Improvement of analysis on presence of accident

Author: Yelu
Date: 2024/01/22
Content:

Presence of accident

Correct errors

Errors in processing and transforming variables (normalization & standardization, etc.) were corrected; 'dvfpath' was removed. One-hot encoding of each categorical variables is listed in appendix.

Filtering variables using correlation & VIF

In total there were 91 variables, after filtering the ones with high correlation (14) and high multicollinearity (17), features left are shown as follows.

Features after filtering: 60

'cp', 'cmin', 'ie', 'mew', 'meg', 'megmax', 'dbusl', 'dtraml', 'dtrainl', 'dplight', 'dstation', 'dparkcar', 'dparktw', 'dpedcro', 'dstopsign', 'dtrafficarea', 'droad', 'gvm_dwv', 'speedlimit', 'bicyclecount', 'carcount', 'z_qnr_1', 'z_qnr_2', 'z_qnr_3', 'z_qnr_4', 'z_qnr_6', 'z_qnr_7', 'z_qnr_9', 'z_qnr_11', 'z_qnr_12', 'z_qnr_13', 'z_qnr_14', 'z_qnr_17', 'z_qnr_21', 'z_qnr_22', 'z_qnr_23', 'z_qnr_24', 'z_qnr_25', 'z_qnr_26', 'z_qnr_28', 'z_qnr_31', 'z_qnr_32', 'z_qnr_33', 'z_knr_2', 'z_knr_3', 'z_knr_7', 'z_knr_9', 'z_knr_10', 'z_knr_11', 'trafficarea_1', 'trafficarea_2', 'r_width_1', 'r_width_3', 'r_width_4', 'r_width_5', 'r_width_6', 'speedlimit_2', 'speedlimit_3', 'speedlimit_5', 'speedlimit_6'

Regression analysis

Regression results without feature selection

The focus is mainly on results of logistic regression. (Results of linear regression are just for comparison.).

According to the following result, adjusted R-squared value (linear) or pseudo R-squared value (logistic) were around 0.627. ‘cp’ was always recognized as significant variables with a negative coefficient. ‘mew’ (Average entropy value for whole scene) was always significant with a positive coefficient. Other significant numeric variables included ‘dpedcro’, ‘droad’, ‘bicyclecount’.

	original dataset of 60											
	Linear						Logistic					
	without splitting			with splitting			without splitting			with splitting		
	feature	sig	coef	feature	sig	coef	feature	sig	coef	feature	sig	coef
1	const	***	0.432	const	***	0.428	cp	*	-0.376	cp	**	-0.541
2	cp	***	-0.034	cp	*	-0.027	mew	***	3.529	mew	***	3.632
3	mew	***	0.266	mew	***	0.267	dpedcro	**	-1.526	dpedcro	*	-1.655
4	dpedcro	*	-0.073	dpedcro	*	-0.103	droad	***	2.832	droad	***	2.878
5	dtrafficarea	*	-0.050	droad	***	0.268	speedlimit	*	-1.008	speedlimit	*	-1.092
6	droad	***	0.258	bicyclecount	***	0.298	bicyclecount	***	6.966	bicyclecount	***	6.991
7	bicyclecount	***	0.303	z_qnr_7	*	-0.179	z_qnr_12	*	1.509	z_knr_10	*	-1.968
8	z_qnr_6	*	0.117	z_qnr_11	**	-0.222	z_knr_10	**	-2.183	r_width_1	*	1.312
9	z_qnr_7	**	-0.189	z_qnr_13	***	0.362						
10	z_qnr_11	**	-0.203	z_qnr_22	*	-0.255						
11	z_qnr_13	***	0.220	z_qnr_26	*	-0.209						
12	z_qnr_17	*	0.126	z_qnr_32	*	-0.172						
13	z_qnr_22	**	-0.223	z_knr_10	***	-0.226						
14	z_qnr_26	**	-0.251									
15	z_qnr_32	*	-0.160									
16	z_knr_10	***	-0.257									
17	z_knr_11	*	-0.104									
rsqu_adj or pseudo rsqu	0.566			0.555			0.693			0.695		

Fig 1 Significant variables and their coefficients - Linear & Logistic regression with 60 variables

Regression results with feature selection

The following parameters were set for feature selection in this report:

a. Direction: forward / backward
b. Estimator: linear regression / logistic regression
c. Metrics:
   i. For linear regression: Neg_mean_squared_error, R2 score, Neg_median_absolute_error, Neg_mean_absolute_error
   ii. For logistic regression: Accuracy, F1, Recall, Roc_auc
d. K-fold cross-validation: k = 5

According to the following linear regression result, after applying both SFFS (sequential forward floating feature selection) and SBFS (sequential backward floating feature selection), average number of selected features was 21, and average adjusted R-squared value was 0.490. ‘cp’ was recognized as significant in half of the models with negative coefficients around -0.036, while ‘mew’ was recognized as significant in 6/8 models with positive coefficients around 0.270. Other numeric variables which were frequently recognized as significant included ‘dtraml’, ‘droad’, ‘bicyclecount’.

	sffs											
	nmse_5			r2_5			nmedae_5			nmeae_5		
	feature	sig	coef	feature	sig	coef	feature	sig	coef	feature	sig	coef
1	const	***	0.446	const	**	0.245	const	***	0.309	const	***	0.438
2	cp	**	-0.037	mew	***	0.276	droad	***	0.328	cp	**	-0.035
3	mew	***	0.255	dtraml	*	-0.071	z_qnr_6	***	0.339	mew	***	0.278
4	dtraml	*	-0.062	dpedcro	*	-0.100	trafficarea_1	***	0.300	dtraml	**	-0.088
5	dtrainl	*	-0.078	droad	***	0.238				dstation	*	-0.110
6	dpedcro	*	-0.109	speedlimit	*	-0.079				dtrafficarea	*	-0.046
7	droad	***	0.257	bicyclecount	***	0.278				droad	***	0.251
8	bicyclecount	***	0.262	z_qnr_1	**	0.222				bicyclecount	***	0.263
9	z_qnr_11	**	-0.249	z_qnr_4	*	0.213				z_qnr_6	***	0.225
10	z_qnr_13	**	0.208	z_qnr_6	***	0.321				z_qnr_11	*	-0.200
11	z_knr_2	*	-0.110	z_qnr_11	*	-0.163				z_qnr_13	***	0.264
12	z_knr_10	***	-0.186	z_qnr_13	***	0.280				z_knr_10	***	-0.151
13				z_qnr_17	**	0.164						
14				z_qnr_21	*	0.144						
15				z_qnr_24	**	0.191						
16				z_knr_3	***	0.212						
17				z_knr_7	**	0.113						
number of selected features	20			33			7			22		
rsqu_adj or pseudo rsqu	0.530			0.539			0.357			0.532		

Fig 2 Significant variables and their coefficients - Linear regression SFFS

	sbfs											
	nmse_5			r2_5			nmedae_5			nmeae_5		
	feature	sig	coef	feature	sig	coef	feature	sig	coef	feature	sig	coef
1	const	***	0.446	const	**	0.245	const	***	0.344	const	***	0.438
2	cp	**	-0.037	mew	***	0.276	droad	***	0.313	cp	**	-0.035
3	mew	***	0.255	dtraml	*	-0.071	z_qnr_6	***	0.350	mew	***	0.278
4	dtraml	*	-0.062	dpedcro	*	-0.100	z_qnr_9	*	0.172	dtraml	**	-0.088
5	dtrainl	*	-0.078	droad	***	0.238	trafficarea_1	*	0.239	dstation	*	-0.110
6	dpedcro	*	-0.109	speedlimit	*	-0.079				dtrafficarea	*	-0.046
7	droad	***	0.257	bicyclecount	***	0.278				droad	***	0.251
8	bicyclecount	***	0.262	z_qnr_1	**	0.222				bicyclecount	***	0.263
9	z_qnr_11	**	-0.249	z_qnr_4	*	0.213				z_qnr_6	***	0.225
10	z_qnr_13	**	0.208	z_qnr_6	***	0.321				z_qnr_11	*	-0.200
11	z_knr_2	*	-0.110	z_qnr_11	*	-0.163				z_qnr_13	***	0.264
12	z_knr_10	***	-0.186	z_qnr_13	***	0.280				z_knr_10	***	-0.151
13				z_qnr_17	**	0.164						
14				z_qnr_21	*	0.144						
15				z_qnr_24	**	0.191						
16				z_knr_3	***	0.212						
17				z_knr_7	**	0.113						
number of selected features	20			33			10			22		
rsqu_adj or pseudo rsqu	0.530			0.539			0.359			0.532		

Fig 3 Significant variables and their coefficients - Linear regression SBFS

According to the following logistic regression result, after applying both SFFS and SBFS, average number of selected features was 21, and average adjusted R-squared value was 0.574. ‘cp’ was recognized as significant in 5/8 models with negative coefficients around -0.528, while ‘mew’ was recognized as significant in 7/8 models with positive coefficients around 3.083. Other numeric variables which were frequently recognized as significant included ‘droad’, ‘dpedcro’, ‘bicyclecount’, ‘speedlimit’.

