Buzz Prediction on Twitter (# of Active Discussions)

Soonmo Seong

Data Structure

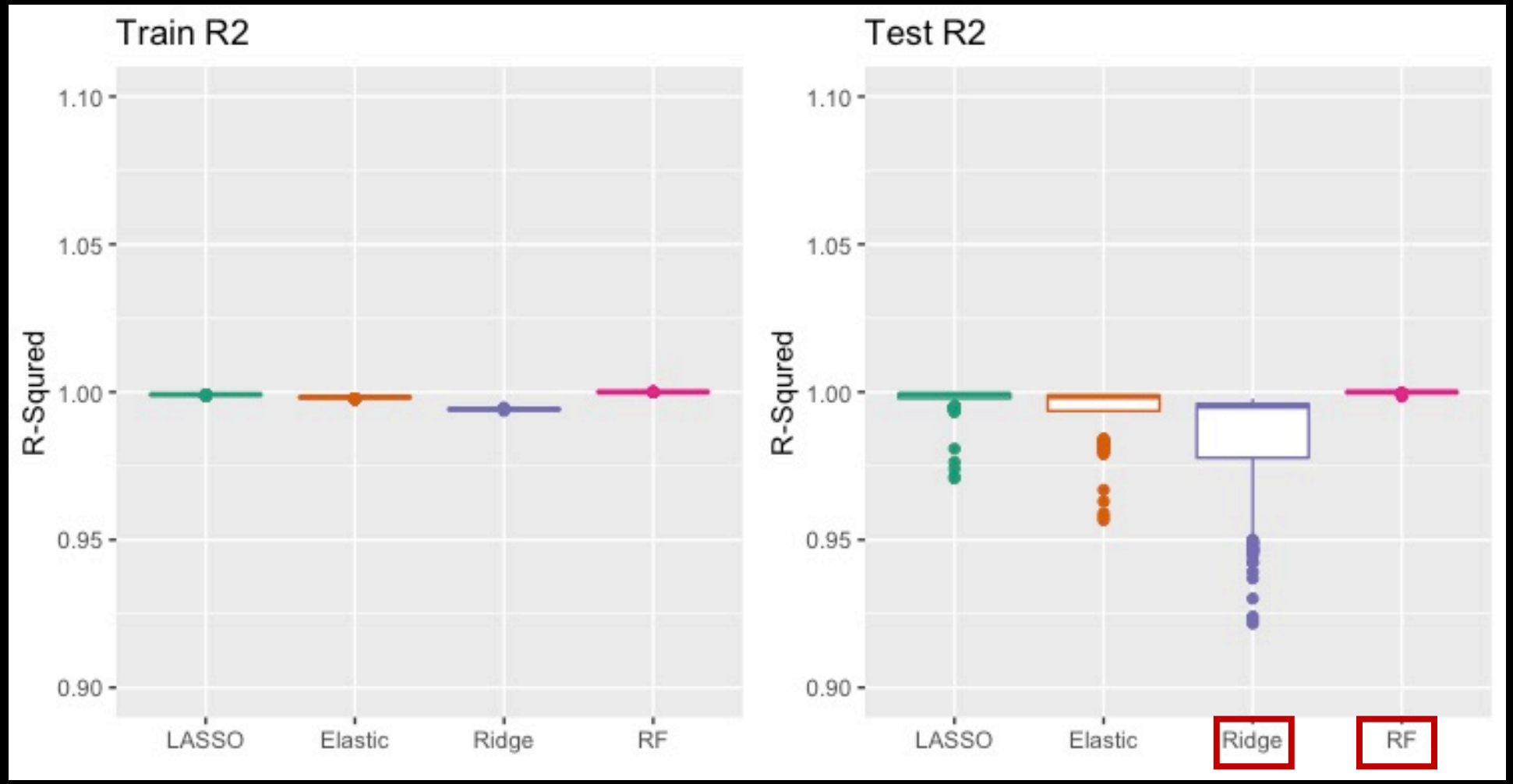
- n = 5832 samples, multiple regression.

- p = 70(except target variable), no missing data
- Predictors have 10 types in different time frames such as Number of Created Discussions, Author Increase, and Attention Level

- The target variable is the number of active discussion(NAD) in 6 weeks after a tweet is created.

