# STAT 452 Project 1

## Introduction & Set-up

In this project, a most-suitable model is used to predict a set of response variables. The training data consists of 20 explanatory variables and 10000 data points. All variables are given numerically, and no additional context has been provided.

| | Y <dbl> | X1 <dbl> | X2 <dbl> | X3 <dbl> | X4 <dbl> | X5 <dbl> | X6 <dbl> | X7 <dbl> | X8 <int> |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 15.443 | 4697.65 | -2.741869 | 70.88615 | 41.60825 | 31.1671 | -15.44438 | 6.6934422 | 106 |
| 2 | 4.876 | 3999.52 | -6.806981 | 42.08085 | 46.00709 | 33.9254 | -14.90708 | 0.0809467 | 25 |
| 3 | 13.834 | 3191.67 | -5.357907 | 71.32753 | 42.67553 | 38.6468 | -18.02603 | 0.7098575 | 16 |
| 4 | 2.611 | 4690.61 | -17.860011 | 121.20172 | 41.36721 | 33.1410 | -19.97657 | 2.1803937 | 42 |
| 5 | 1.519 | 1391.26 | -4.341287 | 53.29422 | 50.99746 | 44.3700 | -45.42413 | 1.6288052 | 17 |
| 6 | 1.866 | 4046.15 | -4.103916 | 42.70136 | 46.99762 | 33.2816 | -11.63259 | 0.4439681 | 11 |

6 rows | 1-10 of 22 columns

A set of fitting techniques are used, and one of the models will be chosen to predict the response values. The model chosen is based on the lowest mean-square predicted error (MSPE). The fitting techniques, in sequence, are:

- Ridge
- LASSO
- Least squares (LS)
- Stepwise
- Partial least squares (PLS)
- Regression tree (RT)
- Random forest (RF)
- Boosting

Initially, all 20 variables will be used, and no interactive terms are considered. Once the model is picked, the pairwise interaction terms are used, if necessary, to attempt to improve the prediction error.

A 5-fold CV is used (each fold has a validation set of 2000 data rows). This is chosen realistically due to hardware limitations and runtime.

The details of each method is explained in each subsection. This report does not include the description of the methods themselves.

## Ridge

The λ values used for testing range from 0 to 100, with increments of 0.05. In total, there would be 2001 values of λ. The λ that gives the least GCV in each fold is used for prediction.

The (average) MSPE is 26.78.

LASSO

Both the $\lambda_{min}$ and $\lambda_{1SE}$ are taken into consideration.

The MSPE for $\lambda_{min}$ is 26.79.
The MSPE for $\lambda_{1SE}$ is 27.19.

Least Squares

The MSPE for the simple linear regression model is 26.78.

Hybrid Stepwise

The stepwise method begins with the null model, and adds variables depending on the importance.

The MSPE for the selected stepwise model is 26.78.

Partial Least Squares

The MSPE for the PLS model is 26.78.

Regression Tree (with bootstrap)

Once a default regression tree is fit, prune it using the minimum complexity error.

The MSPE for the pruned regression tree is 49.52.

Random Forest

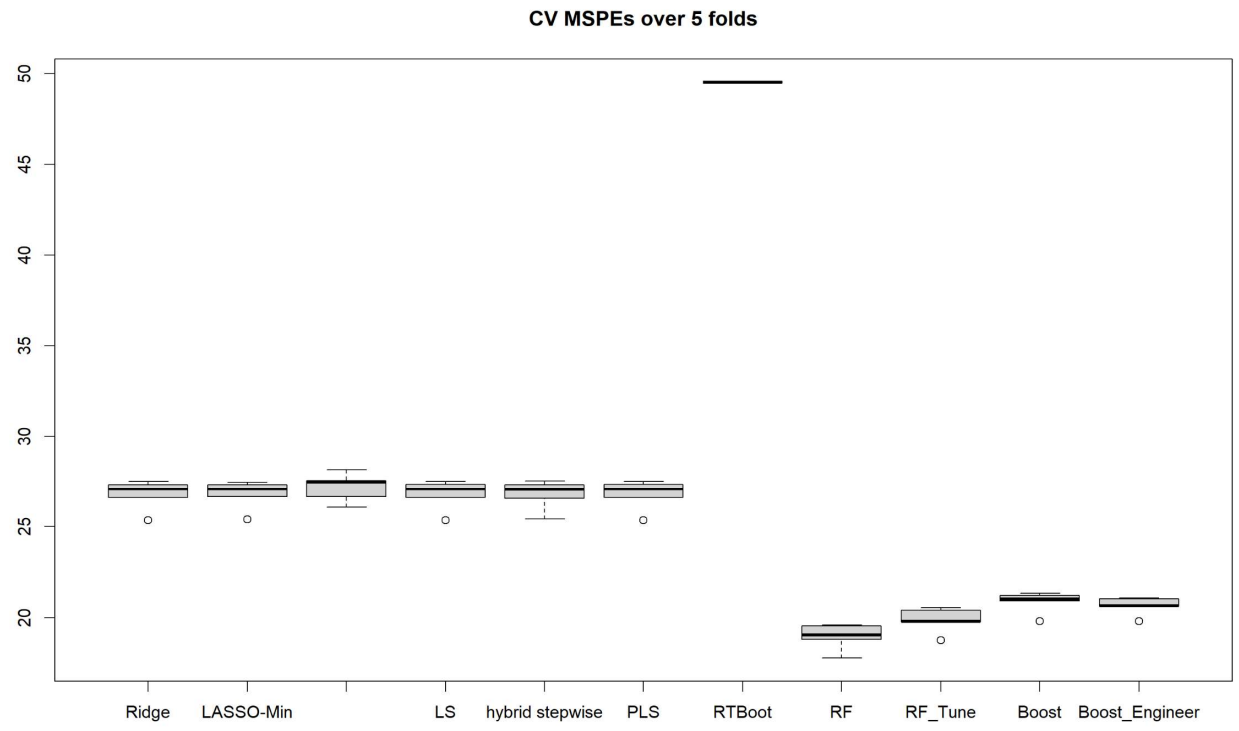
First a default random forest model is fit. Then, some tuning is done to improve the forest. The tuning parameters chosen are 2, 3, and 4, and the 3 node sizes are 3, 5, and 10.

