Homework 4

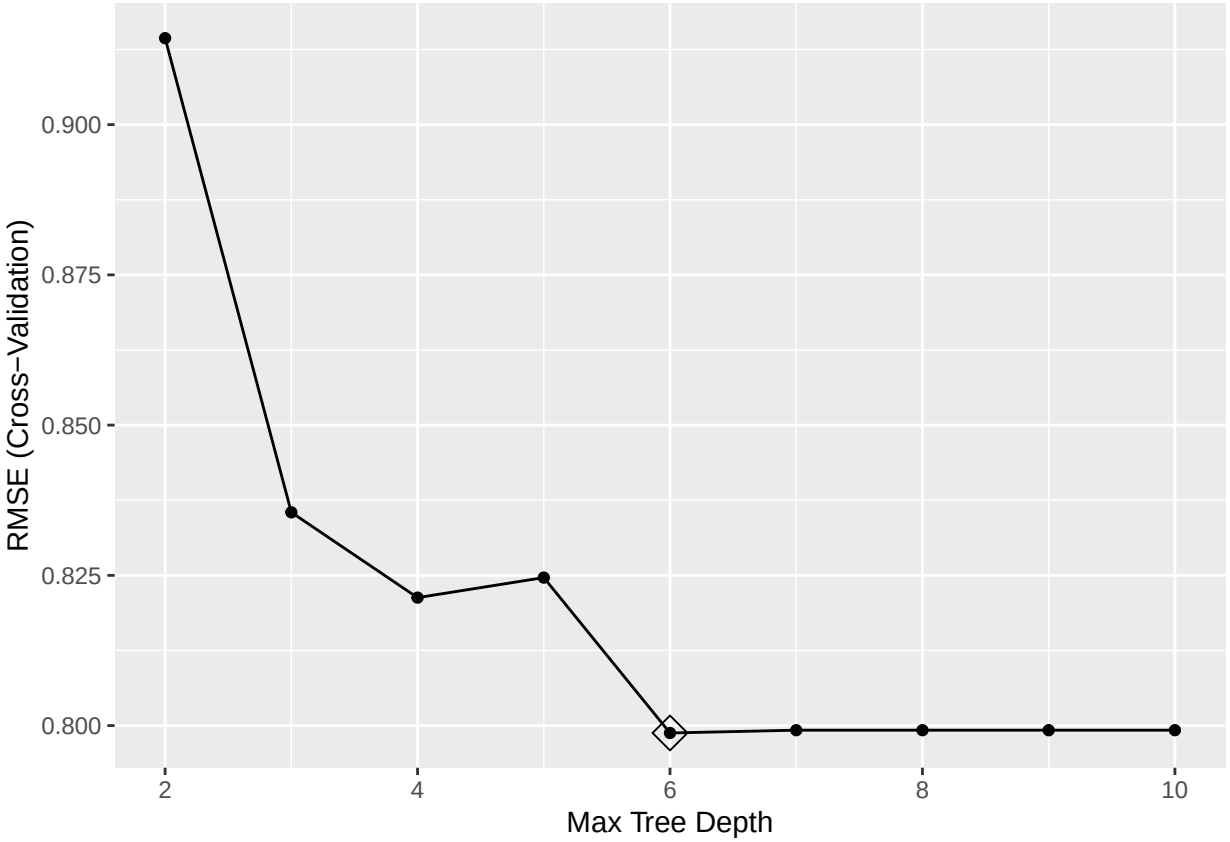
Ziyi Zhao

4/26/2020

Part 1

part a)

Fit a regression tree with lpsa as response variable and the other predictors as predictors. Use CV to determine the optimal tree size. Which tree size corresponds to the lowest cv error? Is this the same as the tree size obtained using 1 SE rules?