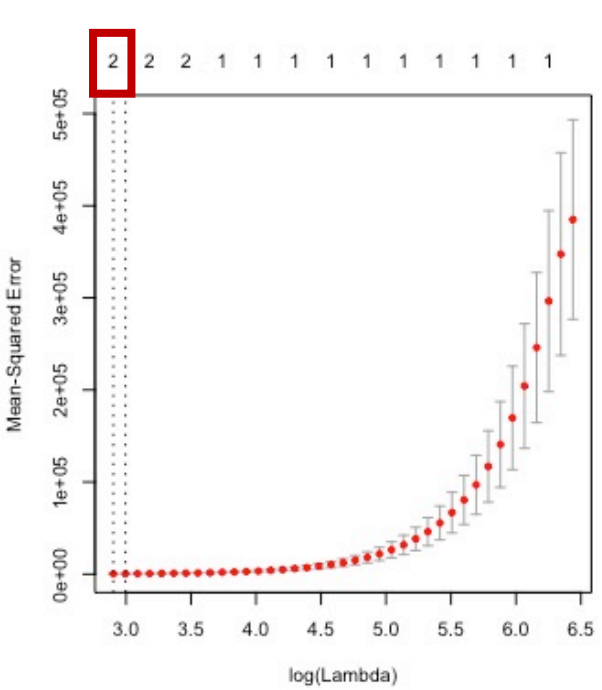
- If the target is well predicted, we can pay attention to emergent issues such as covid-19 and wildfire, act proactively, and minimize possible damages.
- If we set the threshold, this dataset can be used for binary classification: Buzz detection on Twitter