	sffs											
	acc_5			f1_5			pre_5			rec_5		
	feature	sig	coef	feature	sig	coef	feature	sig	coef	feature	sig	coef
1	const	*	-1.067	const	**	-2.519	const	***	-1.115	mew	***	2.066
2	mew	***	2.891	cp	**	-0.525	carcount	*	0.525	dstation	***	-2.021
3	dstation	***	-1.836	mew	***	3.400	z_qnr_6	***	1.669	droad	***	2.437
4	droad	***	2.464	dpedcro	**	-1.806				speedlimit	*	-0.741
5	bicyclecount	***	6.692	droad	***	2.578				bicyclecount	***	6.172
6	z_qnr_12	**	1.961	speedlimit	*	-0.992				z_knr_10	***	-1.934
7	z_qnr_13	*	2.024	bicyclecount	***	6.746				trafficarea_2	***	-1.343
8	z_knr_3	*	1.291	z_qnr_12	**	1.676						
9	z_knr_10	**	-1.862	z_qnr_13	*	1.989						
10	trafficarea_2	***	-1.394	z_knr_10	**	-2.044						
number of selected features	23			36			2			21		
rsqu_adj or pseudo rsqu	0.636			0.666			0.029			0.611		

Fig 4 Significant variables and their coefficients - Logistic regression SFFS

	sbfs											
	acc_5			f1_5			pre_5			rec_5		
	feature	sig	coef	feature	sig	coef	feature	sig	coef	feature	sig	coef
1	const	***	-2.599	const	**	-1.989	const	***	-2.835	const	***	-3.187
2	cp	**	-0.525	cp	**	-0.516	cp	***	-0.551	cp	**	-0.522
3	mew	***	3.314	mew	***	3.261	mew	***	3.200	mew	***	3.451
4	dpedcro	**	-1.574	dtraml	*	-0.780	dpedcro	**	-1.617	dparktw	*	-1.300
5	droad	***	2.726	dpedcro	**	-1.557	droad	***	2.761	dpedcro	**	-1.543
6	speedlimit	**	-1.214	droad	***	2.739	qvm_dwv	*	0.804	droad	***	2.593
7	bicyclecount	***	7.089	speedlimit	*	-1.075	speedlimit	**	-1.105	speedlimit	**	-1.316
8	z_knr_10	***	-2.175	bicyclecount	***	7.095	bicyclecount	***	7.076	bicyclecount	***	7.543
9				z_qnr_12	**	1.778	z_knr_10	***	-2.124			
10				z_qnr_13	*	1.980						
11				z_knr_10	***	-2.147						
number of selected features	24			26			18			22		
rsqu_adj or pseudo rsqu	0.677			0.688			0.678			0.666		

Fig 5 Significant variables and their coefficients - Logistic regression SBFS

Conclusion of regression results

Based on all the regression result, average adjusted R-squared value or pseudo R-squared value is 0.564. ‘cp’ was recognized as significant variables in 13/20 models, while ‘mew’ was recognized as significant variables in 17/20 models.

Random forest classification

Random forest classification was applied to four different variable sets. For each classification, evaluation metrics (accuracy, precision, sensitivity recall, f1 score, mcc score and kappa), confusion matrix, feature importances (Gini importance, mean decrease in accuracy, permutation feature importance) were calculated.

Whole variable set (91)

For comparison, random forest classification was applied with all 91 variables.

```
Train data accuracy: 0.9697368421052631
Test data accuracy: 0.9418960244648318
accuracy 0.9418960244648318
precision [0.944      0.93506494]
sensitivity recall [0.97925311 0.8372093 ]
f1 score [0.96130346 0.88343558]
mcc score 0.8471856226714769
kappa 0.8448960231669871
```

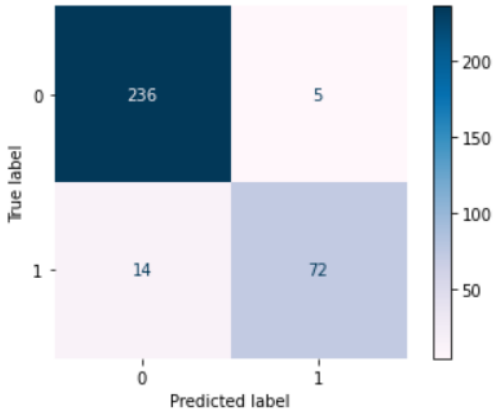

Fig 6 Confusion matrix – all 91 variables

91					
Gini importance		Mean decrease accuracy		Permutation importance	
droad	0.2151	bicyclecount	0.0826	bicyclecount	0.0969
bicyclecount	0.1978	droad_sl	0.0612	droad_sl	0.0777
droad_sl	0.1654	droad	0.0581	droad	0.0755
mew	0.0482	dbusl	0.0092	**mewmax**	0.0083
dparktw	0.0365	**mew**	0.0092	**mewmin**	0.0061
mewmin	0.0322	r_width	-0.0061	z_knr_5	0.0052
dtrainl	0.0190	z_knr_6	-0.0061	dbusl	0.0052
dtraml	0.0189	**mewmax**	0.0061	**mew**	0.0049
z_knr_1	0.0167	z_knr_5	0.0061	**cmean**	0.0043
dparkcar	0.0164	dpedcro	0.0061	dtrainl	0.0043
carcount	0.0154	z_qnr_23	0.0031	dpedcro	0.0043
r_width	0.0134	dplight	0.0031	dparktw	0.0034
dstation	0.0131	r_width_5	0.0031	z_qnr_21	0.0031
mewmax	0.0118	**cmean**	0.0031	**cmax**	0.0031
dbusl	0.0101	dtrainl	0.0031	z_qnr_23	0.0028
dpedcro	0.0092	speedlimit	0.0031	gvm_msp	0.0028
cmean	0.0091	**mewmin**	0.0031	r_width_5	0.0028
iemin	0.0088	gvm_msp	0.0031	speedlimit	0.0028
z_qnr_11	0.0084	**megmax**	-0.0031	dparkcar	0.0024
z_knr_5	0.0084	**ie**	-0.0031	dstopsign	0.0018

Fig 7 Feature importance – features with top 20 highest importances – all 91 variables

Filtered variable set (60; filtered with correlation and VIF)

```
Train data accuracy: 0.9986842105263158
Test data accuracy: 0.9510703363914373
accuracy 0.9510703363914373
precision [0.94820717 0.96052632]
sensitivity recall [0.98755187 0.84883721]
f1 score [0.96747967 0.90123457]
mcc score 0.871811196351556
kappa 0.8688787529447146
```

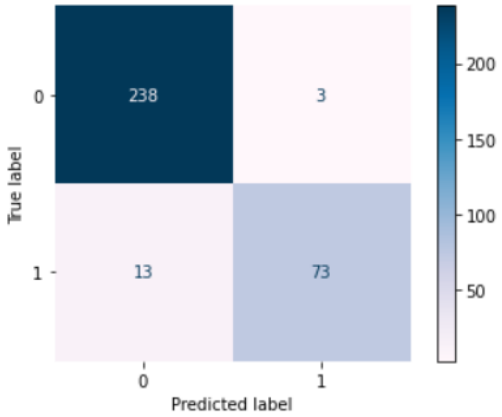

Fig 8 Confusion matrix – 60 variables

60					
Gini importance		Mean decrease accuracy		Permutation importance	
bicyclecount	0.2072	bicyclecount	0.1132	bicyclecount	0.1321
droad	0.1870	droad	0.0642	droad	0.1122
mew	0.0674	**mew**	0.0122	**mew**	0.0220
dparktw	0.0581	dstation	0.0122	dparktw	0.0193
dbusl	0.0373	dparktw	0.0061	dparkcar	0.0135
dtraml	0.0367	gvm_dwv	0.0061	dstation	0.0110
dtrafficarea	0.0364	dbusl	0.0061	**meg**	0.0080
ie	0.0289	dtrafficarea	0.0061	gvm_dwv	0.0067
dstopsign	0.0276	dstopsign	0.0061	dtraml	0.0067
dstation	0.0264	speedlimit_3	0.0031	dbusl	0.0064
meg	0.0255	**meg**	0.0031	dtrafficarea	0.0043
carcount	0.0252	**megmax**	0.0031	z_qnr_12	0.0034
dparkcar	0.0248	dpedcro	-0.0031	speedlimit	0.0034
dpedcro	0.0237	dparkcar	0.0031	dstopsign	0.0034
gvm_dwv	0.0231	z_qnr_12	0.0031	dtrainl	0.0031
dplight	0.0229	speedlimit	0.0031	**cp**	0.0024
megmax	0.0217	**cp**	0.0031	**megmax**	0.0021
dtrainl	0.0175	z_qnr_4	0.0000	dplight	0.0018
cp	0.0086	**ie**	0.0000	speedlimit_3	0.0012
z_qnr_6	0.0084	dtraml	0.0000	dpedcro	0.0009

Fig 9 Feature importance – features with top 20 highest importances – 60 variables (filtered variable set)

Filtered variable set + ‘cmean’ (61)

```
Train data accuracy: 0.9973684210526316
Test data accuracy: 0.9541284403669725
accuracy 0.9541284403669725
precision [0.9484127  0.97333333]
sensitivity recall [0.99170124 0.84883721]
f1 score [0.96957404 0.9068323 ]
mcc score 0.8802062170940408
kappa 0.8765944599592422
```

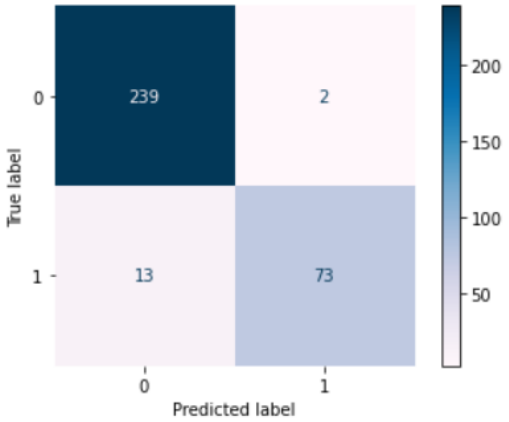

Fig 10 Confusion matrix – 61 variables (filtered set + ‘cmean’)

