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=0.40656565656565666, pvalue=2.2447396953763978e-32, statistic_locatio n=-1.083670242208321, statistic_sign=-1)

'cmin':

KstestResult(statistic=0.9864406779661017, pvalue=2.1498339547925026e-256, statistic_locatio on=-0.1172420763521098, statistic_sign=1)

'cmax':
```

KstestResult(statistic=0.40656565656565656, pvalue=2.2447396953763978e-32, statistic_locatio n=-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
Feature set 2

	Logis	tic regre	ession
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
ср	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

	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 <u>mostly</u> determined to be significantly correlated with presence of accident by using features selected from feature set 1, curb-related variables were <u>never</u> 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).

Methods	el Logis Sequential forward floating selection								Logistic regression Sequential backward selection								
coring	acci	accuracy f1 precision recall						call	accu	ıracy		1	prec		rec	call	
ross 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	
0	***	***	***	***	***	***		***	***		***	***				***	
mean	***	***	***	***	***	***		***	***		***	***				***	
min																	
new									*								
neg									1								
negmax									1								
lbusl																	
ltraml																	
ItrainI																	
lplight																	
Istation																	
lparkcar																	
lparktw																	
pedcro																	
Istopsign								**									
Itrafficarea																	
Iroad						*											
vfpath					*								*				
vm_dwv									-								
icyclecount									-								
arcount	***	***	***					*	***			**				***	
_width_value peedlimit_value	***	***	***				*	***	***	***	***	**			***	xxx	
peeaiimit_value _qnr_1	***		**		**		*	200	*		****						
_qnr_1 _qnr_2	*			**					1"								
_qrii_2 _qnr_3				† <u>"</u>													
_qnr_4	*								1								
_qnr_5			*	***					*	*							
_qnr_6																	
_qnr_7	***	***	***	***		**	***	***	**	**	***	***			**	***	
_qnr_8	***		***	***					***								
_qnr_9																	
_qnr_10	*						**	***			*	**			**	***	
_qnr_11														*			
_qnr_12			**													*	
_qnr_13						*	*	***									
_qnr_14																	
_qnr_15																	
_qnr_16	*															*	
_qnr_17																	
_qnr_18																	
_qnr_19																	
_qnr_20																	
_qnr_21					*												
_qnr_22					*				*					*			
:_qnr_23 :_qnr_24	*				*									*			
_qnr_25							**	*							*		
_qnr_26			*				**				**	**			^		
_qnr_27																	
_qnr_28																	
_qnr_29																	
_qnr_30																	
_qnr_31																	
_qnr_32																	
_knr_1	**		*		**		**										
_knr_2				**													
_knr_3							*										
_knr_4	***				***		*	*	*		***	***	***	***		***	
_knr_5	**							***			***	***				***	
_knr_6		***	***			**	***	**	**	**				*	***		
_knr_7															**		
_knr_8							***	**							***	*	
_knr_9	***		***	***	***				***	***	***	***	***	***		***	
_knr_10																	
_knr_11															*		
afficarea_1																	
afficarea_2	***	**	***	**	**	**			***		***	***	***	***	***	**	
width_1	**		***			**			l				*				
width_2	***	**		**	*	***	*	**	**	*	**	*				*	
width_3			***			***			*	4-4	***		atastas.			*	
width_4			***	**	***			*	**	**	***	***	***	***		**	
width_5			***			*			***	*		***	**	*		***	
width_6			***			*			**							**	
surface_1										atustusts	***						
peedlimit_1				***	*	*				***	***	**				ded	
peedlimit_3	***	***	de de la	***	www	***	***	de de de	***	destrate	de de de	**	***	***	dedede	**	
peedlimit_4	~ * *	~**	***	~**	***		***	***	488	***	***	***	0.55	~**	***	488	
									1								
eedlimit_5																	

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.

