Improvement of regression model

Author: Yelu
Date: 2024/10/10
Content: 1. Checking distribution; 2. Replacing/filtering features; 3. Feature selection; 4. Logistic regression; 5. Calculating VIF; 6. Next steps.

Checking distribution

Distributions of curb-related variables of accident-present locations (acp = 1) and accident-absent locations (acp = 0) were compared using Kolmogorov-Smirnov test.

'cp':

```
KstestResult(statistic=0.084959767163157, pvalue=0.08376157447812585, statistic_location=0, statistic_sign=1)
```

'cmean':

```
KstestResult(statistic=0.4065656565656566, pvalue=2.2447396953763978e-32, statistic_location=-1.083670242208321, statistic_sign=-1)
```

'cmin':

```
KstestResult(statistic=0.9864406779661017, pvalue=2.1498339547925026e-256, statistic_location=-0.1172420763521098, statistic_sign=1)
```

'cmax':

```
KstestResult(statistic=0.4065656565656566, pvalue=2.2447396953763978e-32, statistic_location=-1.1269889934709814, statistic_sign=-1)
```

According to the p-value, we can reject the hypothesis that the distributions of 'cmean', 'cmin', 'cmax' of accident-present locations (acp = 1) and accident-absent locations (acp = 0) are the same.

Replacing features

After filtering variables (remove the n_{th} dummy variables, variables with high correlation, variables from 'unsuitable' dataset), there were **78** features (instead of 98) in total after the filtering process.

Two feature sets were built:

Set 1: Add 'cmean' to the feature set – 79 features
Set 2: Replace 'cp' with 'cmean' – 78 features

Feature selection

Feature set 1 (including curb-related variables: cp, cmean, cmin)

From the left table below, curb-related variables (**'cp', 'cmean', 'cmin'**) were often selected after feature selection using sffs/sbfs with logistic regression model.

(In total, 16 feature selections were performed. According to the table, for example, 'cmean' was selected 8 times out of the 8 sequential forward feature selections (sffs), and 4 times out of 8 sequential backward feature selections (sbfs), and therefore was selected 12 times out of 16 selection results.)

Feature set 2 (including curb-related variables: cmean, cmin)

From the right table below, only one curb-related variables (**'cmin'**) was often selected after feature selection using sffs/sbfs with logistic regression model.

(In total, 16 feature selections were performed. According to the table, for example, 'cmin' was selected 8 times out of the 8 sequential forward feature selections (sffs), and 1 times out of 8 sequential backward feature selections (sbfs), and therefore was selected 9 times out of 16 selection results.)

Table: Features that are selected in more than half selections

Feature set 1

	Logistic regression		
Features	sffs	sbfs	Total
r_surface_1	8	8	16
speedlimit_4	8	8	16
r_width_2	8	6	14
z_qnr_7	7	6	13
trafficarea_2	6	7	13
cmean	8	4	12
ie	8	4	12
z_qnr_23	7	5	12
z_knr_9	5	7	12
r_width_4	5	7	12
cp	7	4	11
cmin	8	3	11
z_qnr_5	6	5	11
z_knr_4	5	6	11
r_width_5	4	7	11
droad	6	4	10
z_qnr_26	5	5	10
mew	5	4	9
dvfpath	6	3	9
z_knr_6	5	4	9
r_width_6	5	4	9
meg	6	2	8
megmax	6	2	8
carcount	4	4	8
speedlimit_value	3	5	8
z_qnr_4	7	1	8
z_qnr_16	7	1	8

Feature set 2

	Logistic regression		
Feature	sffs	sbfs	Total
speedlimit_4	8	8	16
trafficarea_2	7	8	15
mew	8	6	14
ie	8	5	13
dvfpath	7	6	13
r_width_5	5	8	13
dtraml	7	5	12
droad	7	5	12
dtrafficarea	7	4	11
z_qnr_26	5	6	11
z_knr_4	3	8	11
r_surface_1	4	7	11
r_width_value	5	5	10
z_qnr_8	6	4	10
cmin	8	1	9
megmax	8	1	9
dparkcar	7	2	9
dpedcro	6	3	9
z_qnr_3	7	2	9
z_qnr_5	5	4	9
z_qnr_6	6	3	9
z_qnr_7	4	5	9
z_qnr_23	6	3	9
z_knr_9	2	7	9
r_width_3	5	4	9
r_width_4	2	7	9
r_width_6	5	4	9
dstation	6	2	8
speedlimit_value	4	4	8
z_qnr_4	7	1	8
z_qnr_10	4	4	8
z_qnr_29	6	2	8
r_width_2	4	4	8

Logistic regression

Logistic regression models were then applied with the 16 feature selection results for both feature set 1 and feature set 2. Features that were recognized significantly correlated with presence of accident were labels with ***/**/*, which was corresponding to the following significance code:

*Significance codes: 0 *** 0.001 ** 0.01 * 0.05.*

While curb-related variables ('cp', 'cmean') were mostly determined to be significantly correlated with presence of accident by using features selected from feature set 1, curb-related variables were never recognized to be significantly correlated with presence of accident by using features selected from feature set 2.