61					
Gini importance		Mean decrease accuracy		Permutation importance	
bicyclecount	0.2408	bicyclecount	0.1254	bicyclecount	0.1477
droad	0.2041	droad	0.0581	droad	0.1398
mew	0.0534	**cp**	0.0092	**meg**	0.0086
dtraml	0.0393	**cmean**	0.0092	**mew**	0.0080
carcount	0.0360	**meg**	0.0092	dtrainl	0.0076
dparktw	0.0349	dtraml	0.0061	**cmean**	0.0073
dtrafficarea	0.0282	gvm_dwv	0.0061	speedlimit	0.0073
ie	0.0280	dstopsign	0.0031	dtrafficarea	0.0064
dtrainl	0.0276	z_qnr_13	0.0031	dpedcro	0.0061
dpedcro	0.0269	**megmax**	0.0031	dtraml	0.0055
dstation	0.0260	**mew**	0.0031	dparktw	0.0049
dbusl	0.0256	trafficarea_1	0.0031	dbusl	0.0046
meg	0.0251	z_knr_11	0.0031	dstation	0.0040
gvm_dwv	0.0242	dpedcro	0.0031	dplight	0.0040
dparkcar	0.0193	dparktw	0.0031	r_width_4	0.0037
dplight	0.0182	r_width_5	0.0031	r_width_5	0.0037
dstopsign	0.0170	dplight	0.0031	trafficarea_1	0.0037
megmax	0.0153	r_width_4	0.0031	gvm_dwv	0.0034
speedlimit	0.0126	trafficarea_2	-0.0031	z_qnr_13	0.0034
cmean	0.0126	dparkcar	-0.0031	dstopsign	0.0024

Fig 11 Feature importance – features with top 20 highest importances – 61 variables (filtered variable set + ‘cmean’)

Filtered set + curb related variables [‘csum’, ‘cmax’, ‘cmean’] (63)

```
Train data accuracy: 0.9986842105263158
Test data accuracy: 0.9480122324159022
accuracy 0.9480122324159022
precision [0.94094488 0.97260274]
sensitivity recall [0.99170124 0.8255814 ]
f1 score [0.96565657 0.89308176]
mcc score 0.864075582404673
kappa 0.8590410021046225
```

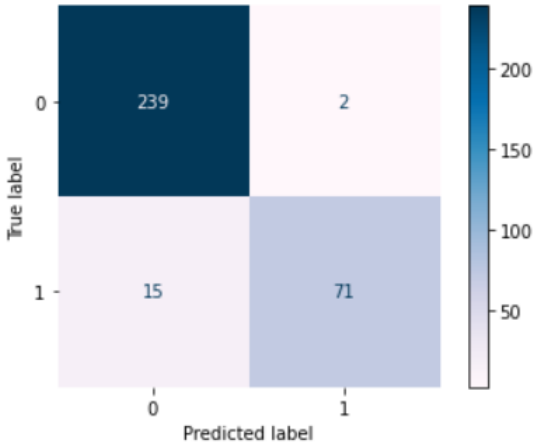

Fig 12 Confusion matrix – 63 variables (filtered set + curb-related variables)

63					
Gini importance		Mean decrease accuracy		Permutation importance	
bicyclecount	0.2146	bicyclecount	0.1162	bicyclecount	0.1333
droad	0.1875	droad	0.0795	droad	0.1174
mew	0.0660	**mew**	0.0153	**mew**	0.0214
dtraml	0.0491	dparktw	0.0122	dparktw	0.0187
dparktw	0.0398	gvm_dwv	0.0122	gvm_dwv	0.0092
dstopsign	0.0335	**meg**	0.0092	dstation	0.0083
carcount	0.0309	dtraml	0.0092	**megmax**	0.0070
dpedcro	0.0281	r_width_5	0.0092	dtraml	0.0064
dplight	0.0271	dparkcar	0.0092	dbusl	0.0064
dbusl	0.0268	**ie**	-0.0092	**meg**	0.0064
dstation	0.0263	**cmax**	0.0061	trafficarea_1	0.0061
ie	0.0255	trafficarea_1	0.0061	trafficarea_2	0.0058
gvm_dwv	0.0249	**megmax**	0.0061	z_knr_11	0.0046
dtrafficarea	0.0244	trafficarea_2	0.0061	dparkcar	0.0043
dparkcar	0.0220	dtrainl	-0.0061	z_qnr_13	0.0031
meg	0.0182	dstation	-0.0061	**csum**	0.0024
megmax	0.0173	dstopsign	-0.0061	z_qnr_6	0.0024
dtrainl	0.0167	**cmean**	-0.0031	**cp**	0.0021
cmean	0.0154	speedlimit	-0.0031	**cmax**	0.0006
csum	0.0133	dbusl	-0.0031	z_qnr_2	0.0006

Fig 13 Feature importance – features with top 20 highest importances – 61 variables (filtered variable set + curb-related variables: ‘cmean’, ‘cmax’, ‘csum’)

Conclusion of random forest classification results

Based on all the results listed above, the performances of random forest classification were with accuracy values higher than 0.949, average precision values around 0.953, average sensitivity recall around 0.914, average f1 score around 0.931, mcc scores around 0.866, kappa values around 0.862.

According to the rank of importances, three variables were always among the top 20 important

features, including ‘bicyclecount’, ‘droad’, ‘mew’. Many traffic-transport numeric variables were often with high importances. As for curb-related variables, for each classification model, at least one of them would be among the top 20 important variables.

PCA

For four different variable set, PCA was performed to explore the feature importance. The following chart is a summary of the most important features in each PC. For PCA result of all four variable sets, only around 17% of the total variance is explained by the first principal component. And ‘cp’ was always found the most important feature in the first component.

	80%											
	91			60			61			63		
	PC	Feature	explained variance ratio	PC	Feature	explained variance ratio	PC	Feature	explained variance ratio	PC	Feature	explained variance ratio
0	PC1	cp	18.47%	PC1	cp	15.42%	PC1	cp	19.36%	PC1	cp	14.66%
1	PC2	droad_sl	15.27%	PC2	dtrafficarea	15.10%	PC2	dtrafficarea	14.32%	PC2	cp	14.37%
2	PC3	droad_sl	7.05%	PC3	bicyclecount	7.15%	PC3	bicyclecount	6.76%	PC3	bicyclecount	6.87%
3	PC4	iemax	5.50%	PC4	droad	6.49%	PC4	droad	6.13%	PC4	droad	6.17%
4	PC5	iemax	5.08%	PC5	bicyclecount	4.33%	PC5	bicyclecount	4.09%	PC5	iemax	5.46%
5	PC6	dtraml	3.73%	PC6	bicyclecount	3.69%	PC6	bicyclecount	3.49%	PC6	bicyclecount	3.96%
6	PC7	r_width	2.94%	PC7	r_width_4	3.36%	PC7	r_width_4	3.17%	PC7	bicyclecount	3.53%
7	PC8	bicyclecount	2.67%	PC8	megmax	3.04%	PC8	megmax	2.88%	PC8	r_width_5	3.08%
8	PC9	speedlimit_4	2.50%	PC9	ie	2.92%	PC9	ie	2.76%	PC9	megmin	2.84%
9	PC10	dtraml	2.35%	PC10	speedlimit	2.36%	PC10	speedlimit	2.23%	PC10	speedlimit	2.39%
10	PC11	dbusl	2.23%	PC11	dtraml	2.28%	PC11	dtraml	2.16%	PC11	dtraml	2.18%
11	PC12	r_width_2	2.05%	PC12	carcount	2.16%	PC12	carcount	2.04%	PC12	carcount	2.06%
12	PC13	z_knr_11	1.68%	PC13	carcount	2.06%	PC13	carcount	1.94%	PC13	dtrainl	1.95%
13	PC14	carcount	1.52%	PC14	r_width_5	1.87%	PC14	r_width_5	1.77%	PC14	r_width_5	1.77%
14	PC15	dtraml	1.52%	PC15	dtrainl	1.74%	PC15	dtrainl	1.64%	PC15	mewmin	1.64%
15	PC16	dtrafficarea	1.40%	PC16	z_knr_7	1.61%	PC16	z_knr_7	1.52%	PC16	z_knr_7	1.55%
16	PC17	carcount	1.36%	PC17	z_knr_2	1.56%	PC17	z_knr_2	1.48%	PC17	z_knr_7	1.50%
17	PC18	z_knr_5	1.23%	PC18	z_knr_7	1.49%	PC18	z_knr_7	1.41%	PC18	r_width_3	1.49%
18	PC19	dtraml	1.19%	PC19	speedlimit_3	1.42%	PC19	speedlimit_3	1.34%	PC19	speedlimit_3	1.34%
19	PC20	dtrainl	1.14%							PC20	dparktw	1.27%

Fig 14 PCA analysis result of four different variable sets

Other results and plots of PCA analysis are in Appendix.

Next steps (?)