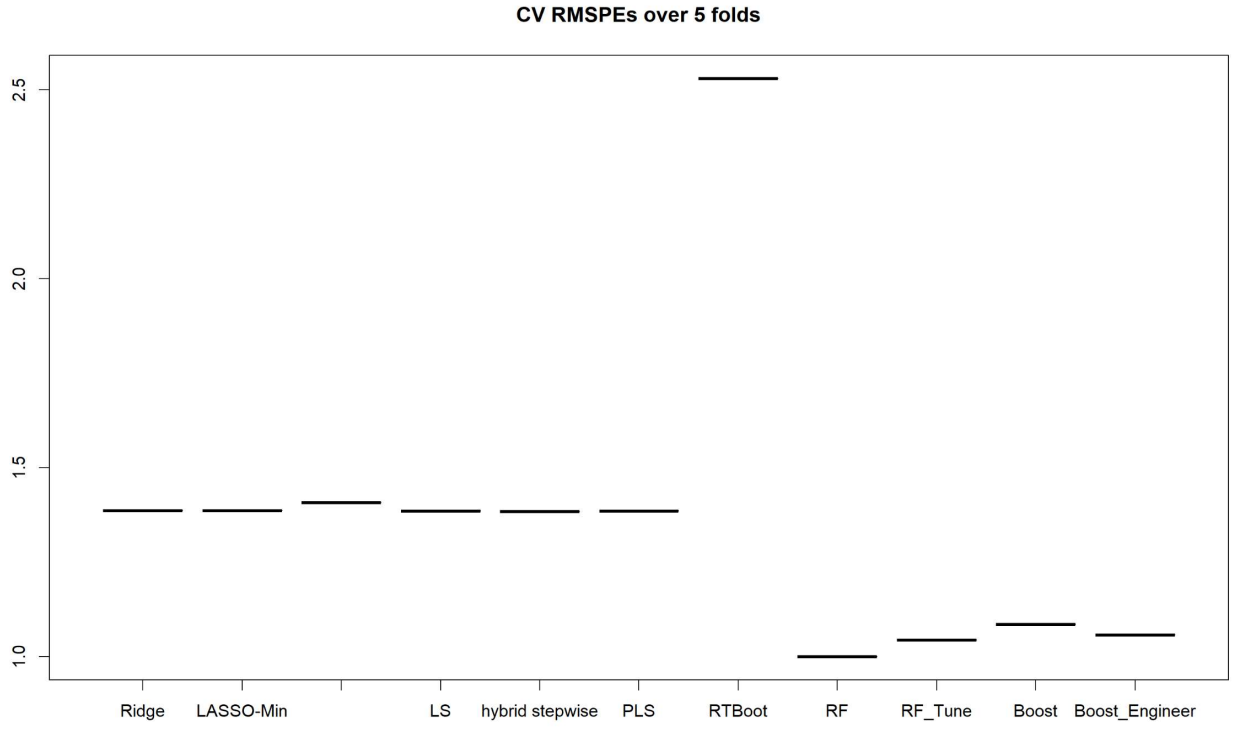
The MSPE for the default random forest is 18.98, and for the pruned forest it is 19.88.