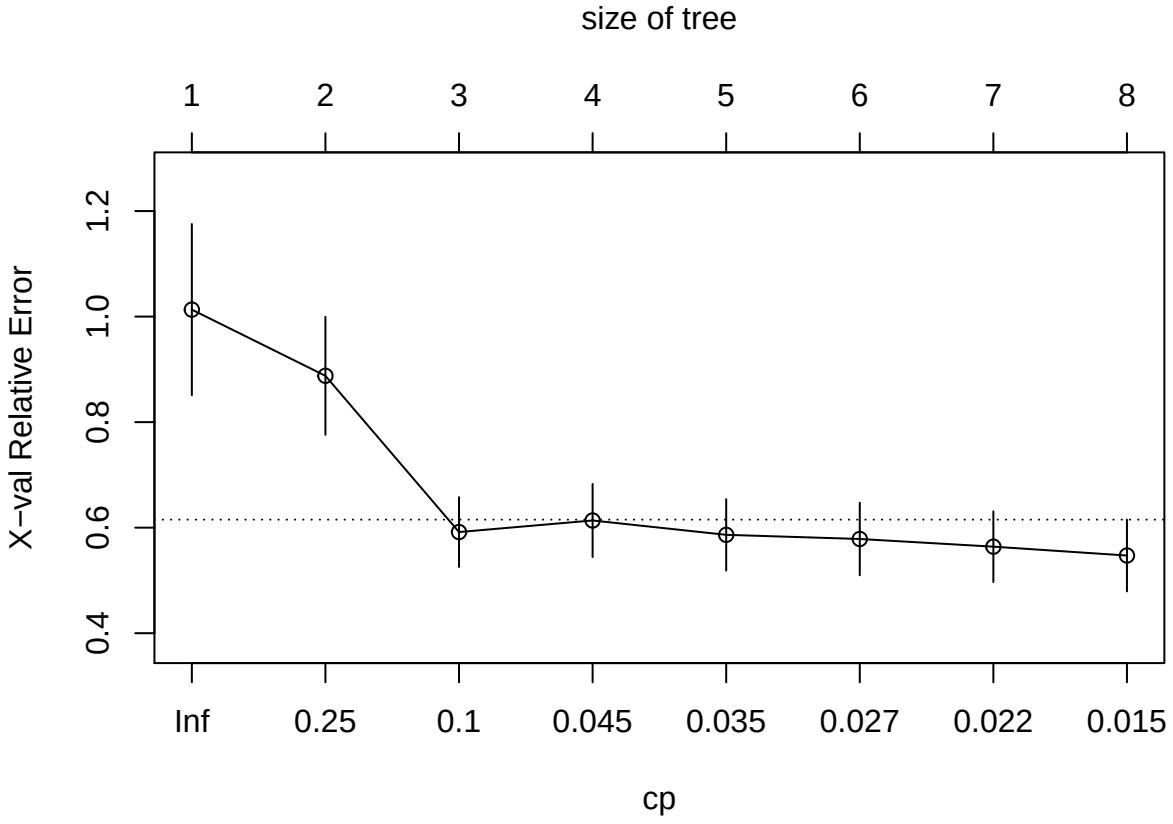


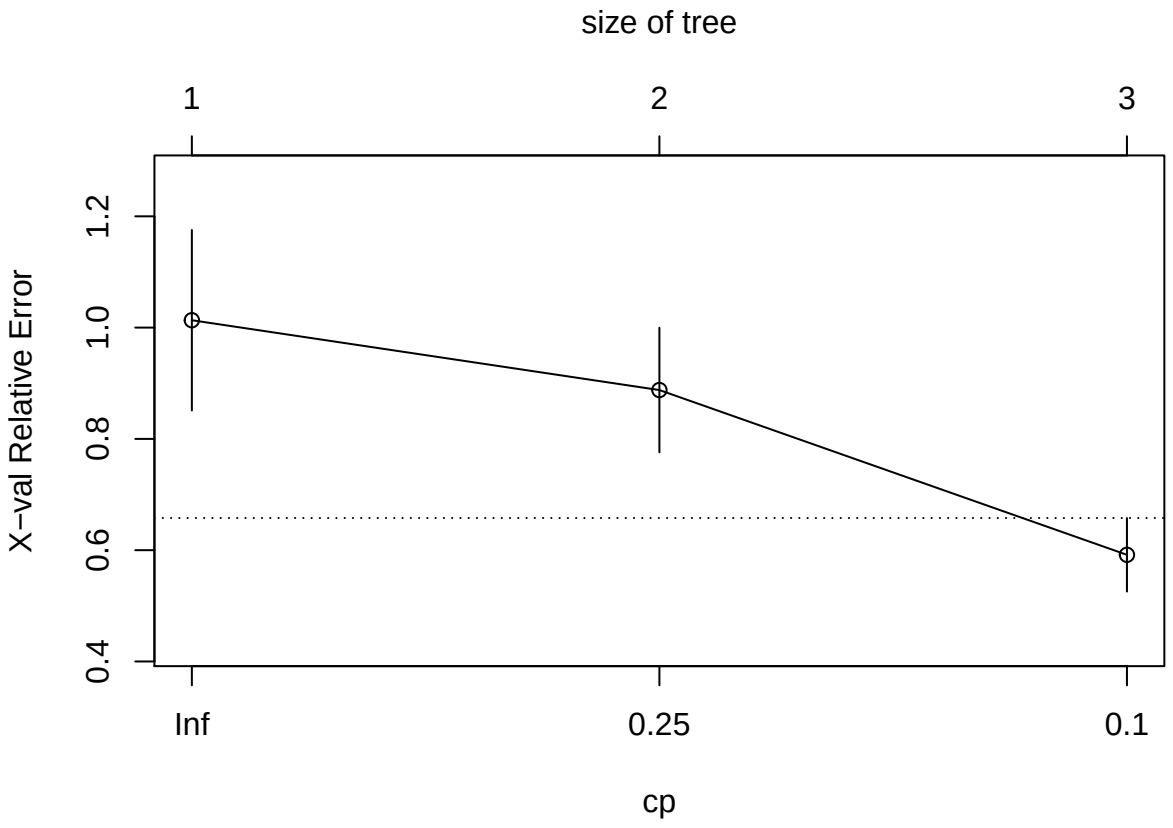

```
##   maxdepth
## 5        6
```

Using the cross validation, we can find out that the tree size = 6 correspond to the lowest CV error.

```
##
## Regression tree:
## rpart(formula = lpsa ~ ., data = dat)
##
```

```
## Variables actually used in tree construction:
## [1] lcavol  lweight pgg45
##
## Root node error: 127.92/97 = 1.3187
##
## n= 97
##
##         CP nsplit rel error  xerror     xstd
## 1 0.347108      0   1.00000 1.01323 0.162162
## 2 0.184647      1   0.65289 0.88779 0.111915
## 3 0.059316      2   0.46824 0.59168 0.066102
## 4 0.034756      3   0.40893 0.61359 0.069269
## 