1. …
2. Writing the paper!
3. (For severity of accidents, avoid overlapping of locations with ‘severe person injury’/‘light person injury’)

Appendix

Categorical variables

z_qnr	z_qnr_1	12	25
	z_qnr_2	34	34
	z_qnr_3	51	22
	z_qnr_4	14	28
	z_qnr_5	82	14
	z_qnr_6	42	55
	z_qnr_7	72	32
	z_qnr_8	92	87
	z_qnr_9	44	32
	z_qnr_10	11	18
	z_qnr_11	81	29
	z_qnr_12	24	58
	z_qnr_13	13	31
	z_qnr_14	122	16
	z_qnr_15	61	50
	z_qnr_16	52	31
	z_qnr_17	119	58
	z_qnr_18	41	16
	z_qnr_19	115	56
	z_qnr_20	123	18
	z_qnr_21	111	44
	z_qnr_22	83	15
	z_qnr_23	102	30
	z_qnr_24	21	42
	z_qnr_25	91	29
	z_qnr_26	74	23
	z_qnr_27	33	27
	z_qnr_28	63	23
	z_qnr_29	101	60
	z_qnr_30	73	19
	z_qnr_31	31	27
	z_qnr_32	71	20
	z_qnr_33	121	10
	(z_qnr_34)	23	8
z_knr	z_knr_1	1	102
	z_knr_2	3	88
	z_knr_3	5	53
	z_knr_4	8	58
	z_knr_5	4	103
	z_knr_6	7	94
	z_knr_7	9	116
	z_knr_8	2	108
	z_knr_9	12	44
	z_knr_10	6	73
	z_knr_11	11	158
	(z_knr_12)	10	90
trafficarea	trafficarea_1	T0	46
	trafficarea_2	T30	961
	(trafficarea_3)	T20	80
r_width	r_width_1	8	96
	r_width_2	4	342
	r_width_3	3	86
	r_width_4	6	204
	r_width_5	10	264
	r_width_6	2	59
	(r_width_7)	1	36
speedlimit	speedlimit_1	50	439
	speedlimit_2	20	31
	speedlimit_3	30	417
	speedlimit_4	0	170
	speedlimit_5	60	17
	speedlimit_6	80	7
	(speedlimit_7)	100	6

Fig 15 Correspondence of categorical variables and their one-hot encoding variables