oring oss validation k conean nin ew * eg egmax susl raml raml raml ramin light tation sarkcar sarktw seedero topsign rafficarea oad frpath m_dwv cyclecount rcount width_value egnr_1 qnr_2 qnr_3 qnr_4 qnr_5 qnr_6 qnr_7 qnr_8 qnr_9 qnr_10 qnr_11 qnr_12 qnr_10 qnr_11 qnr_12 qnr_12 qnr_13 qnr_14 qnr_13 qnr_14 qnr_15 qnr_19 qnr_10 qnr_11 qnr_12 qnr_10 qnr_11 qnr_12 qnr_13 qnr_14	v = 5	uracy cv = 10	cv = 5	ntial forward f1 cv = 10		cv = 10	re-cv = 5	call cv = 10	acct	cv = 1	f	1		ision		call cv = 1
oss validation k conean nin ew * eg egmax susl raml rainl slight tation sarkcar sarktw sedcro topsign rafficarea oad fpath rount width_value egnr_1 egnr_3 egnr_4 egnr_5 egnr_6 egnr_7 egnr_8 ew egnr_10 egnr_11 enr_12 enr_12 enr_12 enr_12 enr_12 enr_10 enr_11 enr_12 enr_12 enr_13 enr_11 enr_12 enr_13 enr_11 enr_12 enr_13 enr_11 enr_12 enr_13 enr_11 enr_12 enr_13 enr_13 enr_11	***		* *		cv = 5			cv = 10			*	cv = 1		cv = 1		
ew * eg egmax usl raml raml raml rainI Jiight tation aarktw bedero topsign rafficarea oad ffpath m_dwv syclecount rcount width_value eqnr_1 qnr_2 qnr_3 qnr_4 qnr_5 qnr_6 qnr_7 qnr_8 qnr_9 qnr_10 qnr_11 qnr_12 qnr_12 qnr_11 qnr_12 qnr_13 qnr_11 qnr_11 qnr_11			*	*		*	*		**	*			**			**
ew * eg egmax pust raml ramin lalight tation parktw pedcro topsign rafficarea oad ffpath roount width value eedlimit value qnr.1 qnr.2 qnr.3 qnr.4 qnr.6 qnr.7 qnr.8 ** qnr.9 qnr.1		×	*	*		*	*		***	*			**			**
ew * eg eg eg max pusl raml raml ramil light tation parkcar parktw pedcro topsign rafficarea oad fpath m_dwv pyclecount rcount width_value eddlimit_value qnr_1 qnr_2 qnr_3 qnr_4 qnr_5 qnr_6 qnr_7 qnr_8 qnr_9 qnr_10 qnr_11 qnr_12 qnr_11 qnr_12 qnr_11 qnr_12 qnr_11 qnr_11 qnr_11 qnr_11 qnr_11 qnr_11 qnr_11 qnr_11		*	*	*		*	*		**	*			**			**
eg egmax pusl raml raml raml rainl light tation parkcar parktw pedcro topsign rafficarea oad ffpath m_dwv syclecount rcount width_value eqnr_1 qnr_2 qnr_3 qnr_4 qnr_5 qnr_6 qnr_7 qnr_8 qnr_9 qnr_10 qnr_11 qnr_12 qnr_11 qnr_12 qnr_12 qnr_10 qnr_11 qnr_11 qnr_11 qnr_11		*		*					**	*			**			**
egmax usl raml rainl light tation avarkcar varktw vedero topsign rafficarea oad ffpath m_dwv cyclecount rcount width_value eedlimit_value qnr_1 = ** qnr_2 qnr_3 qnr_4 qnr_5 qnr_6 qnr_7 qnr_8 *** qnr_9 qnr_10 qnr_11 qnr_12 qnr_11 qnr_12 qnr_13 qnr_11 qnr_11		*		*					**	*	*					
ousl ramil ramil ramil rainil plight tation parkcar parktw pedcro topsign rafficarea oad fpath m_dwv cyclecount rcount width_value eedlimit_value qnr_1 qnr_2 qnr_3 qnr_4 qnr_5 qnr_6 qnr_7 qnr_8 qnr_9 qnr_10 qnr_11 qnr_12 qnr_11 qnr_12 qnr_12 qnr_11		*		*					**	*	*					
