-0.064	-0.086		-0.077	-0.37	-0.2	-0.27	1	-0.058	-0.049	
speedlimit_5	-0.091			-0.0071	-0.026	-0.022		-0.019	-0.019		-0.037	-0.023		-0.021	-0.099	-0.054	-0.073	-0.058	1	-0.013	
speedlimit_6	-0.031	0.031	-0.13		-0.022		-0.028	-0.048	0.05	-0.036	0.067	-0.02	-0.037	0.037	-0.083	-0.045	-0.061	-0.049	-0.013	1	
	aurbtype_1	aurbtype_2	afficarea_1	afficarea_2	afficarea_3	r_width_1	r_width_2	r_width_3	r_width_4	r_width_5	r_width_6	r_width_7	surface_1	surface_2	peedlimit_1	eedlimit_2	eedlimit_3	eedlimit_4	peedlimit_5	peedlimit_6	

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.

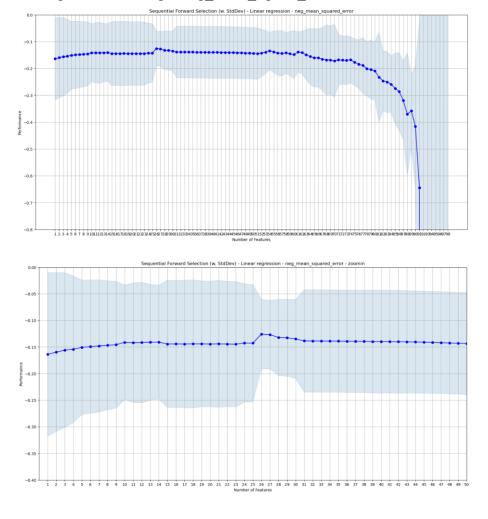
Working on correct feature selection with logistic regression has been paused since 09.04.

Roman Briskine [Science IT] ScienceCluster maintenance 04.09.2	024 06:00 04.09.2024 22:15
Roman Briskine [Science IT] ScienceCluster service interruption 05.09.20	09.09.2024 18:00

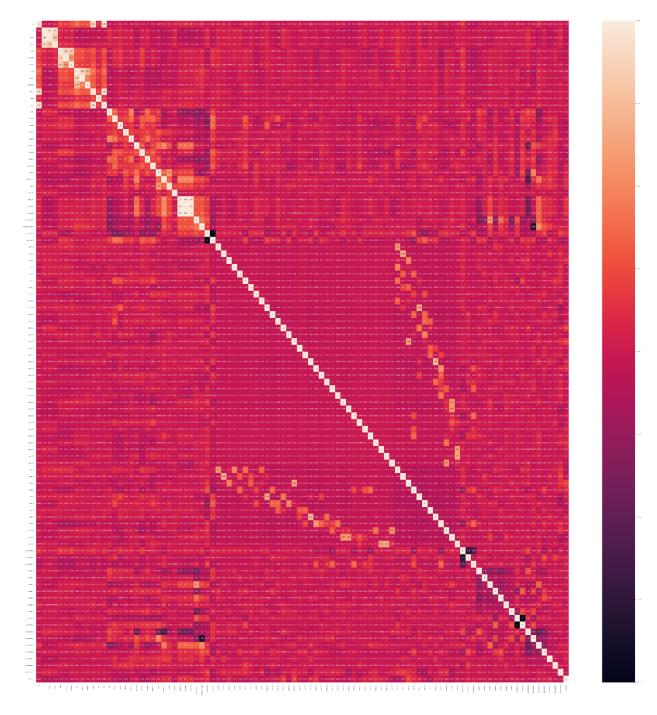
Appendix:

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.



Correlation matrix of all features:



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.