Ozone Level Detection

STA 9891

Yandi (Claire) Chen

Ozone level detection

Number of Instances (N): 1847

- Original dataset: 2356
- Null values: 689

Number of Attributes (P): 73

Sample size (S): 50

Imbalance: (93% : 7%)

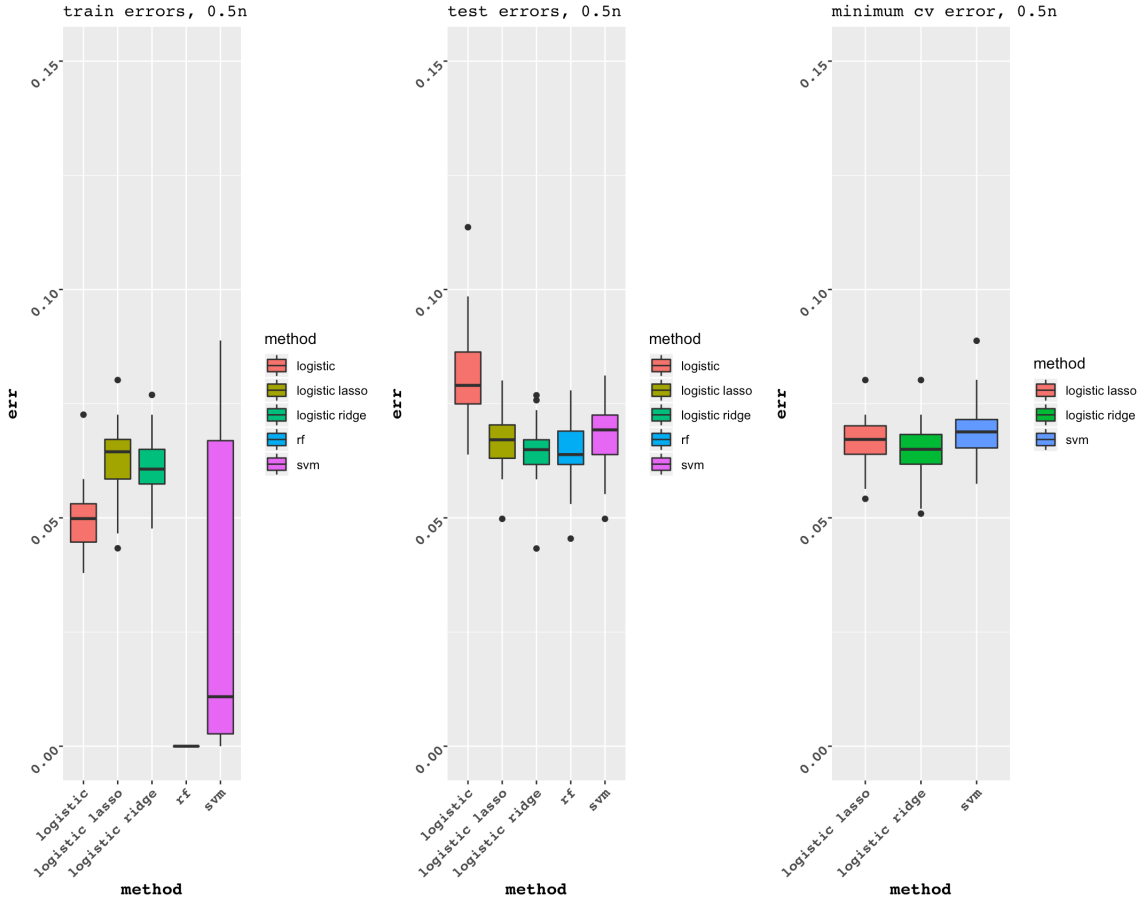
- 0: normal day = 1719
- 1: Ozone day = 128

Motivation:

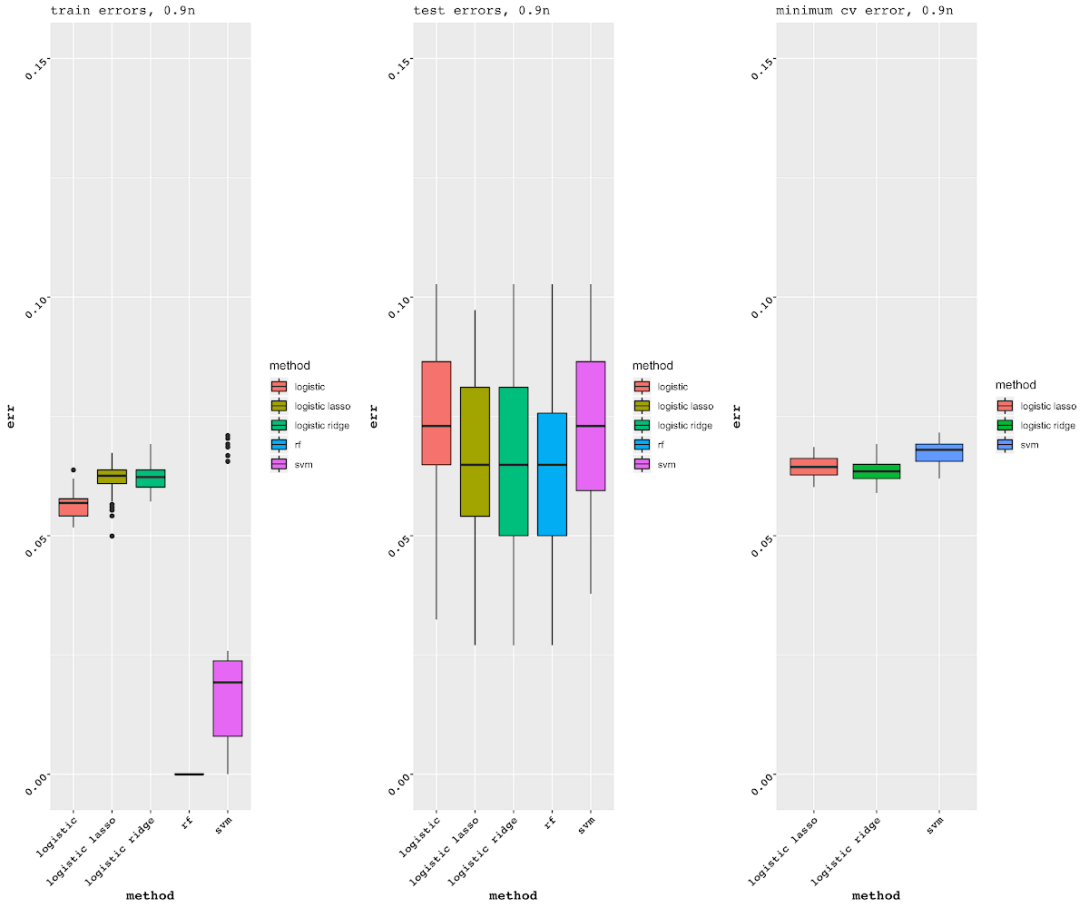
- Determine which attributes are more relevant to detect Ozone level

Model comparison:

Train size = 0.5

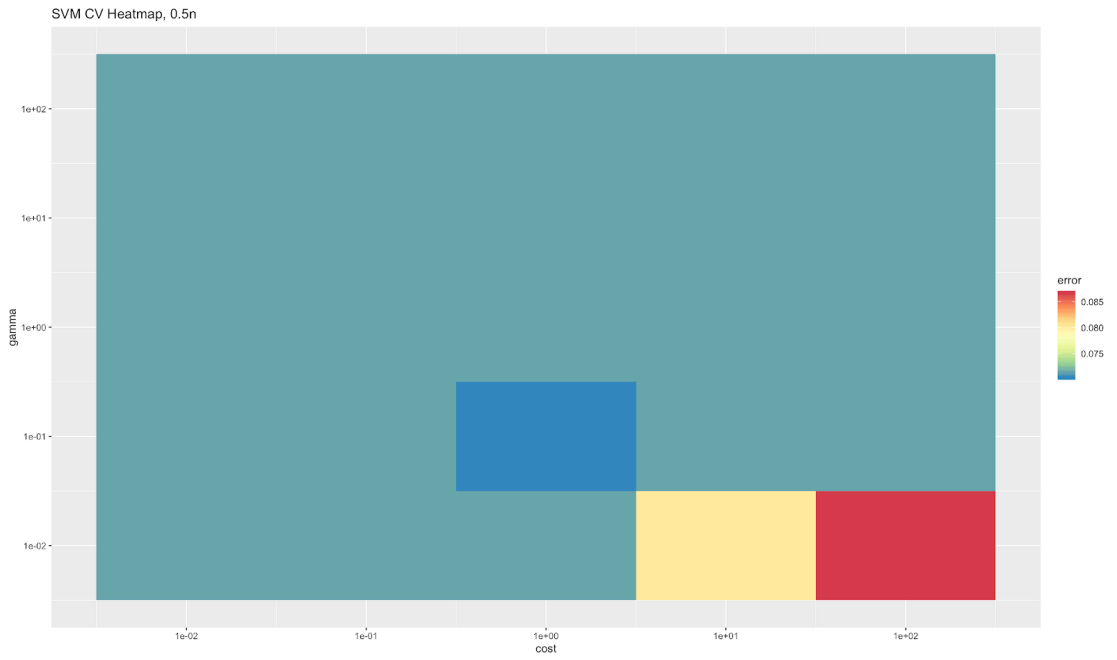


Train size = 0.9

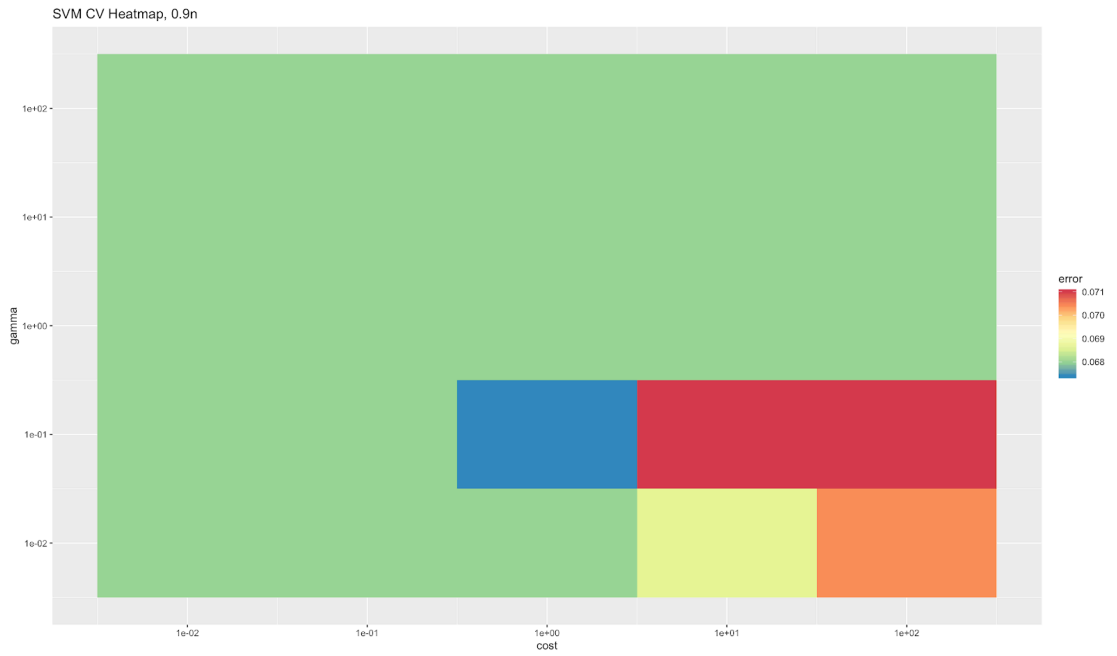


SVM CV Heatmap

Train size = 0.5

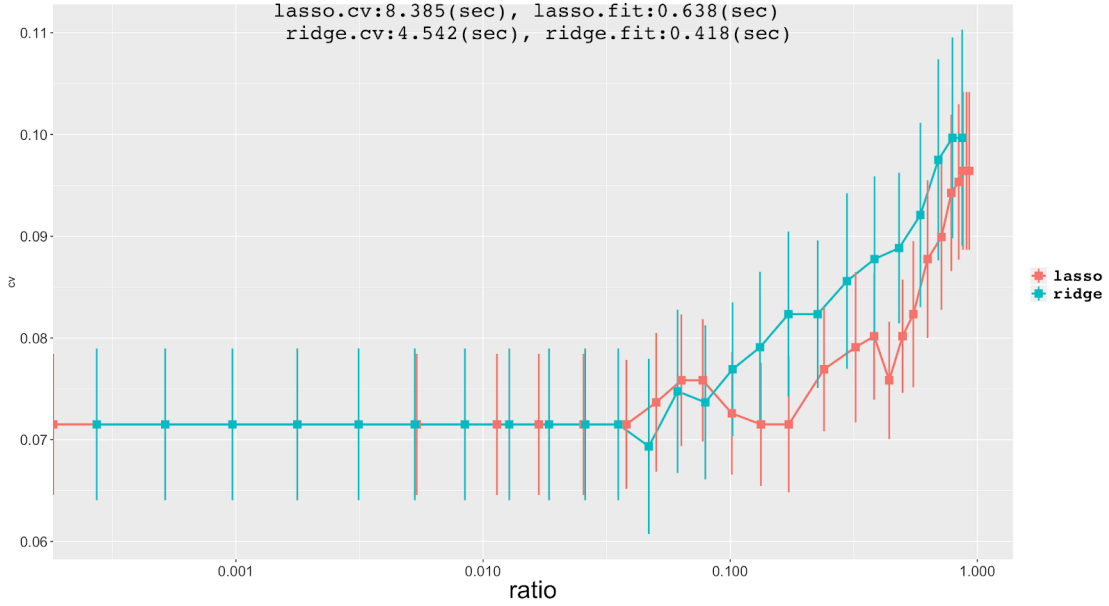


Train size = 0.9

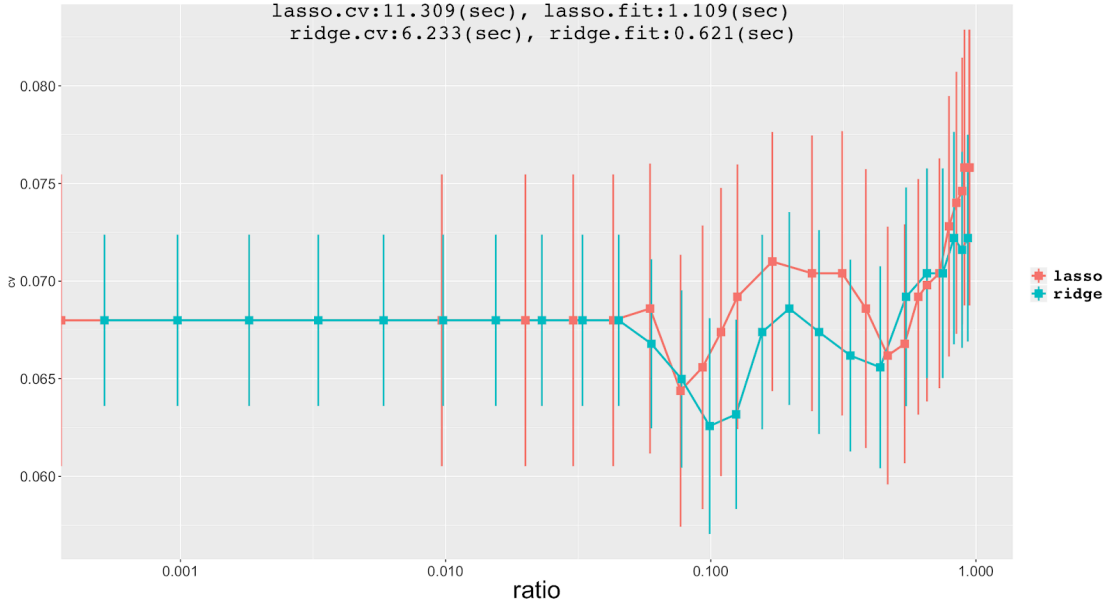


10-fold CV Curves from Lasso and Ridge