5 0.034609      4   0.37417 0.58640 0.067630
## 6 0.021564      5   0.33956 0.57853 0.068772
## 7 0.021470      6   0.31800 0.56398 0.067155
## 8 0.010000      7   0.29653 0.54721 0.068034
```





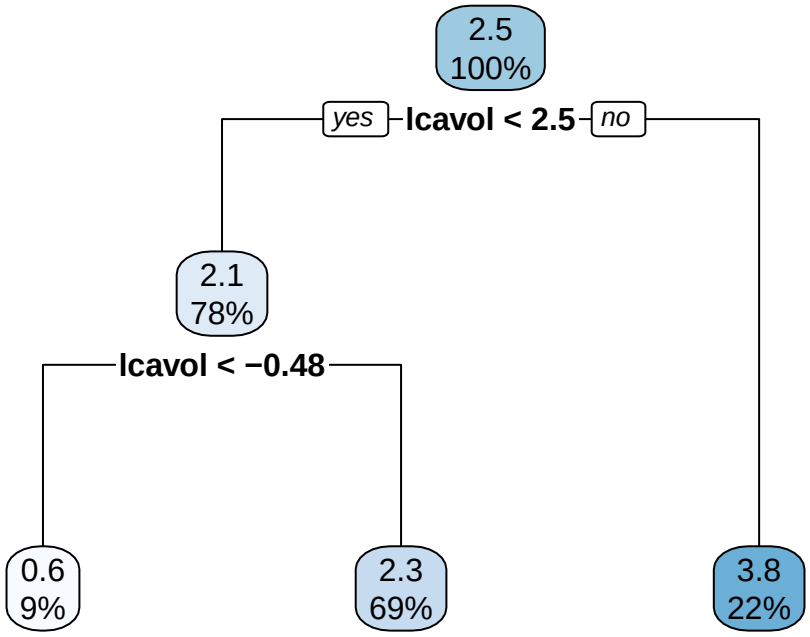
Using the 1 SE rule, we can see the tree size = 3 has the lowest x-error.

The tree size obtained by using cross validation is different from the tree size obtained by using 1 SE rule.

part b)

Create a plot of final tree you choose. Pick one of the terminal nodes, and interpret the information displayed.

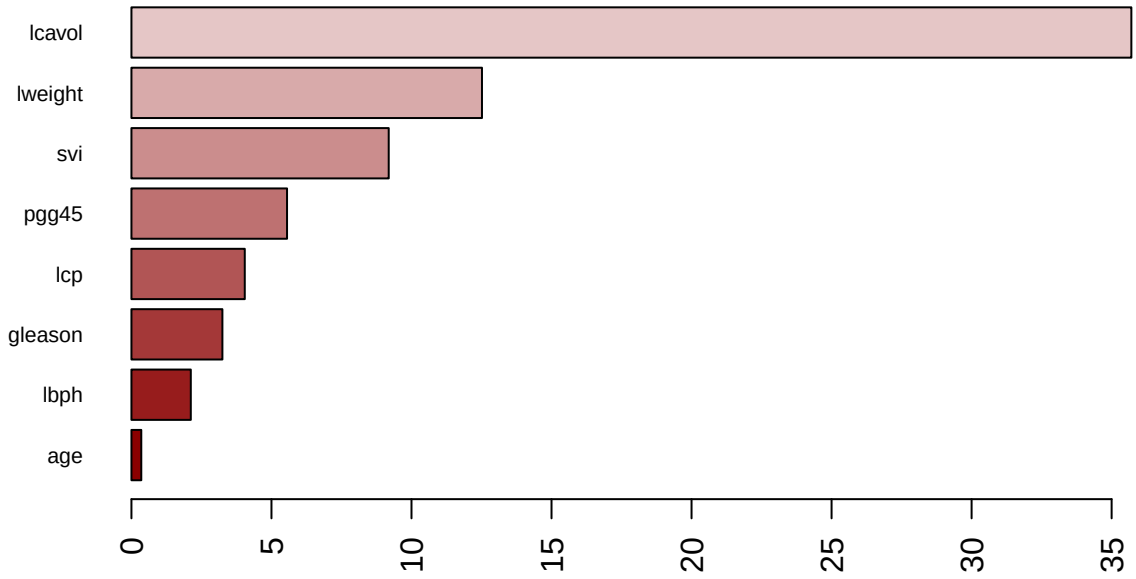
I'd choose the tree created by 1 SE rules due to relatively smaller cross validation error.