raml ramil rainil light tation sarkcar sarktw seedcro topsign rafficarea oad fipath m_dwv syclecount rcount width_value eddlimit_value gnr_1 qnr_2 qnr_3 qnr_4 qnr_5 qnr_6 qnr_7 qnr_8 qnr_9 qnr_10 qnr_11 qnr_12 qnr_10 qnr_11 qnr_12 qnr_12 qnr_13 qnr_14 qnr_15 qnr_16 qnr_17 qnr_18 qnr_18 qnr_19 qnr_19 qnr_10 qnr_11		*		*					**	*	*					
rainI light tation tation parkcar parktw peddro peddro topsign rafficarea oad fipath m_dwv tyclecount rcount rcount ayr.1 ayr.2 ayr.3 ayr.4 ayr.5 ayr.6 ayr.7 ayr.8 ayr.9 ayr.10 ayr.10 ayr.11 ayr.10 ayr.11 ayr.10 ayr.11 ayr.10 ayr.10 ayr.11 ayr.12 ayr.13		*		*					**	*	*					
slight tation sarkcar sarktw sedcro topsign rafficarea oad fpath rount width value sarcal sartination rount side fine fine fine sartination sartinatio		*		*												
tation parkcar parktw pedcro topsign rafficarea oad fpath m_dwv pyclecount reount vioth_value eedlimit_value qnr_1 qnr_2 qnr_3 qnr_4 qnr_5 qnr_6 qnr_7 qnr_8 qnr_9 qnr_10 qnr_11 qnr_12 qnr_10 qnr_11 qnr_12 qnr_11 qnr_12 qnr_13 qnr_11		*		*												
parkcar parktw pedcro topsign rafficarea oad frpath m_dwv pyclecount rcount width_value eedlimit_value qnr_1 qnr_2 qnr_3 qnr_4 qnr_1 qnr_15 qnr_6 qnr_7 qnr_8 qnr_9 qnr_10 qnr_11 qnr_12 qnr_10 qnr_11 qnr_12 qnr_13 qnr_10 qnr_11		*		*												
parktw parktw packero		*		*												
pedcro topsign rafficarea oad fpath m_dwv syclecount roount width_value eedlimit_value qnr_1 qnr_2 qnr_3 qnr_4 qnr_5 qnr_6 qnr_7 qnr_8 qnr_9 qnr_10 qnr_11 qnr_12 qnr_11 qnr_12 qnr_12 qnr_11 qnr_12 qnr_13		*														
topsign rafficarea oad fpath m_dwv syclecount reount width_value eedlimit_value qnr_1 qnr_2 qnr_3 qnr_4 qnr_5 qnr_5 qnr_6 qnr_7 qnr_8 qnr_9 qnr_10 qnr_11 qnr_10 qnr_11 qnr_12 qnr_12 qnr_12 qnr_13		*														
rafficarea oad fipath m_dwv syclecount rccount rccunt width_value eedlimit_value qnr_1 qnr_2 qnr_3 qnr_4 qnr_5 qnr_6 qnr_7 qnr_8 *** qnr_9 qnr_10 qnr_11 qnr_11 qnr_11 qnr_11 qnr_12 qnr_13 qnr_11		*														
oad fpath m_dwv cyclecount roount width_value eedlimit_value qnr_1 qnr_2 qnr_3 qnr_4 qnr_5 qnr_6 qnr_7 qnr_8 qnr_9 qnr_10 qnr_11 qnr_12 qnr_12 qnr_12 qnr_13 qnr_13		*														
frpath m_dwv ycyclecount rcount vioth_value eedlimit_value qnr_1 qnr_2 qnr_3 qnr_4 qnr_5 qnr_6 qnr_7 qnr_8 qnr_9 qnr_10 qnr_11 qnr_12 qnr_12 qnr_13 qnr_13 qnr_13 qnr_11		*								*					*	
m_dwv syclecount recount recount width_value eedlimit_value qnr_1 qnr_2 qnr_3 qnr_4 qnr_5 qnr_6 qnr_7 qnr_8 qnr_9 qnr_10 qnr_11 qnr_12 qnr_13		*														
cyclecount recount width_value eedlimit_value qnr_1		*											*			
roount width_value eedlimit_value qnr_1		*														