m = 25
Train size = 0.5



m = 25
Train size = 0.9



Model Procedure Time

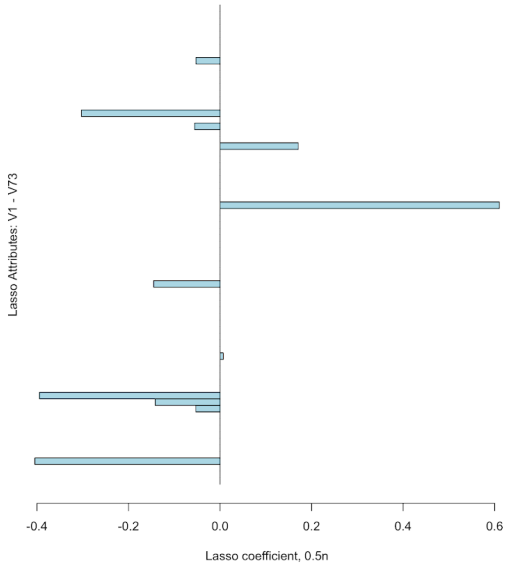
MODEL	Logistic	Lasso	Ridge	Random Forest	SVM
Time (0.5n)	0.13334	13.27668	0.06326	0.86644	71.74018
Time (0.9n)	0.18442	14.12958	3.27736	1.749	210.5241

