STAT452 Project 1

Name Jackson Yuan Student ID:301387501

Abstract(What models did I try to use for this problem)

This is a report that I used several models to predict Y of test data by analyze from the given training data. I choose **LS,PLS,Ridge& LASSO, Random Forest, and Boosting** to get MSPE to decide which model is the best model to predict Y of test data.

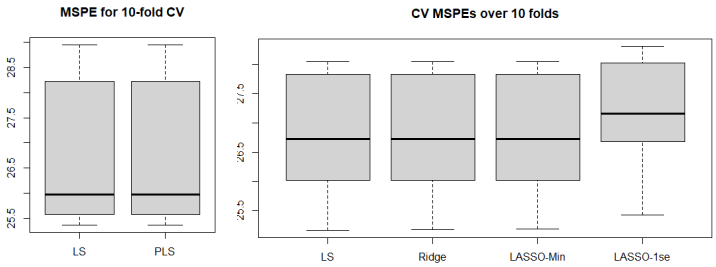
First, I read the csv file and check the dimension of the training data.

```
data <- read.table("training_data.csv",header = TRUE, sep = ",", na.strings = " ")
dim(data)
```

How did I evaluate and compare models.

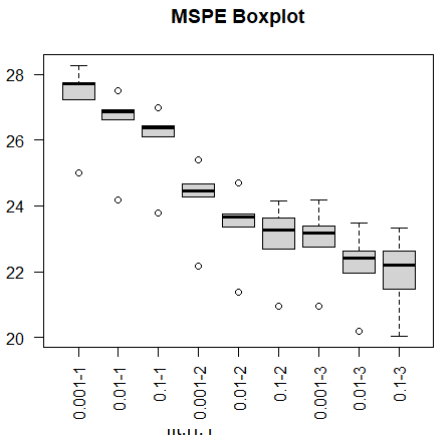
Then I start from the LS and PLS Method and LASSO and Ridge to **compare which model give me a lower MSPE**

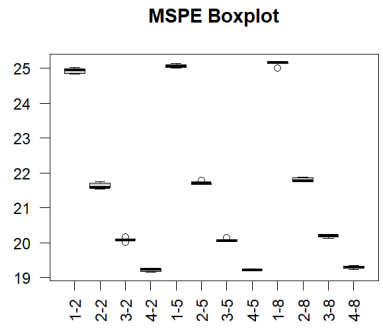
It give the minimum MSPE for 25.5-26 after compiling the **LS and PLS** and **LASSO and ridge** method.

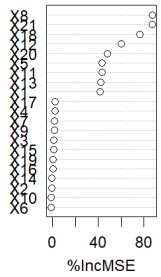


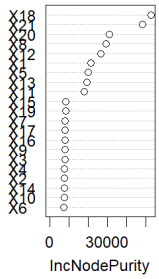
But after using Boosting and Random Forest at some initial value, I decide to compared them to choose one as my best model. Because it give me a less MSPE compared to LASSO& Ridge, LS, PLS. The parameters as the code show:

```
##Boosting method
set.seed(12345678)
max.trees <- 10000
all.shrink <- c(0.001, 0.01, 0.1)#I choose 0.001,0.01 and 0.1 as my first guess for shrink
all.depth <- c(1,2,3)#I start from 1 to 3 as my depth
al1.pars <- expand.grid(shrink = al1.shrink, depth = all.depth)
n.pars <- nrow(a11.pars)
##RandomForest
all.mtry <- 1:4 #I start from 1 to 4 as try parameter
all.nodesize <- c(2,5,8)#I choose 2,5,8 as node size
```





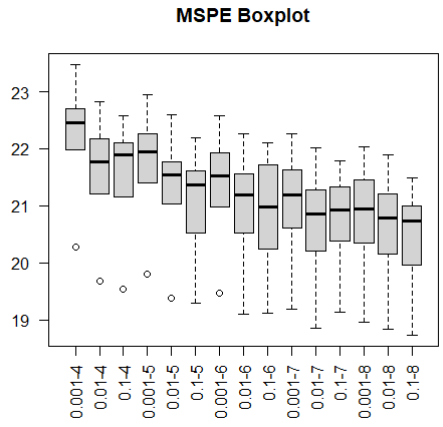


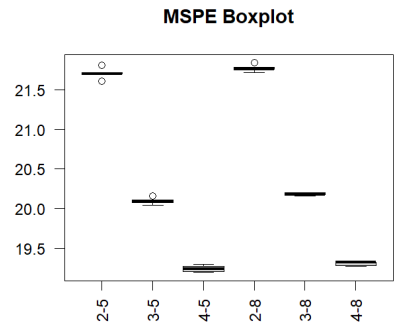


Boosting and Random Forest has tuning parameter, how do I choose those parameter.