I choose the terminal node lcavol < 2.5. If log(cancer volumn) is smaller than 2.5, there is 78% chance for log(prostate specific antigen) to be 2.1. If the log(cancer volumn) is greater than 2.5, there is 22% chance for log(prostate specific antigen) to be 3.8.

part c)

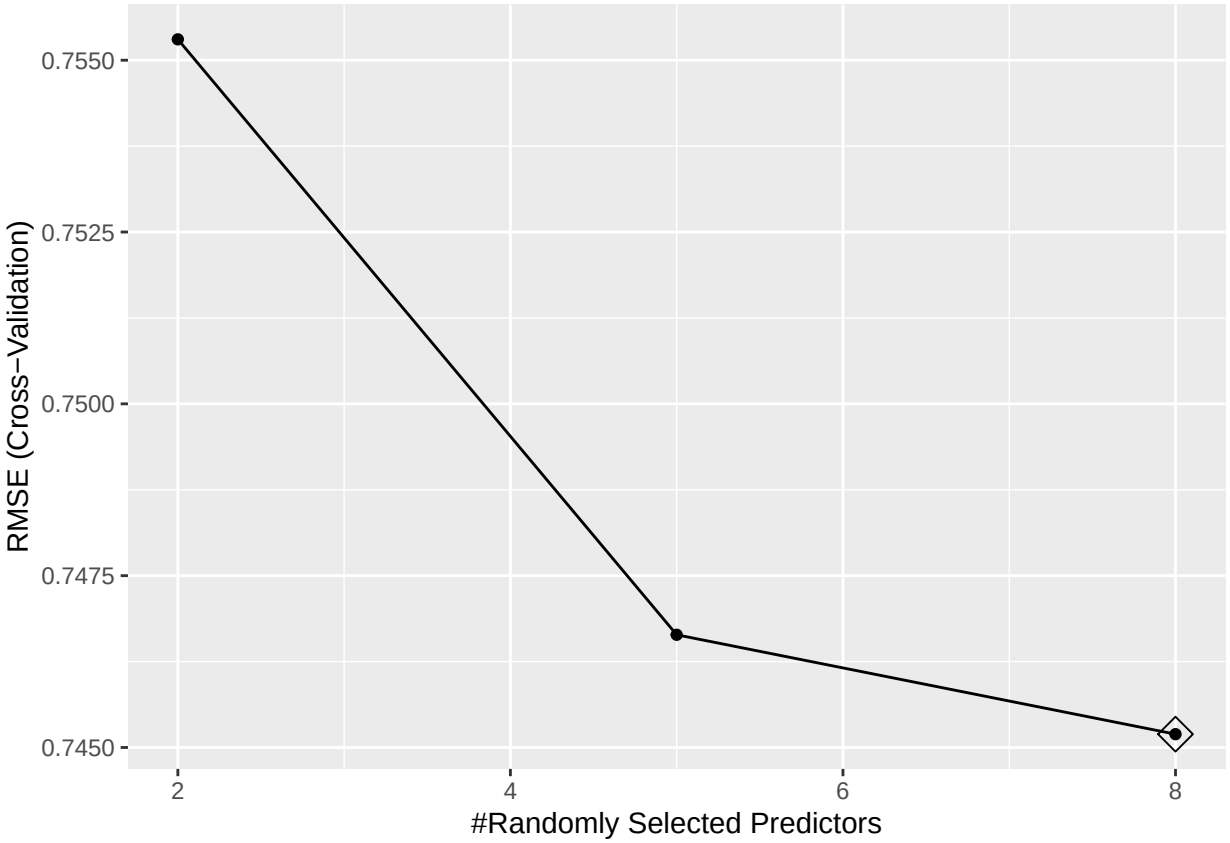
Perform bagging and report the variable importance.