Feature set 1 (including curb-related variables: cp, cmean, cmin)

Based on the following table, 'cp' and 'cmean' were frequently recognized to be significantly correlated with presence of accident with '***' (a pvalue smaller than 0.001).

Regression model	Logistic regression																	
Methods	Sequential forward floating selection								Sequential backward selection							
Scoring	accuracy		f1		precision		recall		accuracy		f1		precision		recall	
Cross validation k	cv = 5	cv = 10	cv = 5	cv = 10	cv = 5	cv = 10	cv = 5	cv = 10	cv = 5	cv = 1	cv = 5	cv = 1	cv = 5	cv = 1	cv = 5	cv = 1
cp	***	***	***	***	***	***		***	***		***	***				***
cmean	***	***	***	***	***	***		***	***		***	***				***
cmin																
ie																
mew									*							
meg																
megmax																
dbusl																
dtraml																
dtrainl																
dplight																
dstation																
dparkcar																
dparktw																
dpedcro																
dstopsign								**								
dtrafficarea																
droad						*										
dvfpath					*								*			
gvm_dwv																
bicyclecount																
carcount																
r_width_value	***	***	***					*	***			**				***
speedlimit_value							*	***		***	***	**			***	
z_qnr_1	***		**		**		**		*							
z_qnr_2	*			**												
z_qnr_3																
z_qnr_4	*															
z_qnr_5			*	***					*	*						
z_qnr_6																
z_qnr_7	***	***	***	***		**	***	***	**	**	***	***			**	***
z_qnr_8	***		***	***					***							
z_qnr_9																
z_qnr_10	*						**	***			*	**			**	***
z_qnr_11														*		
z_qnr_12			**													*
z_qnr_13						*	*	***								
z_qnr_14																
z_qnr_15																
z_qnr_16	*															*
z_qnr_17																
z_qnr_18																
z_qnr_19																
z_qnr_20																
z_qnr_21																
z_qnr_22					*				*							
z_qnr_23	*				*									*		
z_qnr_24																
z_qnr_25							**	*							*	
z_qnr_26			*				**				**	**				
z_qnr_27																
z_qnr_28																
z_qnr_29																
z_qnr_30																
z_qnr_31																
z_qnr_32																
z_knr_1	**		*		**		**									
z_knr_2				**												
z_knr_3							*									
z_knr_4	***				***		*	*	*		***	***	***	***		***
z_knr_5	**							***			***	***				***
z_knr_6		***	***			**	***	**	**	**				*	***	
z_knr_7															**	
z_knr_8							***	**							***	*
z_knr_9	***		***	***	***				***	***	***	***	***	***		***
z_knr_10																
z_knr_11															*	
trafficarea_1																
trafficarea_2	***	**	***	**	**	**			***		***	***	***	***	***	**
r_width_1	**		***			**							*			
r_width_2	***	**		**	*	***	*	**	**	*	**	*				*
r_width_3			***			***			*							*
r_width_4			***	**	***			*	**	**	***	***	***	***		**
r_width_5			***			*			***	*		***	**	*		***
r_width_6			***			*			**							**
r_surface_1																
speedlimit_1					*	*				***	***					
speedlimit_3	***			***					***			**				**
speedlimit_4	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
speedlimit_5																
	20	8	19	13	14	14	14	16	20	10	14	16	8	9	10	19
prsquared	0.607698	0.414132	0.625552	0.573014	0.49317	0.447174	0.572131	0.548494	0.58795	0.488295	0.542318	0.604309	0.420023	0.410257	0.468542	0.568589

Feature set 2 (including curb-related variables: cmean, cmin)

Based on the following table, with feature set 2 (replacing 'cp' with 'cmean'), 'cmean' and 'cmin' were never recognized to be significantly correlated with presence of accident.