Boosting and Random Forest shows much less MSPE than the other four models., especially Random Forest gives as value 19 as low MSPE. And `I found that shrinkage goes from 0.001 to 0.1, and MSPE decreases as shrinkage gets larger` . `From 1 to 4, MSPE also decreases as depth grows` . I find that the MSPE of Random Forest model at the tuning parameter of 1 has much difference from the other three so I use the second round by reduce some parameter. **And I consider x18, x21, x20, x8, x12, x1, x5, x13, x11 is important predictor as Random Forest give a apparent result.**(9 true predictors)

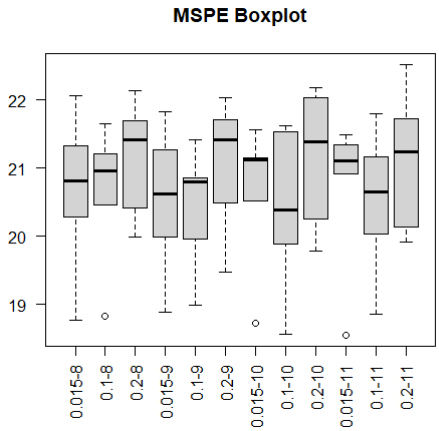
```
##Boosting method
all.shrink <- c(0.001, 0.01, 0.1)#Hold 0.001,0.01 and 0.1
all.depth <- c(4,5,6,7,8)#continue increase depth
##RandomForest
all.mtry <- c(2,3,4) #I start from 1 to 4 as try parameter
all.nodesize <- c(5,8)#I choose 2,5,8 as node size
```

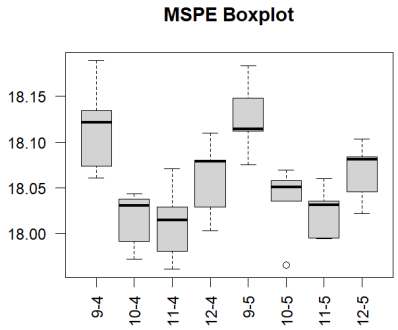


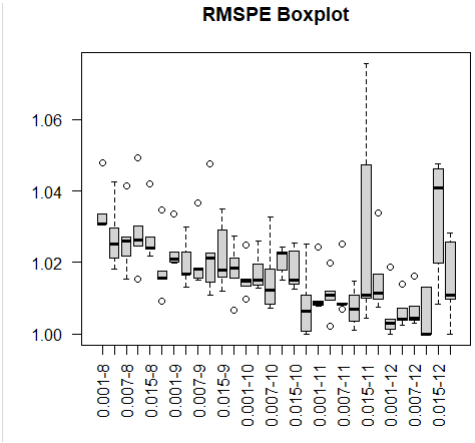


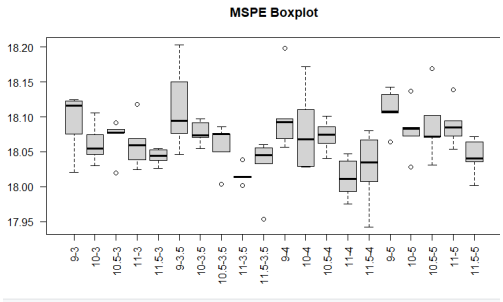
After I increase depth parameter, **Boosting model** give decrease MSPE and we can find MSPE continue decreasing as depth parameter increase. **Random Forest** give the better MSPE than before and it apparently lower when try parameter increase and node size is lower when it is 5 than 8. So I test value of try parameter greater than 8 and node size is 4,5 to see whether the model give me a lower parameter. And I change the range of shrinkage and continue to increase the value of depth to see new MSPE.

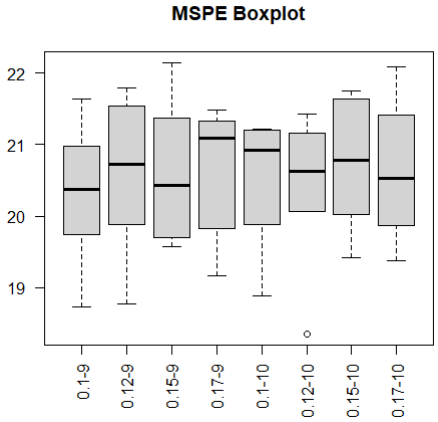
```r
##Boosting method
all.shrink <- c(0.015,0.1,0.2)#Change the range of shrinkage
all.depth <- c(8,9,10,11)#continue increase depth
## Random Forest
all.mtry <- c(9,10,11,12)
all.nodesize <- c(4,5)
```

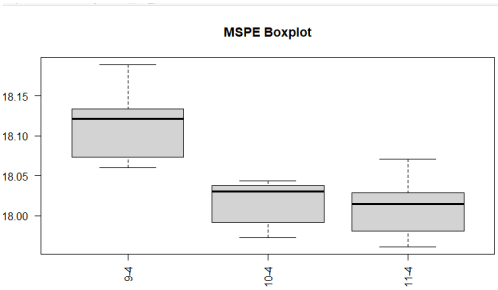













```
Approximate significance of smooth terms:
          edf Ref.df       F p-value
s(X21) 8.679  8.968 114.171  <2e-16 ***
s(X8)  7.453  8.275  39.413  <2e-16 ***
s(X18) 7.515  8.482  10.501  <2e-16 ***
s(X12) 8.202  8.843  15.323  <2e-16 ***
s(X20) 8.104  8.739  16.466  <2e-16 ***
s(X1)  7.992  8.691   9.907  <2e-16 ***
s(X5)  8.216  8.853   7.999  <2e-16 ***
s(X13) 8.582  8.878   6.997  <2e-16 ***
s(X11) 7.783  8.626  10.898  <2e-16 ***
---
```

After test different sets of tuning parameter by binary search concept, **I consider Random Forest has lowest MSPE with tuning parameter as try parameter is 11 and node size is 4**. And estimate of the number of true predictors is on the last picture according to GAM model.