Improvement of regression model

Author: Yelu
Date: 2024/10/29
Content:

Presence of accident

Filtering features using VIF

Calculated VIF value for each feature and removed features with high VIF values (>=5) until all remaining features have VIF values less than 5. In total, there were 72 features after the filtering process.

Feature	VIF				
cmin	1.106	z_qnr_17	1.636	z_qnr_24	2.187
dvfpath	1.131	z_qnr_22	1.648	z_qnr_19	2.268
ie	1.172	dstation	1.687	z_qnr_7	2.473
speedlimit_5	1.222	dparkcar	1.716	speedlimit_1	2.477
dplight	1.223	gvm_dwv	1.722	z_qnr_12	2.488
z_qnr_16	1.297	z_qnr_23	1.738	megmax	2.504
mew	1.346	z_qnr_29	1.744	r_width_value	2.513
droad	1.349	r_width_2	1.751	dtrafficarea	2.570
z_qnr_30	1.366	speedlimit_value	1.752	z_qnr_14	2.593
z_qnr_27	1.367	speedlimit_4	1.800	z_knr_4	2.684
r_surface_1	1.376	dbusl	1.851	z_qnr_3	2.691
cmean	1.390	r_width_5	1.902	meg	2.800
r_width_6	1.423	z_qnr_8	1.905	z_qnr_5	2.887
carcount	1.426	dstopsign	1.928	r_width_1	2.901
z_qnr_31	1.431	z_qnr_20	1.938	z_knr_8	3.234
z_qnr_21	1.485	z_qnr_4	1.964	z_qnr_11	3.478
dpedcro	1.491	z_qnr_13	2.023	z_qnr_10	3.510
z_qnr_32	1.511	dtraml	2.024	z_qnr_2	3.531
bicyclecount	1.526	z_knr_7	2.048	z_knr_5	3.769
z_qnr_26	1.537	z_qnr_6	2.052	trafficarea_2	3.775
z_qnr_28	1.572	z_knr_10	2.082	z_knr_2	3.998
z_qnr_9	1.580	r_width_4	2.141	z_qnr_15	4.047
dparktw	1.599	z_qnr_25	2.141	z_knr_3	4.337
dtrainl	1.626	z_qnr_1	2.154	trafficarea_1	4.596

Feature selection

By using sequential forward/backward floating feature selection methods, features which were selected in more than half selections were listed as follows:

feature	sffs	sbfs	total
speedlimit_4	8	8	16
trafficarea_2	7	8	15
dpedcro	8	6	14
ie	8	5	13
dvfpath	7	6	13
r_width_5	6	7	13
dtraml	8	4	12
z_qnr_10	6	6	12
cmean	8	3	11
z_qnr_7	4	7	11
z_knr_4	4	7	11
r_width_2	5	6	11
speedlimit_1	5	6	11
z_knr_5	5	5	10
r_width_4	4	6	10
meg	8	1	9
dtrafficarea	7	2	9
r_width_value	3	6	9
z_qnr_3	7	2	9
z_qnr_6	8	1	9
z_qnr_26	4	5	9
z_qnr_29	7	2	9
z_knr_10	4	5	9
cmin	8	0	8
mew	8	0	8
megmax	8	0	8
dbusl	6	2	8
dplight	8	0	8
dparkcar	6	2	8
carcount	6	2	8
speedlimit_value	3	5	8
z_qnr_2	8	0	8
z_qnr_4	8	0	8
z_qnr_5	4	4	8
z_qnr_11	5	3	8
z_qnr_16	7	1	8
z_qnr_22	5	3	8
z_qnr_31	7	1	8
r_width_6	5	3	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.*