R-Sqaures

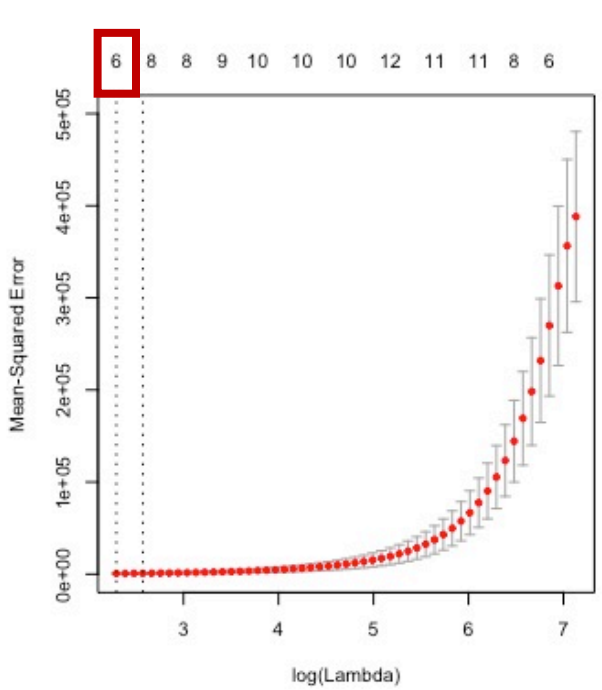

Train R2
Test R2
R-Squred
1.10
1.05
1.00
0.95
0.90
LASSO
Elastic
Ridge
RF

CV Curves

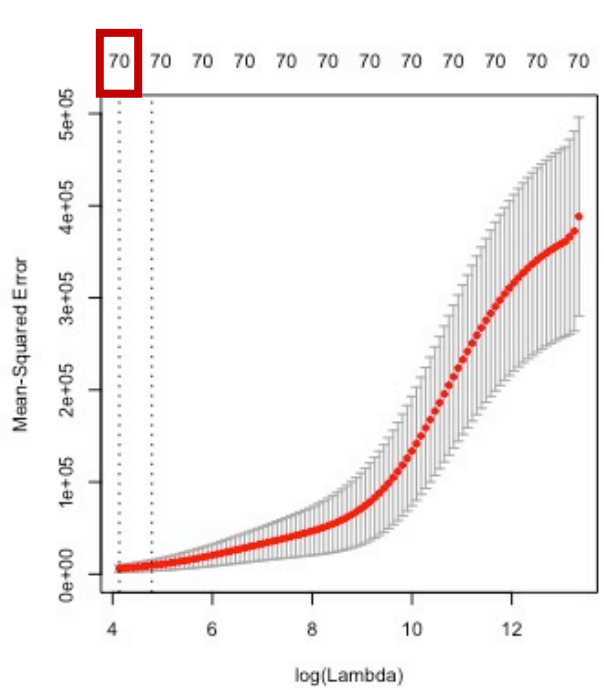
LASSO



Elastic Net

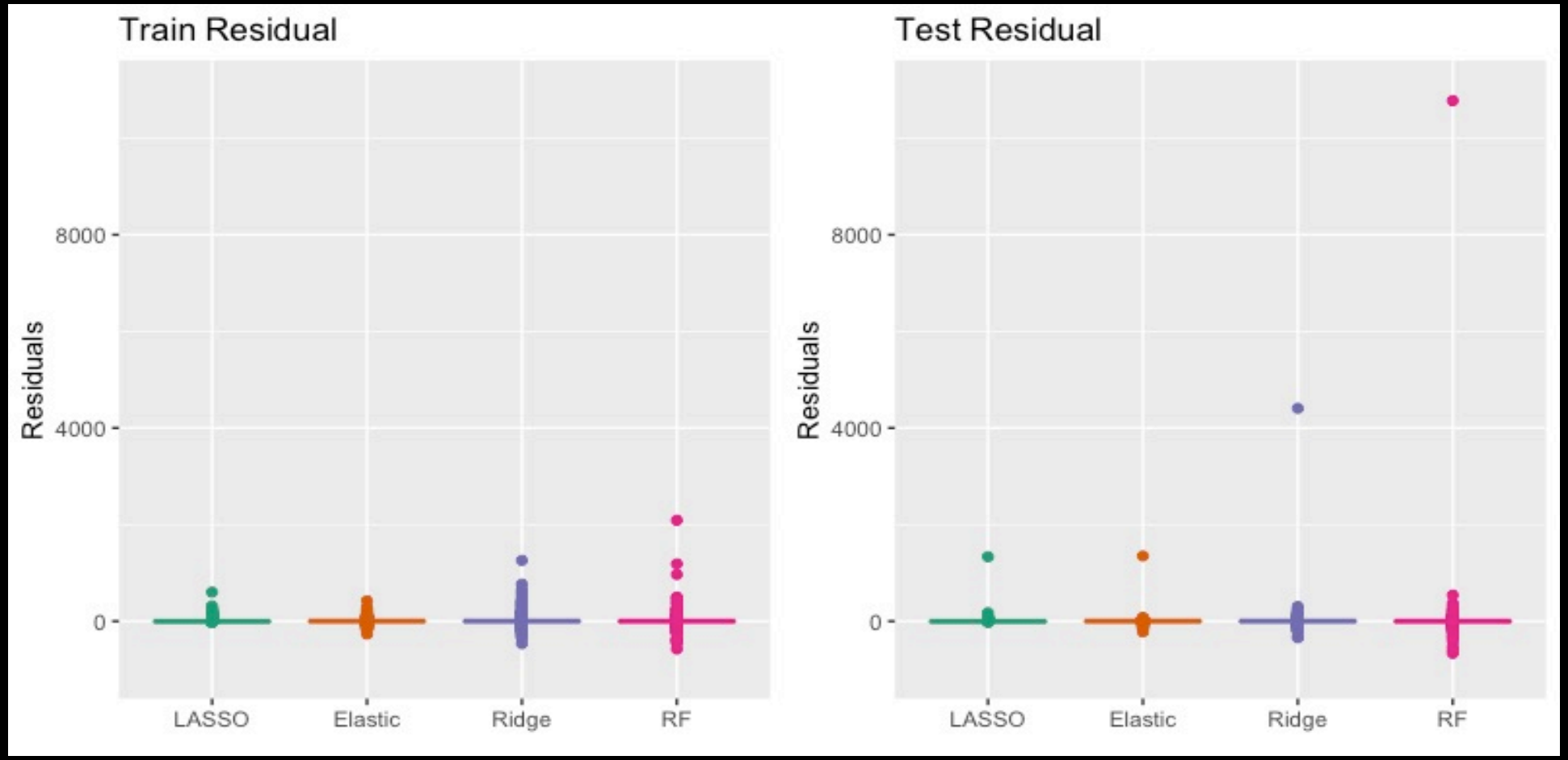


Ridge



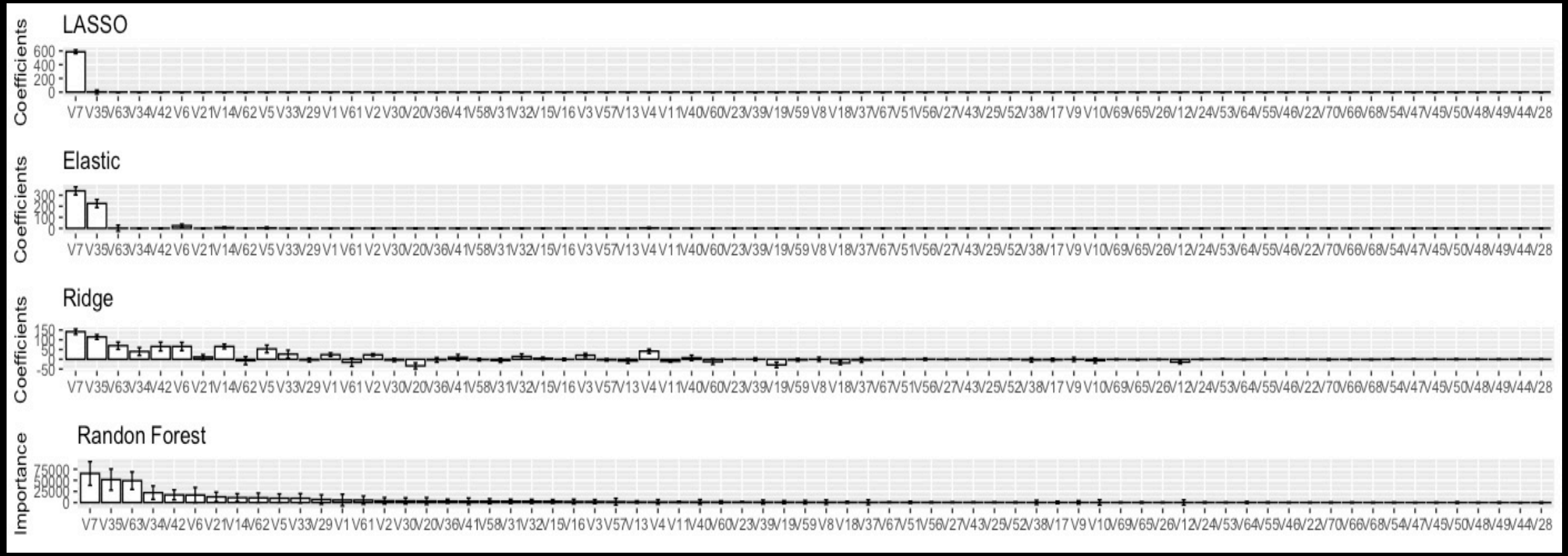
- Optimized lasso only has 2 non-zero coefficients
- Optimized elastic net has 6 non-zero coefficients
- As expected, ridge doesn’t have zero coefficient and show the worst cv mse
- In general, elastic net has the lowest variance in cv mse and ridge the largest

Residuals



- Models tend to predict the target much lower than the true value as the absolute value of positive residuals is way larger than that of negative ones.
- Size of train residual is 4665 and that of test residual is 1167(0.2*n)
- Test residual has more spread. Variance would be much greater because it's squared.

Estimated Coefficients(with bootstrap)



- 4 models have in common Top 2 variables that has the largest coefficient: V7 - # of created discussions, V35 – # of atomic containers in the same topic
- Ridge and Random Forest select V63 as the third: contribution sparseness
- Ridge has lower coefficients than lasso and elastic net
- Random Forest has more spread in variable importance than others

Conclusion

	Train R-Square		Test R-Square		Time (s)	
	Mean	SD	Mean	SD	Mean	SD
LASSO	0.99912	5.652e-05	0.99842	0.00259	0.262	0.133
Elastic Net	0.99902	4.625e-05	0.99826	0.00204	0.270	0.072
Ridge	0.99417	1.804e-04	0.98659	0.01814	0.470	0.111
Random Forest	0.99999	6.267e-07	0.99993	0.00012	65.998	166.554

- In general, all the models result in good performance.
- The best model is Random Forest considering the trade-off between time and performance because average time to model is just around 1 minute and Random Forest has the largest mean and least SD in test R-square.
- Ridge is the worst in terms of performance.
- Top 2 variable that the most predict the target are in common.