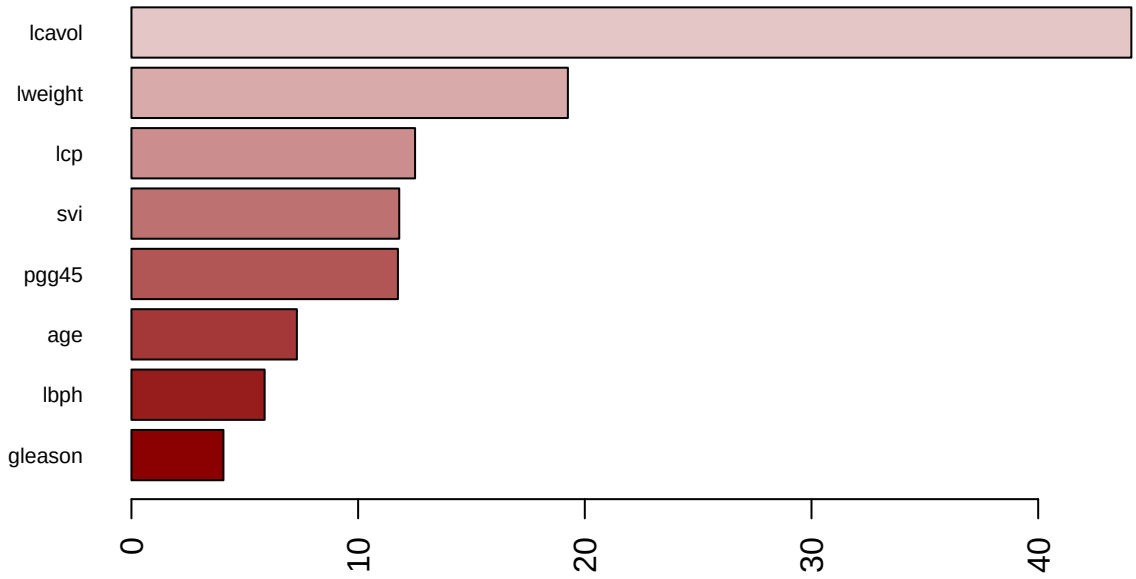


The log(cancer volumn) has the greatest relative importance (apporximate to 35). The age has the least imporance in the model.

part d)

Perform random forest and the variable importance.

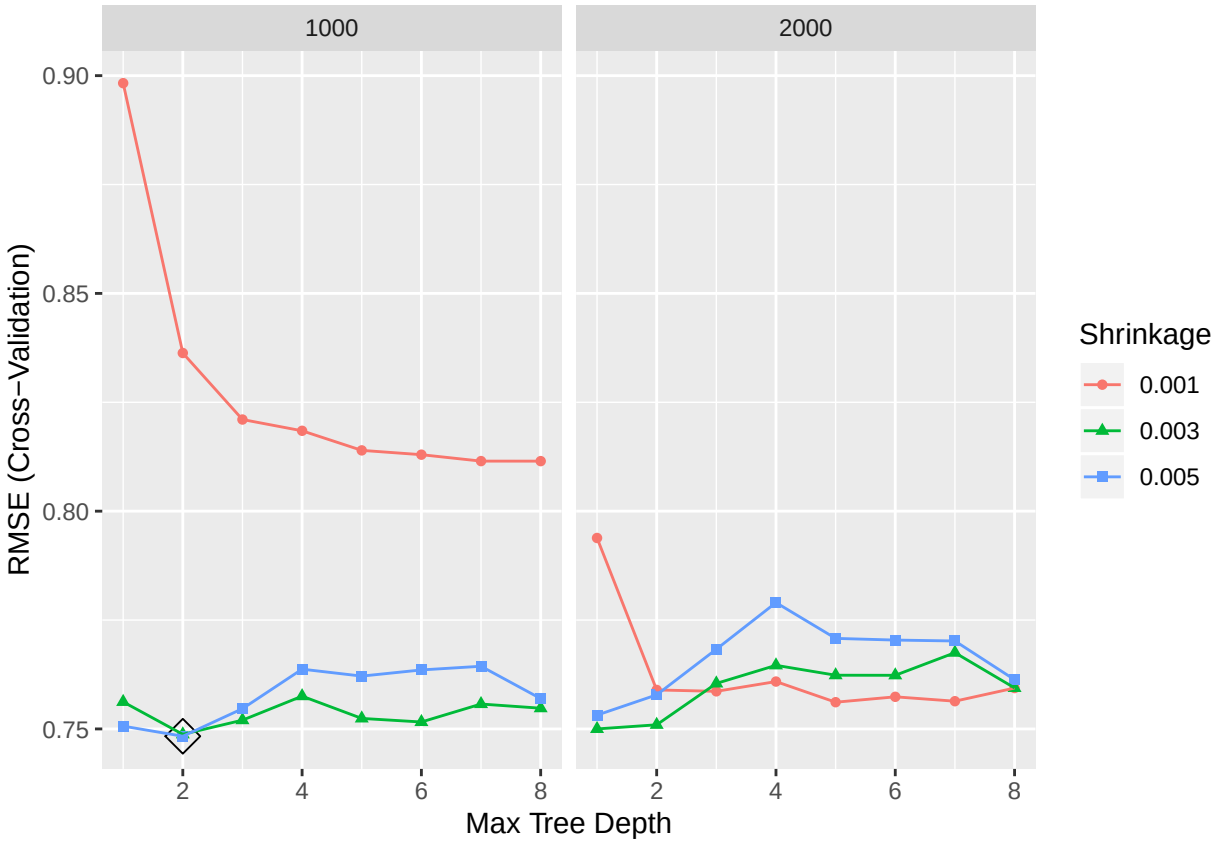

0.7550
0.7525
0.7500
0.7475
0.7450
RMSE (Cross−Validation)
2
4
6
8
#Randomly Selected Predictors