Feature importance in Random Forest classification

91 variables

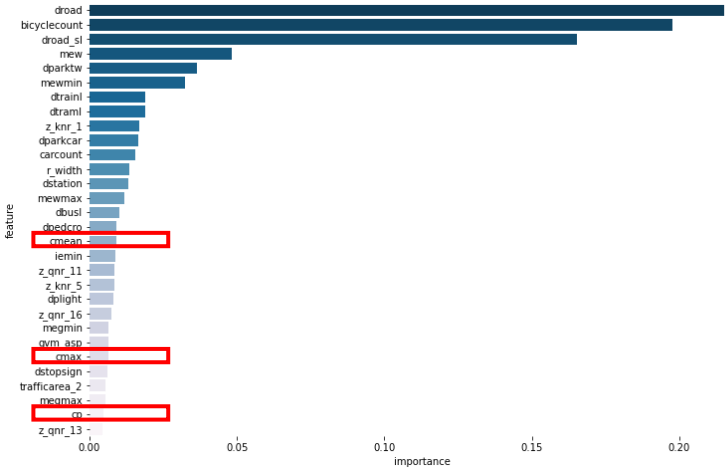

Fig 16 Gini feature importance top 30 – all 91 variables

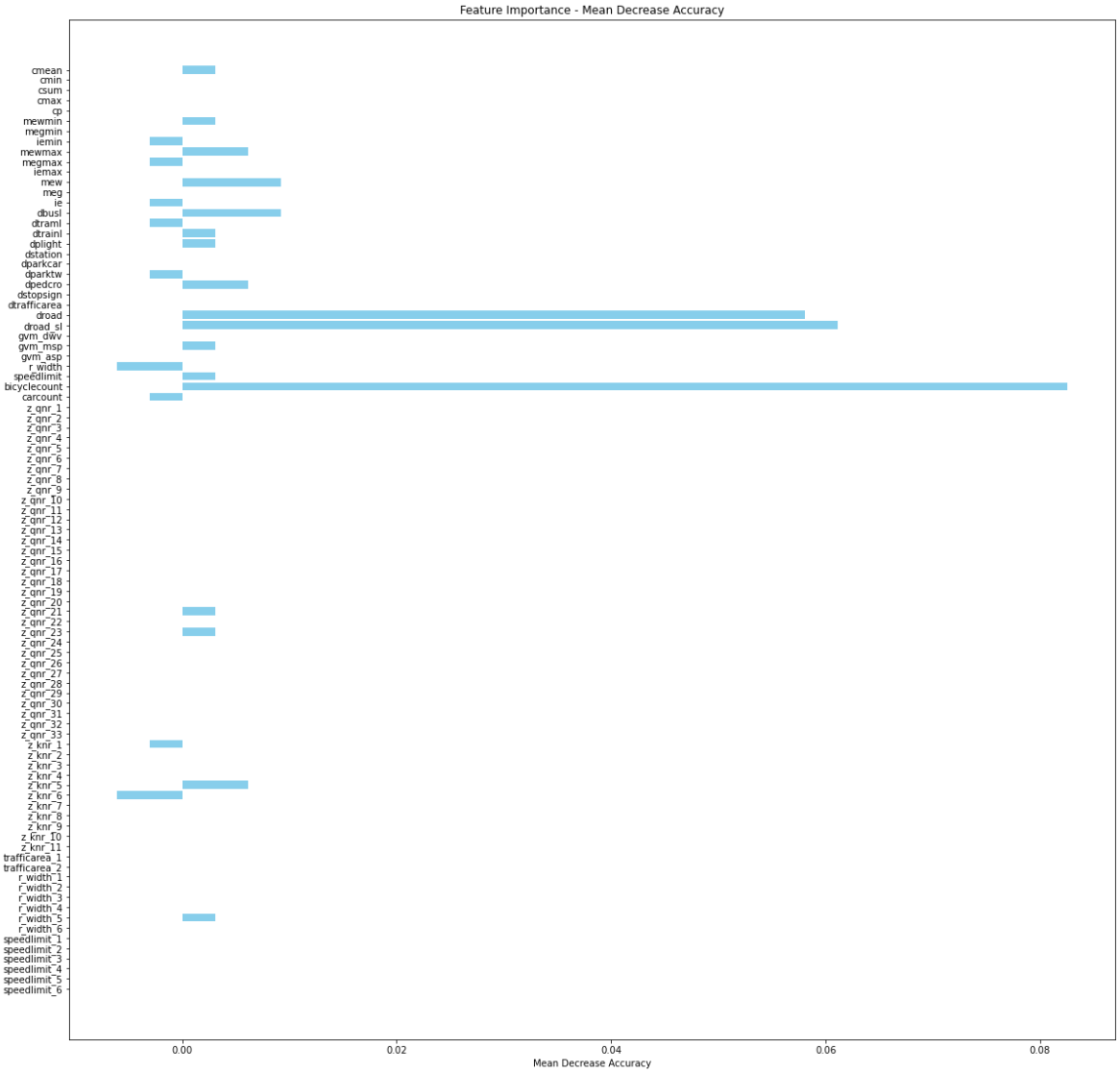

Fig 17 Mean decrease accuracy - all 91 variables

60 variables

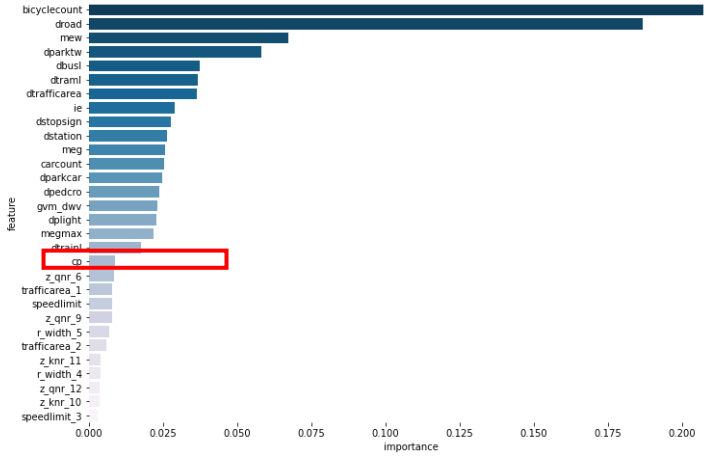

Fig 18 Gini feature importance top 30 – 60 variables

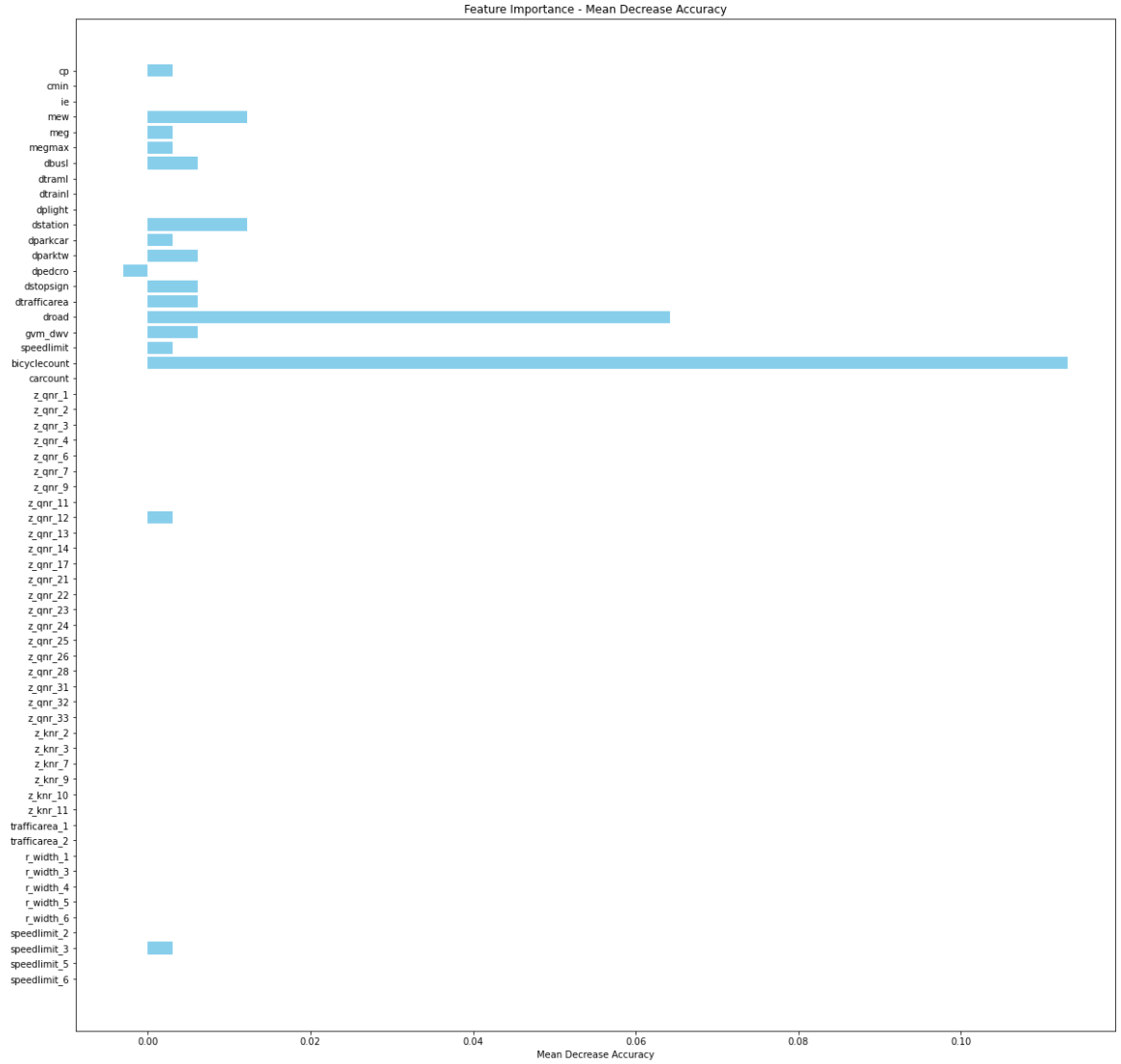

Fig 19 Mean decrease accuracy – 60 variables

61 variables

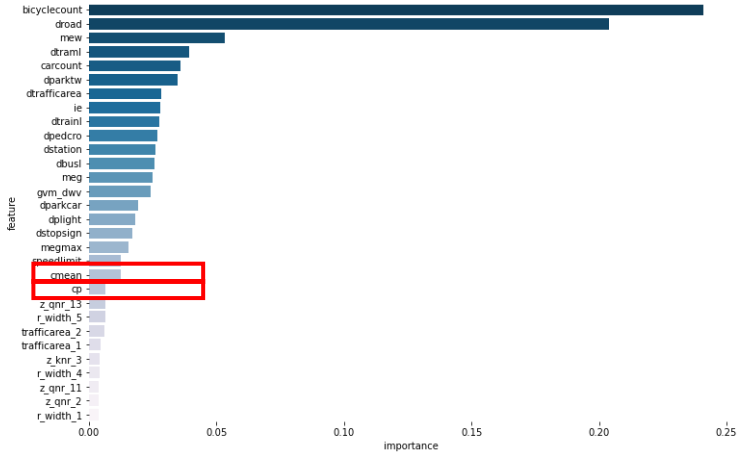

Fig 20 Gini importance top 30 - 61 variables (filtered set + 'cmean')

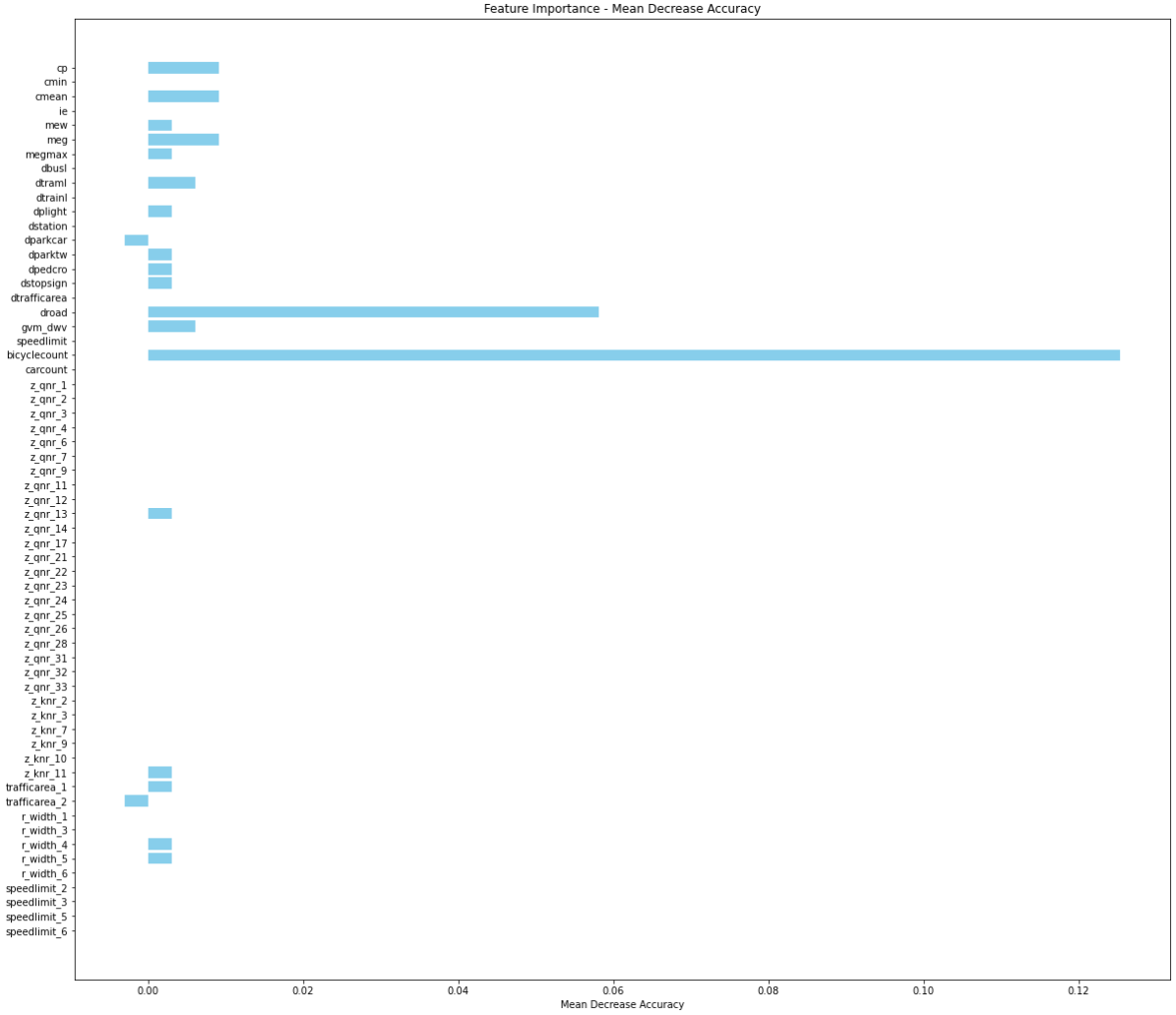

Fig 21 Mean decrease accuracy – 61 variables (filtered set + 'cmean')

63 variables

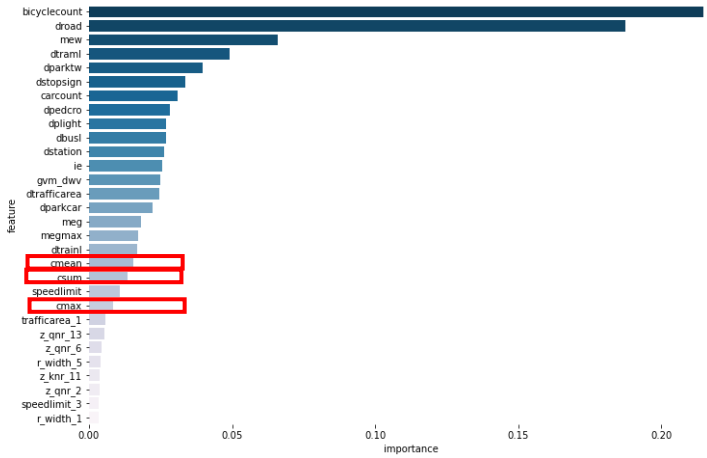

Fig 22 Gini importance top 30 – 63 variables (filtered set + curb-related variables)