Regression model	Logistic regression																	
Methods	Sequential forward floating selection								Sequential backward selection							
Scoring	accuracy		f1		precision		recall		accuracy		f1		precision		recall	
Cross validation k	cv = 5	cv = 10	cv = 5	cv = 10	cv = 5	cv = 10	cv = 5	cv = 10	cv = 5	cv = 1	cv = 5	cv = 1	cv = 5	cv = 1	cv = 5	cv = 1
cmean																
cmin																
ie			*			*	*									
mew	*		*								*		**			**
meg																
megmax																
dbusl																
dtraml									**	*	*					
dtrainl																
dplight																
dstation																
dparkcar				*												
dparktw																
dpedcro																
dstopsign																
dtrafficarea										*					*	
droad																
dvfpath													*			
gvm_dwv																
bicyclecount																
carcount																
r_width_value	***	*	**			***	***				***		*	***	***	***
speedlimit_value			***	***			***	*			*	***			***	*
z_qnr_1	**		*						*							
z_qnr_2				**												
z_qnr_3						*										
z_qnr_4									***							
z_qnr_5		**		***												
z_qnr_6							*									
z_qnr_7		**	**	**				***			**	***		***	**	**
z_qnr_8	***	**	**	**		*	***				***		***		***	**
z_qnr_9									***							
z_qnr_10		*				**	***			**		***			**	***
z_qnr_11													*			
z_qnr_12							*				**		***			*
z_qnr_13							*									
z_qnr_14																
z_qnr_15						*										
z_qnr_16																
z_qnr_17	*															
z_qnr_18																
z_qnr_19																
z_qnr_20																
z_qnr_21																
z_qnr_22	**		**	**									**		*	
z_qnr_23																*
z_qnr_24																
z_qnr_25		*					*									
z_qnr_26	*		*				*					**			**	**
z_qnr_27							*									
z_qnr_28																
z_qnr_29																
z_qnr_30																
z_qnr_31																
z_qnr_32							**									
z_knr_1	**															
z_knr_2				**												
z_knr_3																
z_knr_4	***		**					***		***	**	***	*	***	**	***
z_knr_5										***		***				**
z_knr_6			**				***				***		***		***	**
z_knr_7							**									
z_knr_8							***								***	*
z_knr_9								***		***	***	***	***	***		***
z_knr_10																
z_knr_11						*										
trafficarea_1																
trafficarea_2	***	**	***	***	*	***	***		***	***	***	***	***	***	***	***
r_width_1																
r_width_2		**	***	*			*				**			*	*	*
r_width_3	*				**		**									**
r_width_4							**	***	***	***	**	***		**	**	***
r_width_5	**			**		***	***			*	***	***	*	***	***	***
r_width_6	*					*	**									**
r_surface_1																
speedlimit_1			***			*	***								***	
speedlimit_3		***														***
speedlimit_4	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
speedlimit_5																
	14	10	15	12	3	12	23	6	7	10	15	11	13	9	17	22
prsquared	0.532582	0.473943	0.536686	0.44876962	0.331298	0.418232	0.599347	0.496183	0.408148	0.425548	0.563509	0.553867	0.517529	0.448181	0.560432	0.571066

Calculating VIF

Variance Impact Factor

feature	value
cmin	1.083201168
dvfpath	1.134604506
ie	1.199886311
dplight	1.233585502
mew	1.335104263
droad	1.358461927
z_qnr_30	1.42571221
carcount	1.43872171
dpedcro	1.527087768
bicyclecount	1.534458102
dparktw	1.618559575
dstation	1.706839679
dtrainl	1.729131231
gvm_dwv	1.733747381
dparkcar	1.753985256
z_qnr_27	1.763581354
speedlimit_5	1.821765535
dbusl	1.862022415
z_qnr_26	1.943967563
z_qnr_22	2.024978431
z_qnr_28	2.034404136
dtraml	2.099011085
dstopsign	2.119914749
z_qnr_16	2.22990004
z_qnr_31	2.251549492
z_qnr_29	2.380769849
z_qnr_32	2.391033266
megmax	2.515353626
dtrafficarea	2.664046873
z_qnr_17	2.671171598
meg	2.774207537
cp	2.983429165
z_qnr_8	3.533521469
z_qnr_23	3.534120534
r_width_2	3.626124321
z_qnr_13	3.74557533
z_qnr_7	3.904742644
trafficarea_2	4.003815986
z_qnr_9	4.139843141
z_qnr_25	4.296993256
z_qnr_24	4.306832434
z_qnr_5	4.409050994
z_qnr_21	4.562618818
z_qnr_20	4.648249471
z_qnr_2	4.73777926
speedlimit_value	4.774984911
z_qnr_19	4.84101957

Next steps

Analysis

- For dependent variable, ‘presence of accident’, continue filtering/selecting features considering VIF and applying regression models
- Start/continue working on feature selection and regression analysis for other dependent varialbes, severity of accident

Writing

- Create overleaf project and start basic structure of the paper

Formatting

- Check journal formatting style
- Check recent issues of the journal