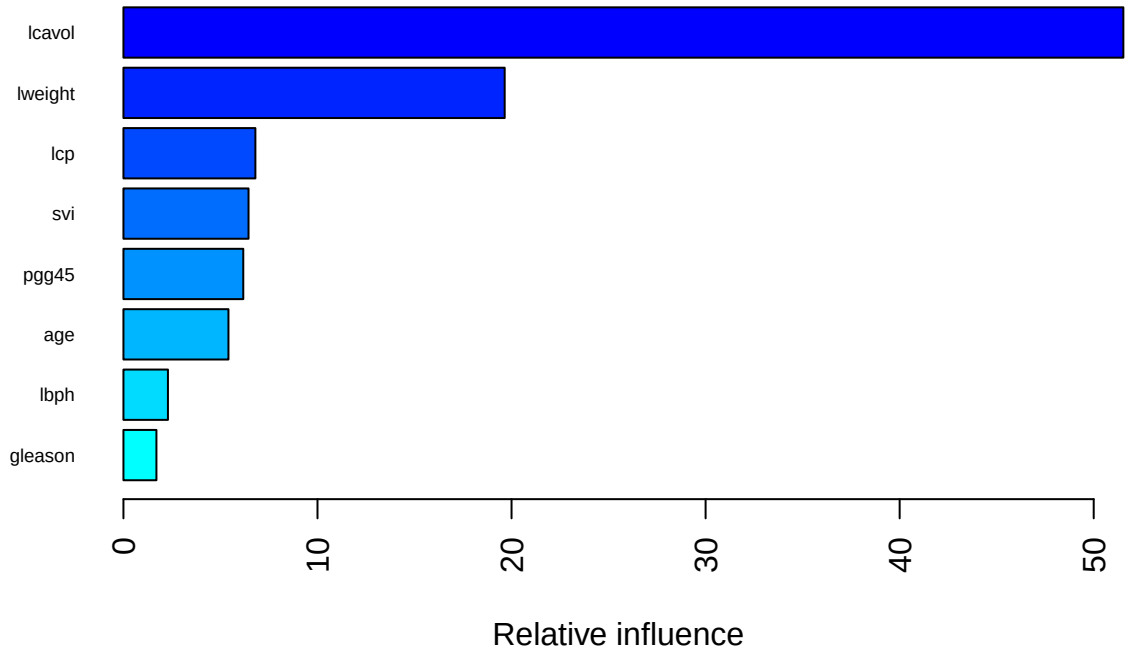


The log(cancer volumn) has the greatest variable importance (>40) and gleason score has the smallest.

part E

Perform boosting and report variable importance


1000
2000
0.90
0.85
0.80
0.75
RMSE (Cross−Validation)
2
4
6
8
Max Tree Depth
Shrinkage
0.001
0.003
0.005




```
##             var   rel.inf
## lcavol   lcavol 51.527840
## lweight lweight 19.646266
## lcp         lcp  6.799666
## svi         svi  6.446269
## pgg45     pgg45  6.175011
## age         age  5.411604
## lbph       lbph  2.293667
## gleason gleason  1.699678
```

The log(cancer volumn) has the greatest relative influence among all predictors (51.53). The gleason score has the smallest (1.70).

Part F

Which model will you select to predict PSA level? Explain.

```
##
## Call:
## summary.resamples(object = resamp)
##
## Models: bag, rf, gbm
## Number of resamples: 10
##
```