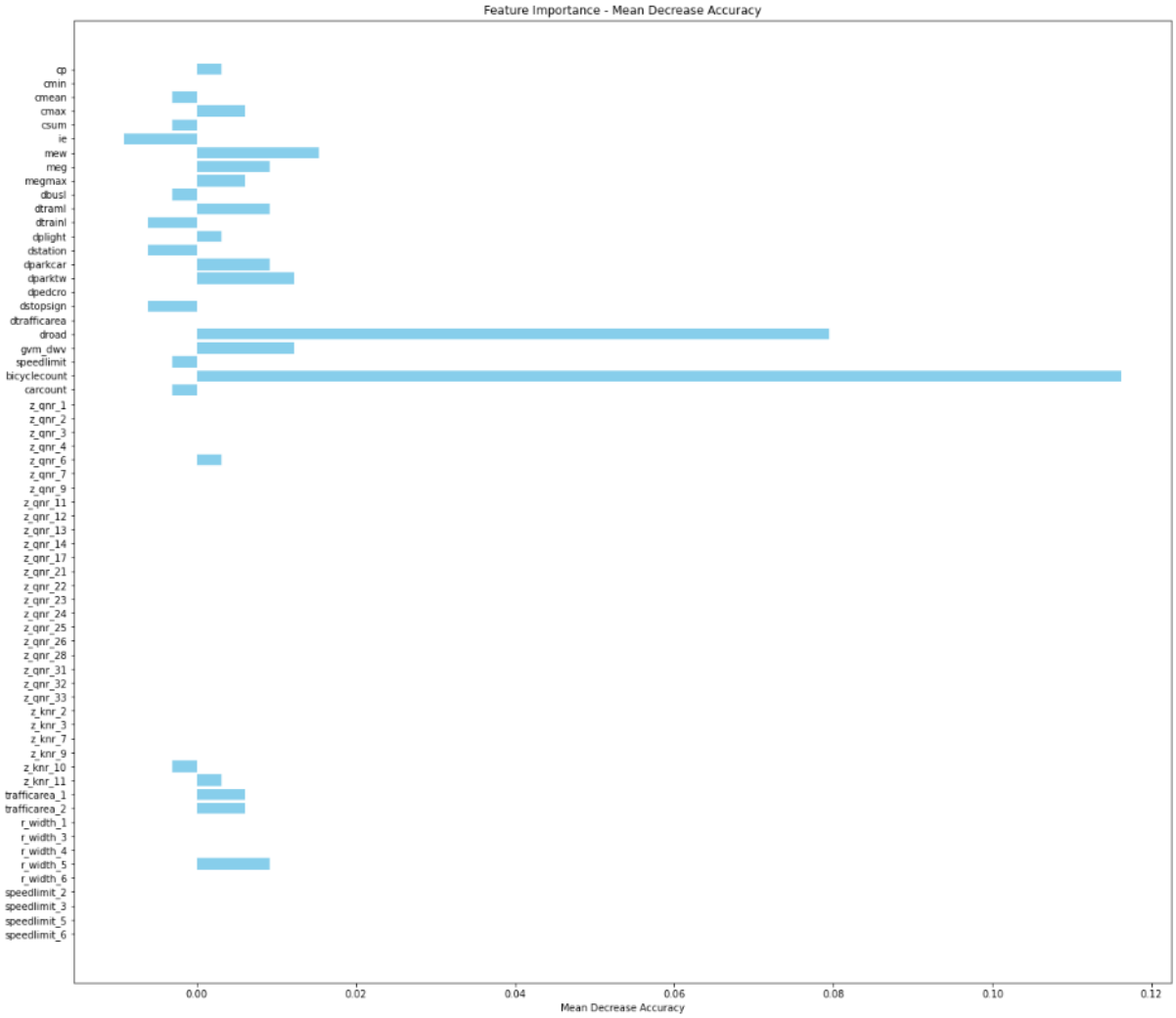

Fig 23 Mean decrease accuracy – 63 variables (filtered set + curb-related variables)

Results of PCA

	95%							
	91		60		61		63	
	PC	Feature	PC	Feature	PC	Feature	PC	Feature
0	PC0	cp	PC0	cp	PC0	cp	PC0	cp
1	PC1	droad_sl	PC1	dtrafficarea	PC1	dtrafficarea	PC1	cp
2	PC2	droad_sl	PC2	bicyclecount	PC2	bicyclecount	PC2	bicyclecount
3	PC3	iemax	PC3	droad	PC3	droad	PC3	droad
4	PC4	iemax	PC4	bicyclecount	PC4	bicyclecount	PC4	iemax
5	PC5	dtraml	PC5	bicyclecount	PC5	bicyclecount	PC5	bicyclecount
6	PC6	r_width	PC6	r_width_4	PC6	r_width_4	PC6	bicyclecount
7	PC7	bicyclecount	PC7	megmax	PC7	megmax	PC7	r_width_5
8	PC8	speedlimit_4	PC8	ie	PC8	ie	PC8	megmin
9	PC9	dtraml	PC9	speedlimit	PC9	speedlimit	PC9	speedlimit
10	PC10	dbusl	PC10	dtraml	PC10	dtraml	PC10	dtraml
11	PC11	r_width_2	PC11	carcount	PC11	carcount	PC11	carcount
12	PC12	z_knr_11	PC12	carcount	PC12	carcount	PC12	dtrainl
13	PC13	carcount	PC13	r_width_5	PC13	r_width_5	PC13	r_width_5
14	PC14	dtraml	PC14	dtrainl	PC14	dtrainl	PC14	mewmin
15	PC15	dtrafficarea	PC15	z_knr_7	PC15	z_knr_7	PC15	z_knr_7
16	PC16	carcount	PC16	z_knr_2	PC16	z_knr_2	PC16	z_knr_7
17	PC17	z_knr_5	PC17	z_knr_7	PC17	z_knr_7	PC17	r_width_3
18	PC18	dtraml	PC18	speedlimit_3	PC18	speedlimit_3	PC18	speedlimit_3
19	PC19	dtrainl	PC19	dstopsign	PC19	dstopsign	PC19	dparktw
20	PC20	z_knr_2	PC20	r_width_3	PC20	r_width_3	PC20	dstopsign
21	PC21	z_knr_5	PC21	trafficarea_2	PC21	trafficarea_2	PC21	dstation
22	PC22	z_knr_2	PC22	mew	PC22	mew	PC22	r_width_3
23	PC23	dparkcar	PC23	dpedcro	PC23	r_width_3	PC23	dpedcro
24	PC24	r_width_1	PC24	mew	PC24	mew	PC24	dparkcar
25	PC25	megmax	PC25	dplight	PC25	dplight	PC25	dplight
26	PC26	carcount	PC26	r_width_6	PC26	cmean	PC26	dplight
27	PC27	dpedcro	PC27	dplight	PC27	z_knr_10	PC27	z_knr_10
28	PC28	megmax	PC28	z_qnr_12	PC28	z_qnr_12	PC28	dplight
29	PC29	dparkcar	PC29	dstation	PC29	cmin	PC29	z_qnr_17
30	PC30	cp	PC30	z_knr_9	PC30	dstation	PC30	z_qnr_19
31	PC31	z_knr_4	PC31	cmin	PC31	z_knr_9	PC31	z_qnr_12
32	PC32	r_width_6	PC32	cmin	PC32	z_qnr_12	PC32	z_qnr_12
33	PC33	z_qnr_17	PC33	z_qnr_17	PC33	z_qnr_17	PC33	iemin
34	PC34	dplight	PC34	z_qnr_6	PC34	cmin	PC34	mewmax
35	PC35	dparktw	PC35	z_qnr_13	PC35	z_qnr_6	PC35	cmin
36	PC36	dplight	PC36	r_width_6	PC36	z_qnr_13	PC36	z_knr_3
37	PC37	dpedcro	PC37	meg	PC37	z_qnr_31	PC37	z_qnr_13
38	PC38	iemax					PC38	z_qnr_13
39	PC39	dstopsign					PC39	z_qnr_2
40	PC40	z_qnr_25					PC40	z_qnr_31
41	PC41	mewmax						
42	PC42	z_qnr_29						