Boosting

Two types of boosting are used: the original boosting and the engineered feature. The shrinkage, $\lambda$, is picked to be 0.001, 0.005, 0.025, and 0.125, while the tree sizes are picked to be 2, 4, and 6, respectively.

The MSPE for the default boosting is 20.88, and for the boosting with engineered features is 20.66.



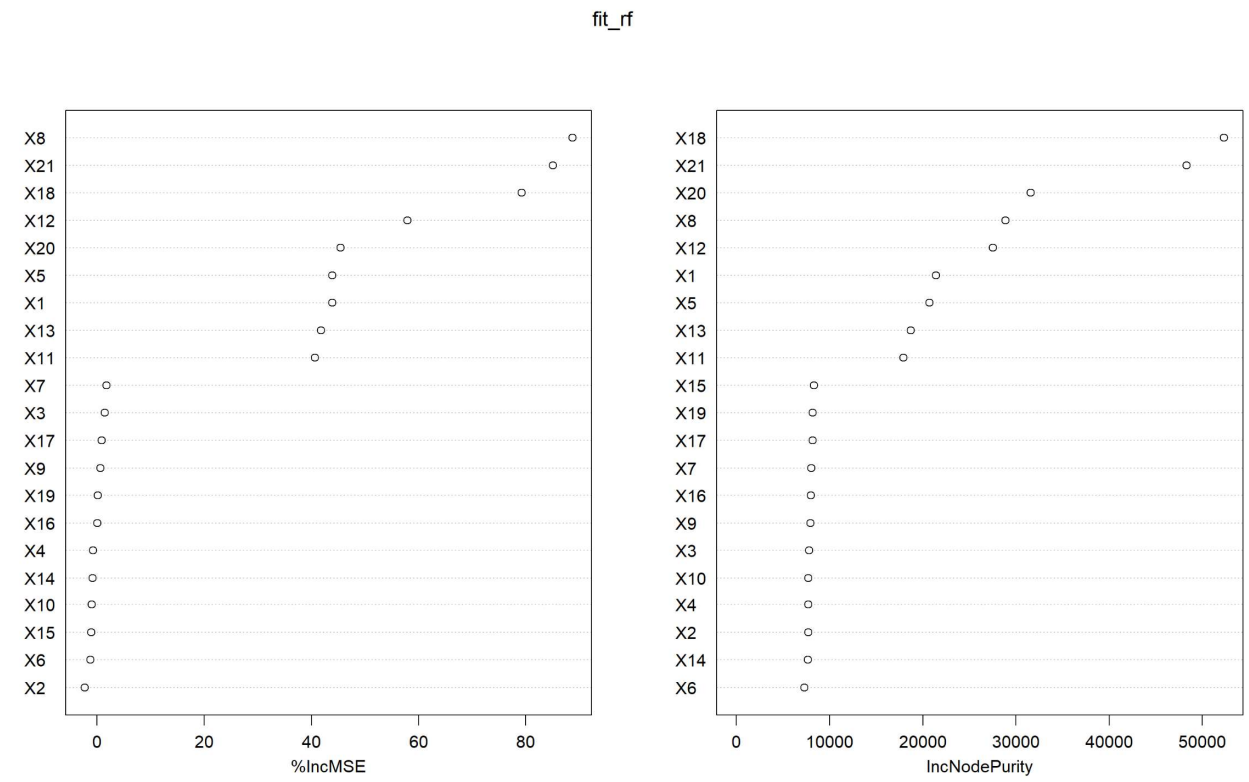

As shown, the ensemble methods perform much better than the non-ensemble methods. However, they are also computationally demanding, taking up a lot of the run time.

In the end, a default random forest model without tuning is used. (Tuning just takes too long to execute)

Random Forest – the chosen model

With the model chosen, it is fit over the entire dataset, i.e. the entire dataset becomes the training set. The important variables will be extracted:

```
        %IncMSE IncNodePurity
X1   43.94150076     21412.754
X2   -2.21659814      7706.937
X3    1.51217806      7804.062
X4   -0.73901619      7720.790
X5   43.94208136     20719.455
X6   -1.20515990      7289.489
X7    1.77714928      8042.508
X8   88.68133137     28884.786
X9    0.68085727      7940.627
X10  -0.94837722      7739.706
X11  40.66727945     17934.636
X12  57.94395423     27517.399
X13  41.80026589     18708.778
X14  -0.82752160      7672.576
X15  -1.02929820      8332.525
X16   0.08890042      8007.541
X17   0.91477212      8178.474
X18  79.21941907     52298.379
X19   0.16850305      8196.739
X20  45.50002799     31573.912
X21  85.09980830     48314.007
```




From the plot and the importance indicators, the important variables seem to be X1, X5, X8, X11, X12, X13, X18, X20, and X21. A final random forest model is fit using these variables only, resulting in the prediction responses.