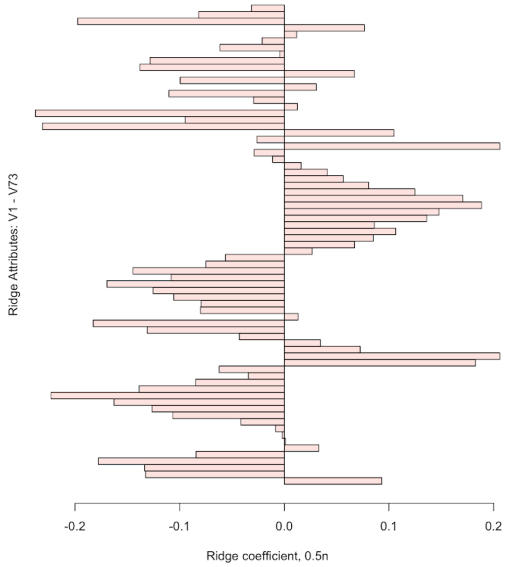
Coefficient Plot

Train size = 0.5

Lasso: V43 > V52 > V20

Ridge: V52 > V20 > V43 > V19





V19: WSR17 – wind speed at 17:00pm

V20: WSR18 – wind speed at 18:00pm

V39: T11 – temperature measured at 11:00am

V42: T14 – temperature measured at 14:00pm

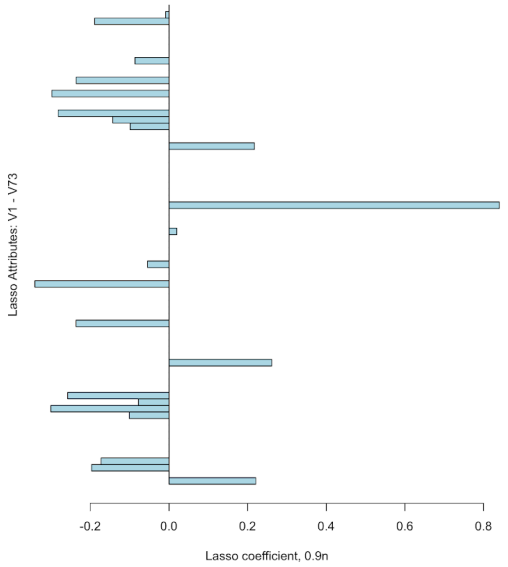
V43: T15 – temperature measured at 15:00pm

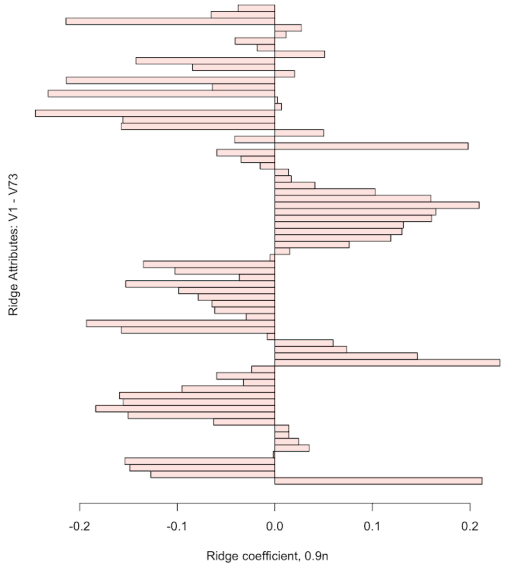
V52: T_PK – peak temperature

Train size = 0.9

Lasso: V43 > V19 > V52 > V39

Ridge: V19> V43> V52 > V42





Summary

- Performance: Ridge > Lasso > Random Forest > Logistic > SVM
- Cost: Logistic < Random Forest < Ridge < Lasso < SVM
- Tune the hyper parameters and tree size in Random Forest to avoid overfitting