Fig 24 PCA 95% result for four different variable sets

91 variables

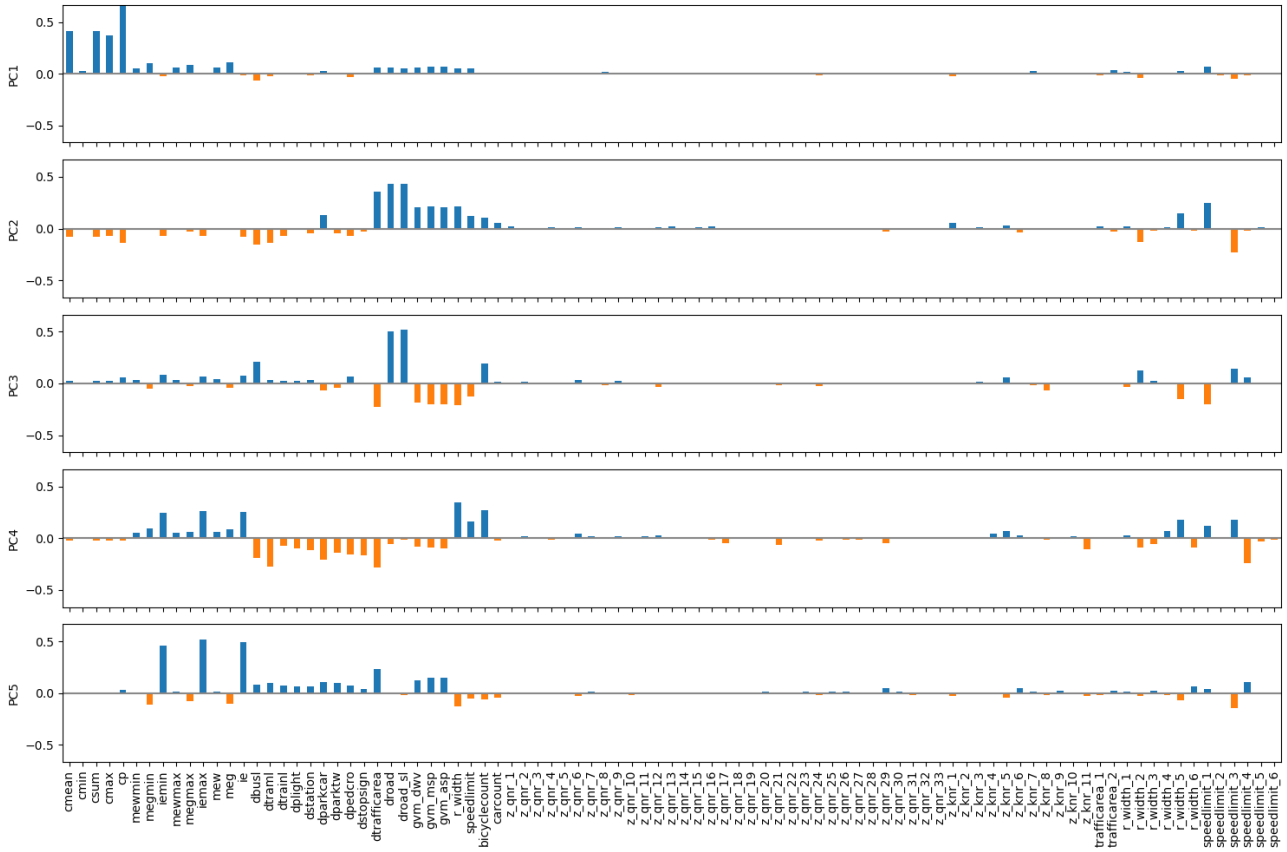

Fig 25 First 5 components – all 91 variables

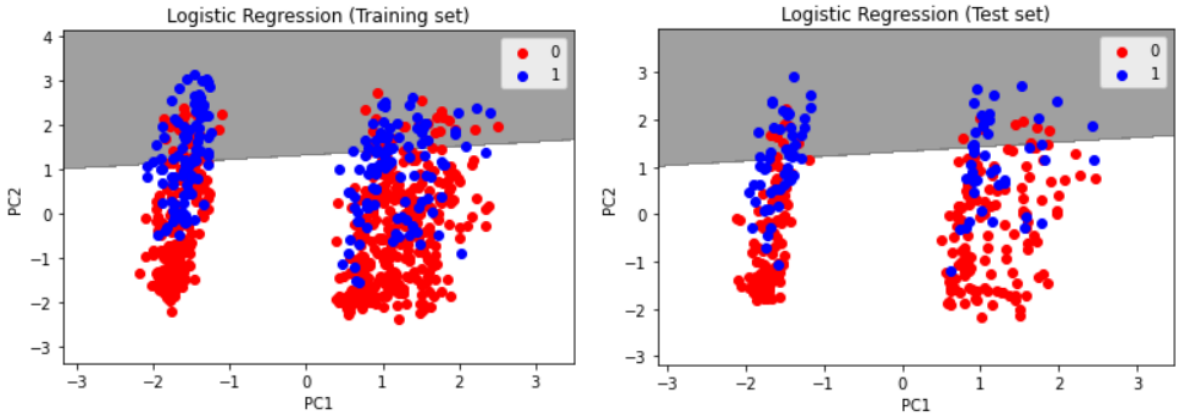

Fig 26 PCA score plot + logistic regression result – all 91 variables

60 variables

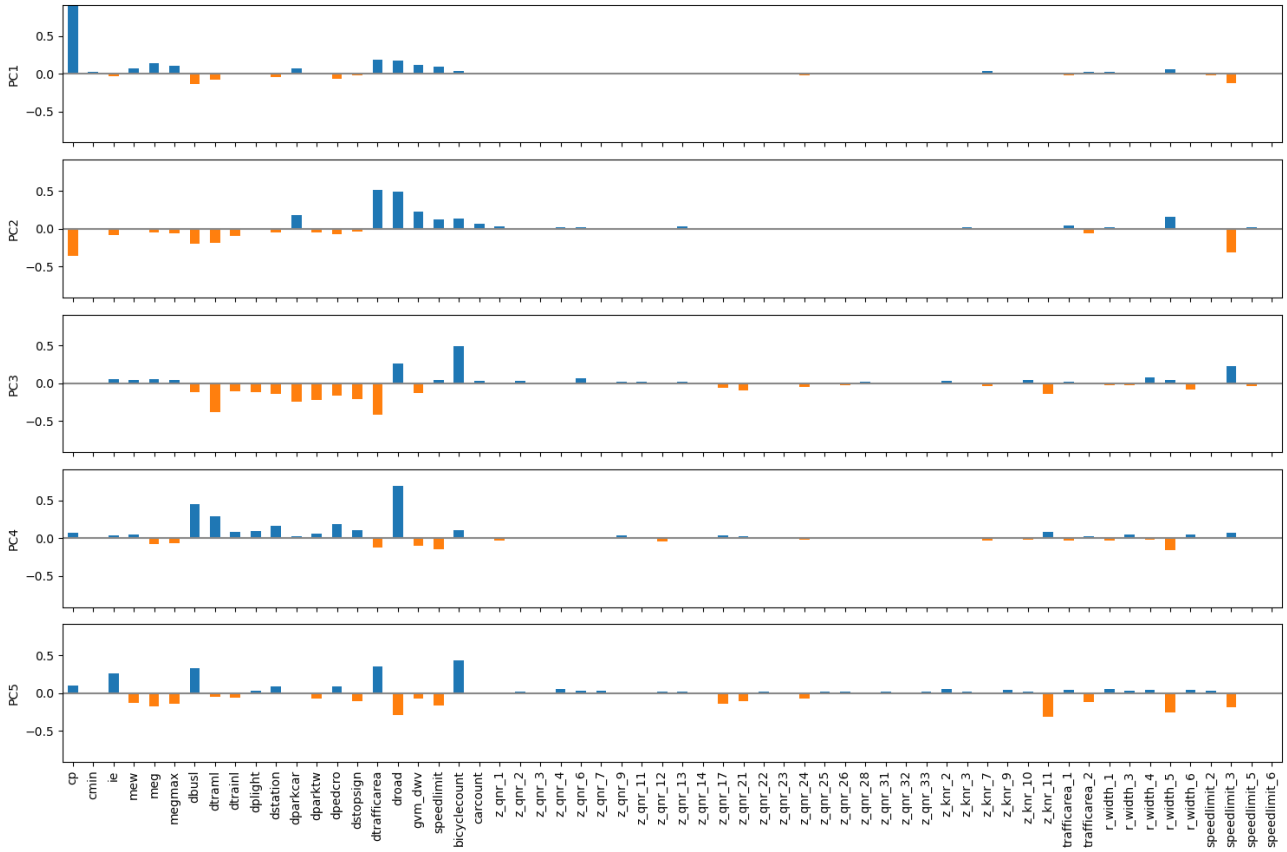

Fig 27 First 5 components – 60 variables

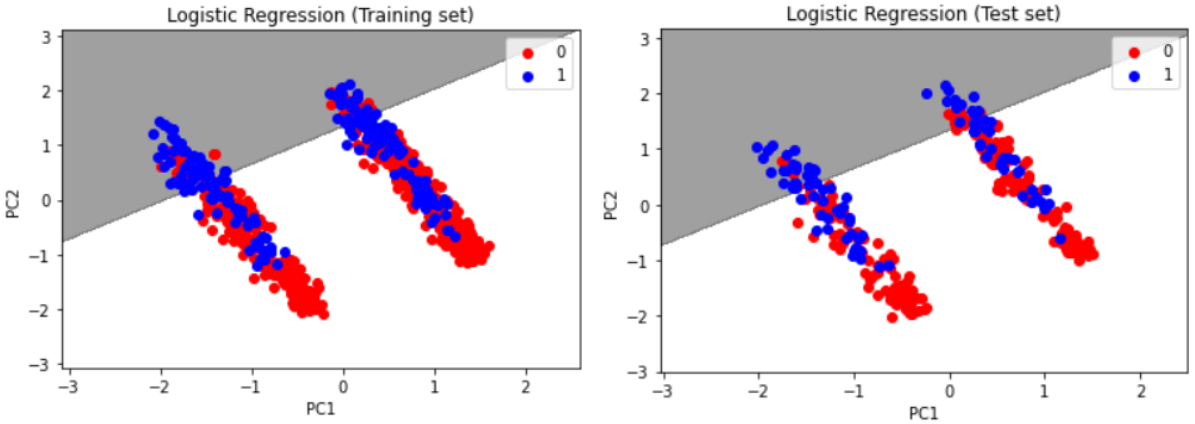

Fig 28 PCA score plot + logistic regression result – 60 variables

61 variables