	Logistic regression																	All 72 features without feature selection			
Methods	Sequential forward floating selection								Sequential backward selection									linear regression ols		logistic regression	
Scoring	accuracy		f1		precision		recall		accuracy		f1		precision		recall			all		all	
Cross validation k	cv = 5	cv = 10	cv = 5	cv = 10	cv = 5	cv = 10	cv = 5	cv = 10	cv = 5	cv = 10	cv = 5	cv = 10	cv = 5	cv = 10	cv = 5	cv = 10		without sp	with split	without sp	with split
cmean																			*		
cmin																					
ie																					
mew																					
meq			*						*												
meqmax																					
dbusl																					
dtraml																					
dtrainl																					
dplight																					
dstation																					
dparkcar				*																	
dparktw																					
dpedcro																					
dstopsign																					
dtrafficarea																		*	**		
droad																					
dvfpath																					
qvm_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_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_2		***																			
z_knr_3	*																				
z_knr_4			***	**			***	***	***	***	***	***		***	***	***		***	***	***	***
z_knr_5	*		**	***			*	***		**	***	**			**	**		***	***	***	***
z_knr_7																					
z_knr_8							**									**		*	*	***	**
z_knr_10					*																
trafficarea_1																		**	*		
trafficarea_2	***	***	***	***	***	***	***		***	***	***	***	***	***	***	***		***	***	***	*
r_width_1			*								*					*		*			
r_width_2		***	***	**			**	**	***	**		***	**	**		***		***	**	**	**
r_width_4	***			**			**	***		**	***		**	**	***				*		
r_width_5				***			**	**		*	***		*	*	***					*	
r_width_6	*										**				*						
r_surface_1																		***	***		
speedlimit_1	*						***		**	***	***	***			***	***		**	**	***	***
speedlimit_4	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***		***	***	***	***
speedlimit_5							*													*	
pseudo r squared	0.438687	0.449274	0.492	0.549	0.360	0.406	0.560	0.480	0.440	0.512	0.511	0.499	0.386	0.433	0.519	0.522	adj/pseudo r squ	0.540	0.554	0.596	0.605

In conclusion, although curb-related variables were included in the result of feature selection, they were not recognized as significantly correlated with presence of accident.

Severity of accident

Filtering features using VIF

feature	VIF
z_qnr_29	1.165401
z_qnr_26	1.201286
cmin	1.232692
z_qnr_28	1.273350
z_qnr_27	1.296253
dplight	1.324291
ie	1.325090
bicyclecount	1.332474
r_width_2	1.347994
z_qnr_9	1.355616
z_qnr_13	1.358190
z_qnr_4	1.362768
z_qnr_14	1.387864
z_qnr_24	1.428709
z_qnr_15	1.436011
z_knr_8	1.439084
z_qnr_25	1.458042
dpedcro	1.458747
r_width_3	1.462649
z_qnr_2	1.466114
z_qnr_3	1.478381
z_qnr_10	1.501965
z_qnr_18	1.505239
r_width_6	1.526526
carcount	1.555751
cmean	1.612371
dstation	1.623488
z_qnr_17	1.632699
r_width_4	1.686983
z_qnr_20	1.783044
dparktw	1.815367
z_qnr_6	1.835370
dtrainl	1.900883
dparkcar	1.963168
gvm_dwv	1.970928
z_qnr_21	1.978967
speedlimit_value	2.000202
mew	2.042815
z_qnr_22	2.075581
dbusl	2.084388
z_knr_7	2.142196
r_width_value	2.155611
dtraml	2.168556
z_knr_1	2.187285
z_qnr_5	2.239490
speedlimit_1	2.253249
trafficarea_1	2.281662
z_qnr_19	2.336563
z_knr_3	2.344158
dvfpath	2.417156
z_qnr_8	2.425914
dtrafficarea	2.503804
dstopsign	2.524836
megmax	2.741690
droad	2.788528
meg	3.248079

In total, there were 56 features after the filtering process.

Checking distribution

For each binary dependent variable y, checked distributions of each curb-related variable between y-present locations and y-absent locations.

Considering severity of accident, there are four binary dependent variables, including ‘presence of person injury’, ‘presence of property damage’, ‘presence of light person injury’, and ‘presence of severe person injury’. Curb-related variables include ‘cp’, ‘cmean’, ‘cmin’, ‘cmax’.

Presence of person injury:

```
check_cv_dist(dfacrsv, 'inp')
```

```
Last executed at 2024-10-29 07:32:28 in 8ms
```

```
cp KstestResult(statistic=0.00980392156862745, pvalue=1.0, statistic_location=0, statistic_sign=-1)
cmean KstestResult(statistic=0.027450980392156862, pvalue=0.9999999999999981, statistic_location=1.7519618970807305, statistic_sign=-1)
cmin KstestResult(statistic=0.01568627450980392, pvalue=1.0, statistic_location=-0.1172420763521098, statistic_sign=-1)
cmax KstestResult(statistic=0.03137254901960784, pvalue=0.999999999995699, statistic_location=1.4358700609298287, statistic_sign=-1)
```