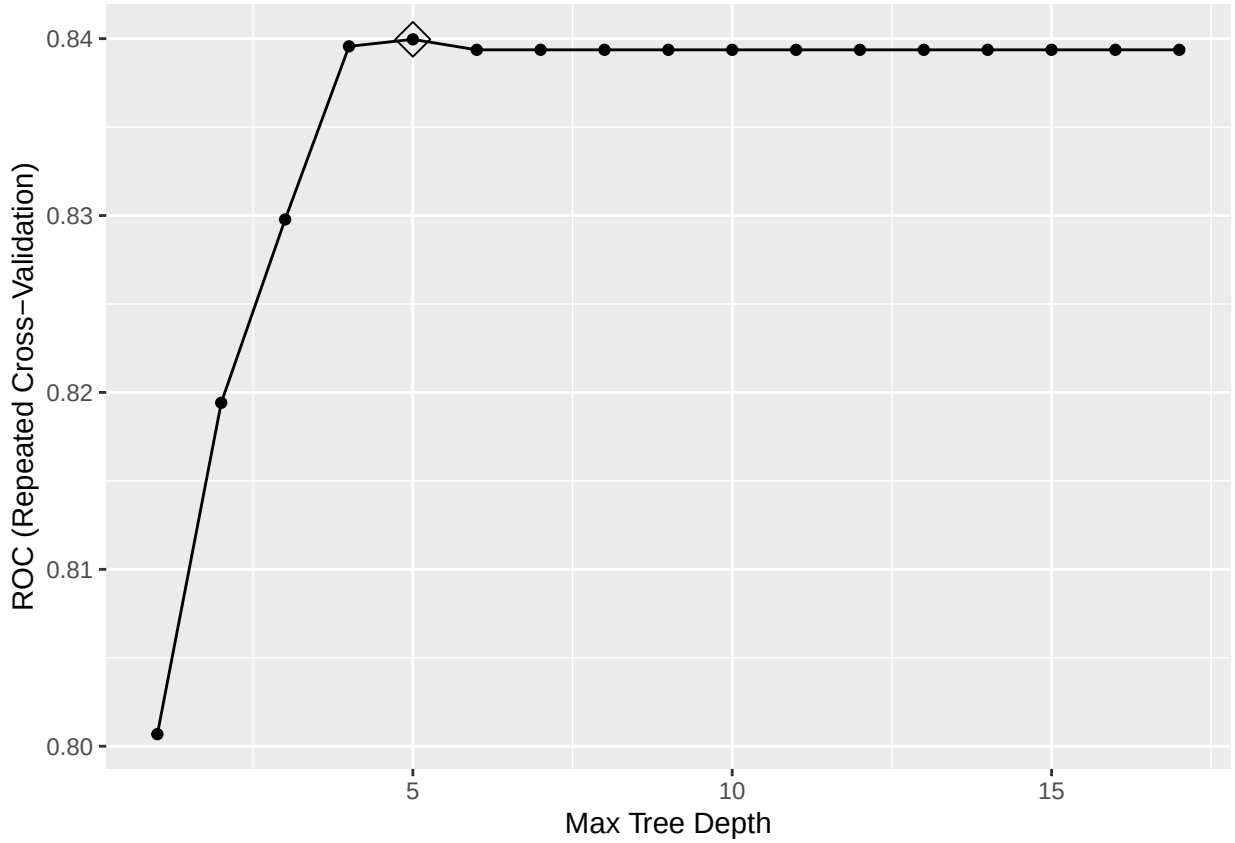
```
## MAE
##          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## bag 0.4925182 0.5154656 0.5953374 0.5997542 0.6798249 0.7056196    0
## rf  0.5046822 0.5188930 0.6090570 0.5990042 0.6461156 0.7177101    0
## gbm 0.5093666 0.5431438 0.5945322 0.5973703 0.6327586 0.6979876    0
##
## RMSE
##          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## bag 0.5990968 0.6342680 0.7197929 0.7451939 0.8317062 0.9796059    0
## rf  0.6092195 0.6307310 0.6951370 0.7405486 0.8418719 1.0134615    0
## gbm 0.6324549 0.6809332 0.7293635 0.7483191 0.8089783 0.9016944    0
##
## Rsquared
##          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## bag 0.4742951 0.5191098 0.5861057 0.6078218 0.7001524 0.8061142    0
## rf  0.3675284 0.5023536 0.5875393 0.5974209 0.7294368 0.7930649    0
## gbm 0.3981008 0.4794449 0.6424552 0.6102653 0.7352487 0.8157170    0
```

The random forest has the smallest median and mean of RMSE; however, the boosting has larger mean and median of Rsquared. I prefer choosing random forest.