*** *** *** *** *** *** *** *** *** *** *** *** * **		*							-							
eedlimit_value qnr_1 qnr_1 qnr_2 qnr_3 qnr_4 qnr_5 qnr_6 qnr_7 qnr_8 qnr_9 qnr_10 qnr_11 qnr_11 qnr_12 qnr_13		*														
qnr_1 ** qnr_2 qnr_3 qnr_4 qnr_5 qnr_6 qnr_7 qnr_6 qnr_7 qnr_8 ** qnr_10 qnr_11 qnr_12 qnr_13 qnr_11	rk		**			***	***		-		***		*			***
qnr_2 qnr_3 qnr_4 qnr_5 qnr_6 qnr_7 qnr_8 qnr_9 qnr_10 qnr_11 qnr_12 qnr_12 qnr_12	nk		***	***			***	*			*	***			***	*
qnr_3 qnr_4 qnr_5 qnr_6 qnr_7 qnr_8			*						*							
qnr_4 qnr_5 qnr_6 qnr_7 qnr_8 ** qnr_9 qnr_10 qnr_11 qnr_12 qnr_12 qnr_13				**												
qnr_5 qnr_6 qnr_7 qnr_8 ** qnr_9 qnr_10 qnr_11 qnr_12 qnr_12 qnr_13						*			l							
qnr_6 qnr_7 qnr_8 ** qnr_9 qnr_10 qnr_11 qnr_12 anr 13									***							
qnr_7 qnr_8 ** qnr_9 qnr_10 qnr_11 qnr_12 qnr_13		**		***												
qnr_8 ** qnr_9 qnr_10 qnr_11 qnr_12 qnr_13		**					*	***			4-4	at at at		aturbut.	4-4	ded
qnr_9 qnr_10 qnr_11 qnr_12 qnr_13		**		**		*		***	-		**	***				**
qnr_10 qnr_11 qnr_12 qnr 13	**	**	**	**		*	***				***		***		***	**
qnr_11 qnr_12 anr 13		*							***							
qnr_12 anr 13		*				**	***			**		***			**	***
anr 13													*			
qnr_13							*				**		***			*
							*									
qnr_14																
qnr_15						*										
qnr_16																
qnr_17 *																
qnr_18																
qnr_19																
qnr_20																
qnr_21																
qnr_22 **	n*		**	**									**		*	
qnr_23																*
qnr_24																
qnr_25		*					*									
qnr_26 *			*				*					**			**	**
qnr_27							*									
qnr_28																
qnr_29																
gnr_30																
qnr_31																
qnr_32							**									
knr_1 **	n#															
knr_2				**												
knr_3																
	nk-k		**					***		***	**	***	*	***		***
knr_5										***		***				**
knr_6			**				***				***		***		***	**
knr_7							**									
knr_8							***									*
knr_9								***		***	***	***	***	***		***
knr_10																
knr_11						*										
officarea_1																
	nk-k	**	***	***	*	***	***		***	***	***	***	***	***	***	***
vidth_1																
vidth_2		**	***	*			*				**			*	*	*
width_3 *					**		**									**
width_4							**	***	***	***	**	***				***
width_5 **	n*			**		***	***			*	***	***	*	***		***
width_6 *	r					*	**									**
surface_1																
eedlimit_1			***			*	***								***	
eedlimit_3		***														***
	nkrik	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
eedlimit_5																
_																

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
ср	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 variables, 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