Presence of property damage

```
check_cv_dist(dfacrsv, 'pdp')
```

```
Last executed at 2024-10-29 07:32:28 in 7ms
```

```
cp KstestResult(statistic=0.0574606968833318, pvalue=0.9504836658059234, statistic_location=0, statistic_sign=-1)
cmean KstestResult(statistic=0.0574606968833318, pvalue=0.9504836658059234, statistic_location=-0.9559721771356244, statistic_sign=-1)
cmin KstestResult(statistic=0.013836535809506298, pvalue=1.0, statistic_location=-0.1172420763521098, statistic_sign=1)
cmax KstestResult(statistic=0.0574606968833318, pvalue=0.9504836658059234, statistic_location=-0.9857218406672054, statistic_sign=-1)
```

Presence of light person injury

```
check_cv_dist(dfacrsv, 'lvp')
```

```
Last executed at 2024-10-29 07:32:28 in 8ms
```

```
cp KstestResult(statistic=0.003409865878608775, pvalue=1.0, statistic_location=0, statistic_sign=-1)
cmean KstestResult(statistic=0.0330188679245283, pvalue=0.9999989166149185, statistic_location=1.7519618970807305, statistic_sign=-1)
cmin KstestResult(statistic=0.018867924528301886, pvalue=1.0, statistic_location=-0.1172420763521098, statistic_sign=-1)
cmax KstestResult(statistic=0.026653784951125255, pvalue=0.999999999428004, statistic_location=0.7630739452214893, statistic_sign=1)
```

Presence of severe person injury

```
check_cv_dist(dfacrsv, 'svp')
```

```
Last executed at 2024-10-29 07:32:29 in 7ms
```

```
cp KstestResult(statistic=0.022399203583872575, pvalue=1.0, statistic_location=0, statistic_sign=1)
cmean KstestResult(statistic=0.028455284552845527, pvalue=0.9999999999973915, statistic_location=1.7519618970807305, statistic_sign=1)
cmin KstestResult(statistic=0.016260162601626018, pvalue=1.0, statistic_location=-0.1172420763521098, statistic_sign=1)
cmax KstestResult(statistic=0.022399203583872575, pvalue=1.0, statistic_location=-0.9857218406672054, statistic_sign=1)
```

In conclusion, distributions of 'cp', 'cmean', 'cmin', 'cmax' are the **same** between person-injury-present and person-injury-absent locations, between property-damage-present and property-damage-absent locations, between lightly-injured-person-present and lightly-injured-person-absent locations, between severely-injured-person-present and severely-injured-person-abresent locations.

Regression without feature selection

Presence of person injury (Logistic regression):

Pseudo R-squ: 0.4131

	feature	0	sig
0	ie	0.006142	**
1	droad	0.014193	*
2	speedlimit_1	0.019451	*

Presence of property damage(Logistic regression):

Pseudo R-squ: 0.2915

	feature	0	sig
0	meg	0.03918	*
1	megmax	0.00387	**

Presence of light person injury (Logistic regression)

Pseudo R-squ: 0.3082

	feature	0	sig
0	ie	0.026392	*
1	dtraml	0.037193	*
2	z_qnr_2	0.038523	*
3	z_qnr_10	0.020944	*

Presence of severe person injury (Logistic regression)

Pseudo R-squ: 0.3503

	feature	0	sig
0	droad	0.012907	*
1	z_qnr_2	0.008748	**
2	z_qnr_6	0.012194	*
3	r_width_3	0.027291	*
4	r_width_4	0.025105	*
5	speedlimit_1	0.049703	*

Number of lightly injured persons (OLS linear regression)

R-squ: 0.389

Adj.R-squ: 0.118

	feature	0	sig
0	const	7.924183e-08	***
1	dstation	1.104676e-02	*
2	dtrafficarea	4.153163e-02	*
3	z_qnr_5	3.022138e-02	*
4	z_qnr_24	3.150583e-02	*
5	z_qnr_28	8.765195e-03	**
6	z_knr_11	6.827384e-04	***
7	trafficarea_1	1.222460e-02	*

Number of severely injured persons(OLS linear regression)

R-squ: 0.326

Adj.R-squ: 0.027

	feature	0	sig
0	z_knr_1	0.024746	*
1	r_width_6	0.015962	*

Value of property damage(OLS linear regression)

R-squ: 0.365

Adj.R-squ: 0.084

	feature	0	sig
0	dstation	0.002152	**
1	dpedcro	0.002711	**
2	z_qnr_20	0.016318	*
3	z_knr_1	0.038539	*

Next steps

1. Focus on presence of accident first, apply Random Forest Classification model and compare the results with results of logistic regression [– if ‘cp’, ‘cmean’ are both included, are they recognized as significant/important in both models?];
2. Try with PCA first and then apply regression model;
(3. Avoid overlapping of locations with ‘severe person injury’/‘light person injury’)