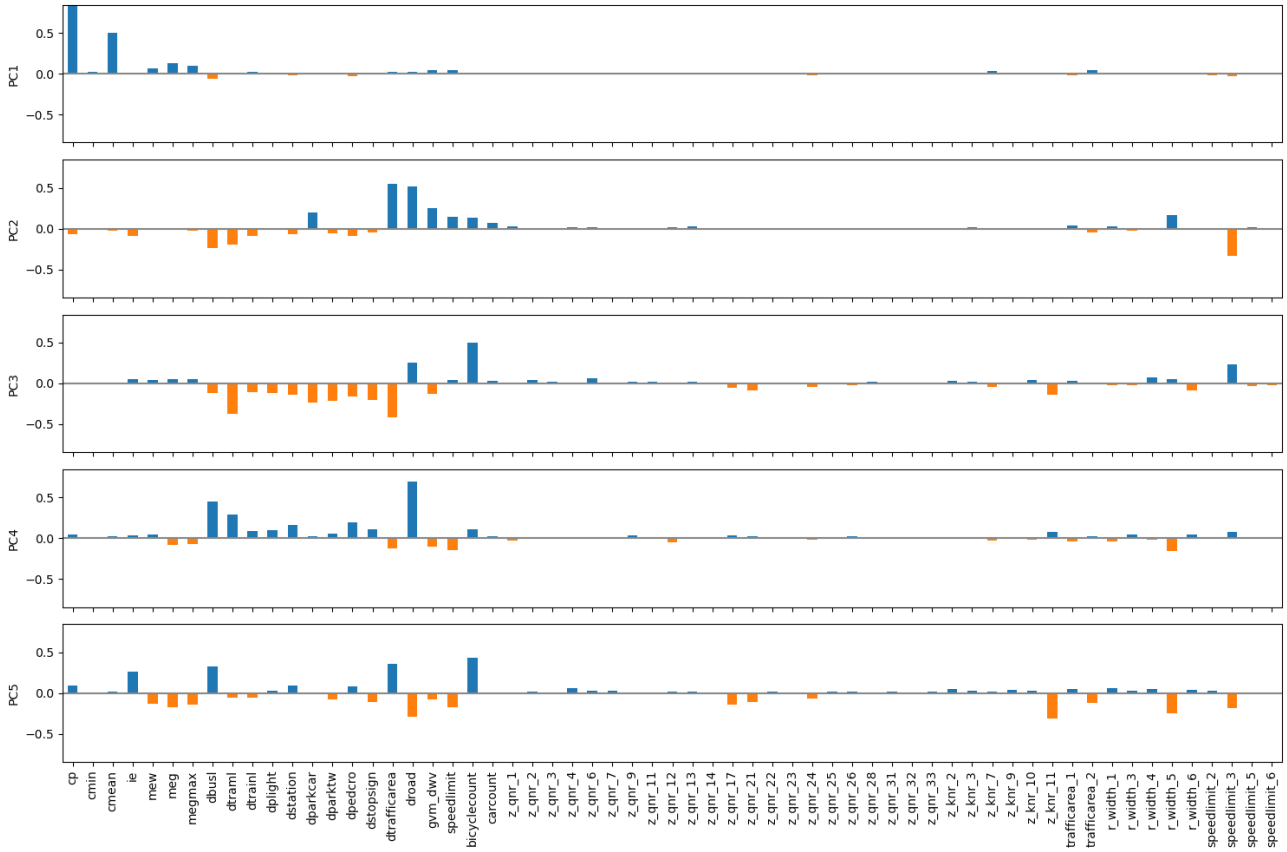

Fig 29 First 5 components – 61 variables (filtered set + 'cmean')

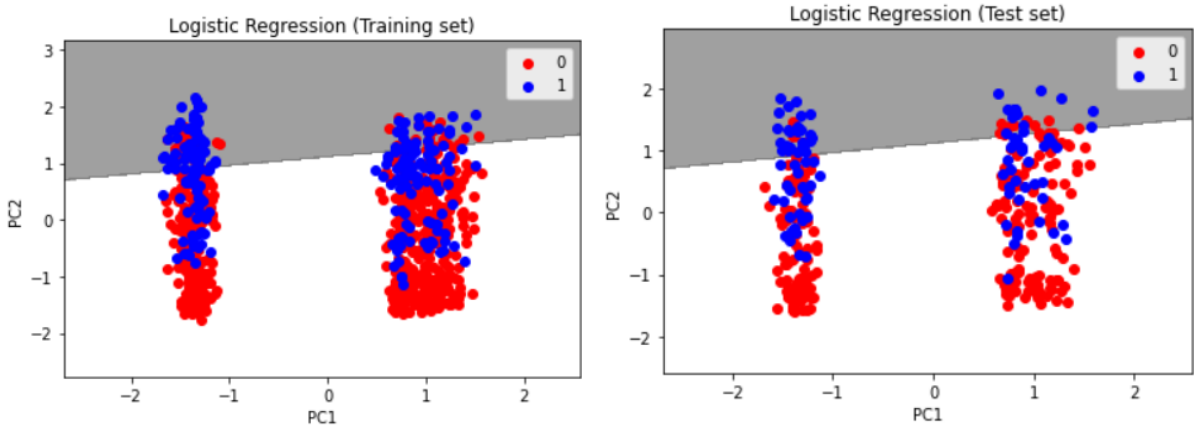

Fig 30 PCA score plot + logistic regression result – 61 variables (filtered set + 'cmean')

63 variables

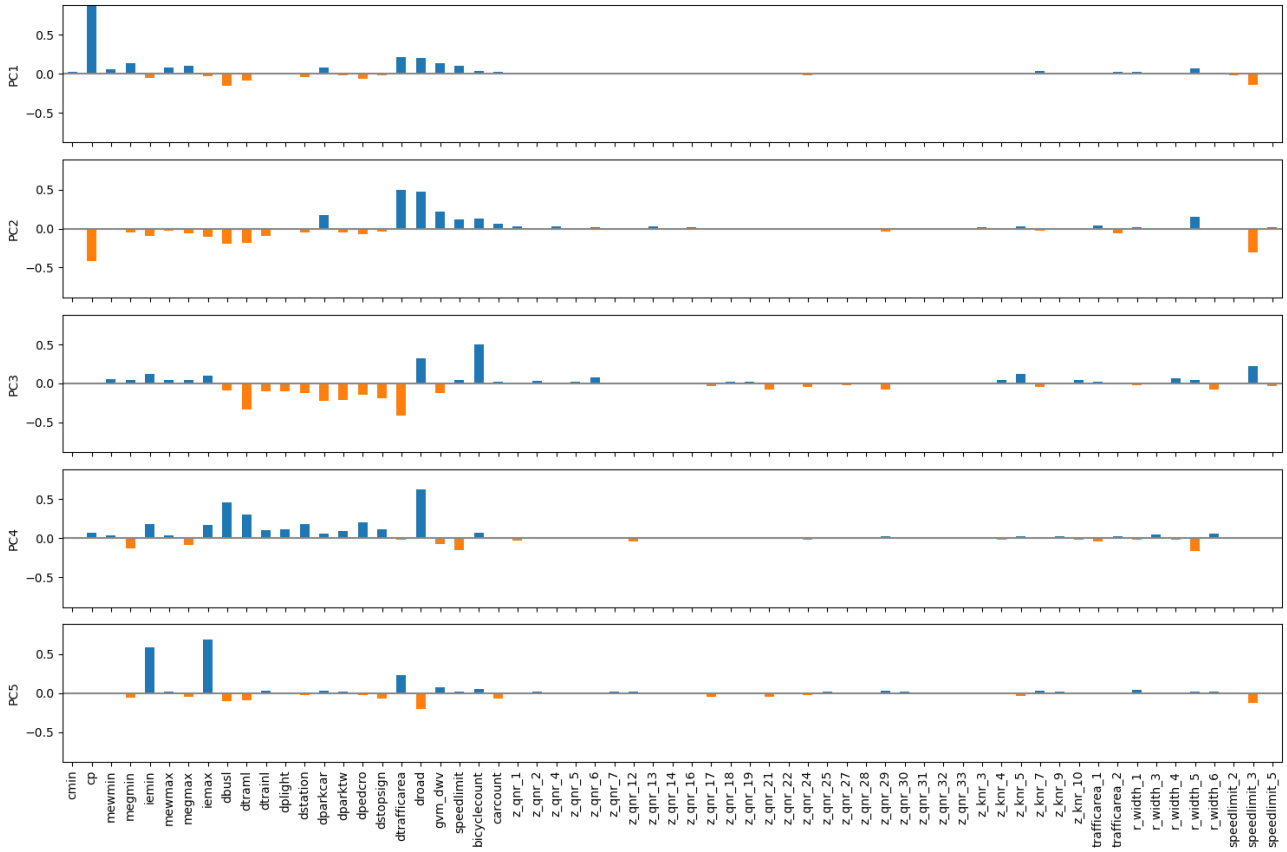

Fig 31 First 5 components – 63 variables (filtered set + curb-related variables)

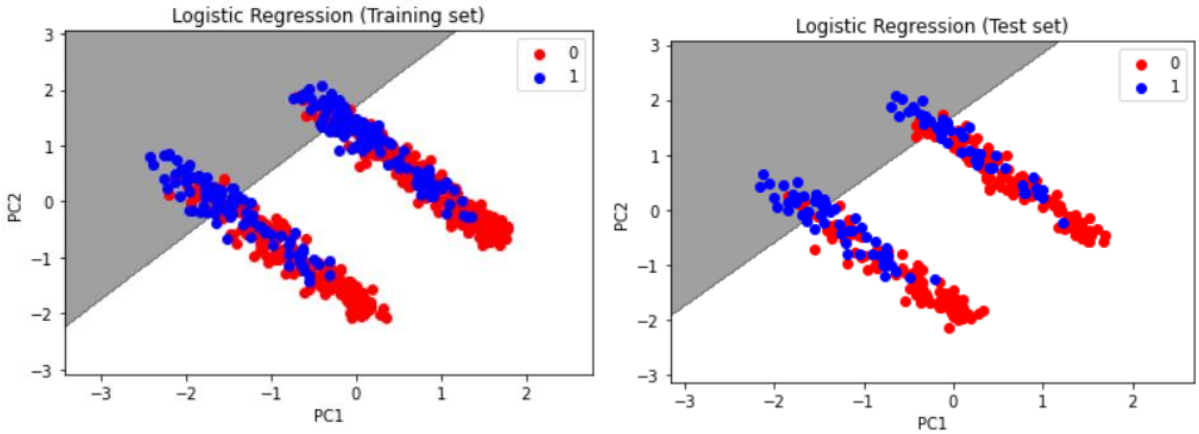

Fig 32 PCA score plot + logistic regression result – 63 variables (filtered set + curb-related variables)