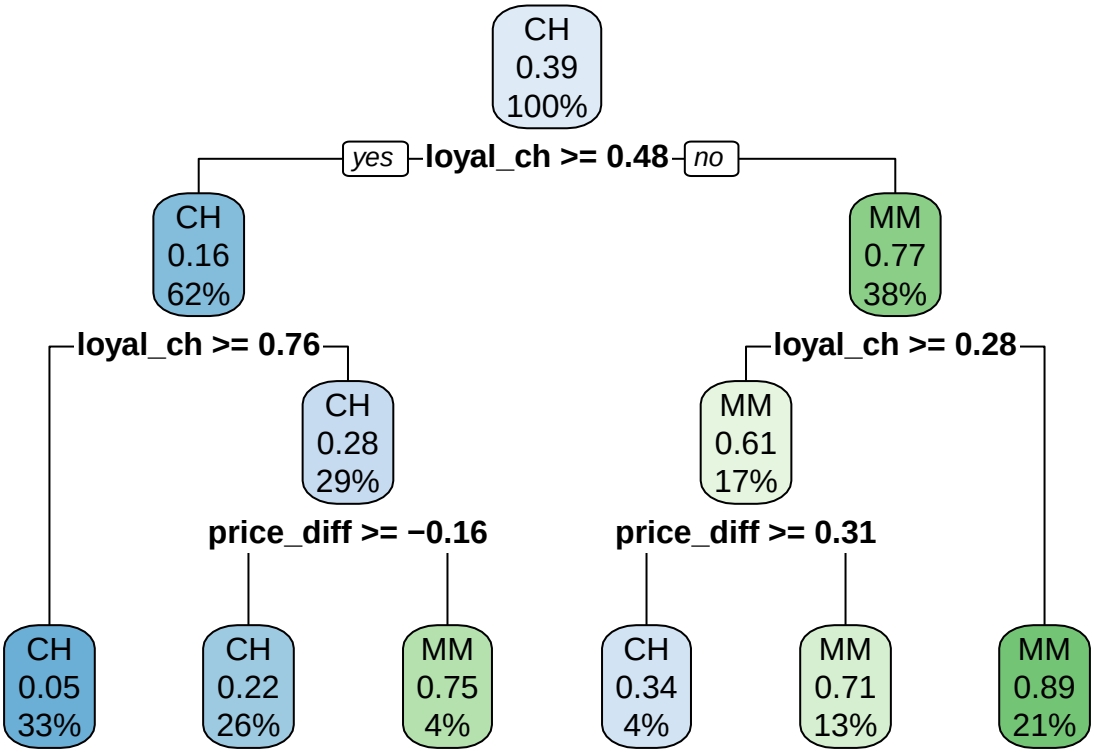
Part 2

part a)

Fit a classification tree to the training set, with Purchase as the response and the other variables as predictors. Use cross validation to determine the tree size and create a plot of the final tree. Predicted the response on the test data. What's the classification error rate?


0.84
0.83
0.82
0.81
0.80
ROC (Repeated Cross−Validation)
5
10
15
Max Tree Depth

```
##   maxdepth
## 5        5
```



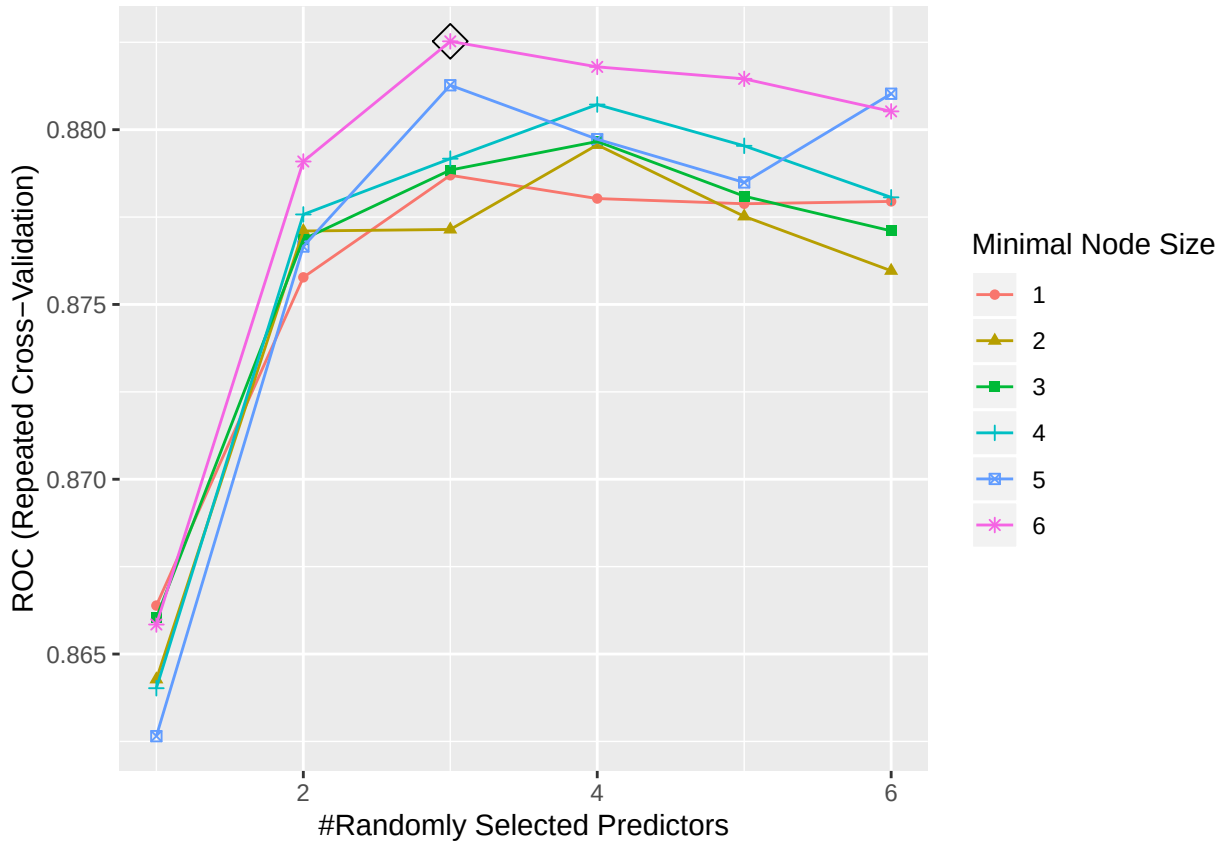

```
## Setting levels: control = CH, case = MM
```

```
## Setting direction: controls > cases
```

The auc of the classification tree using CV is 0.8832323.

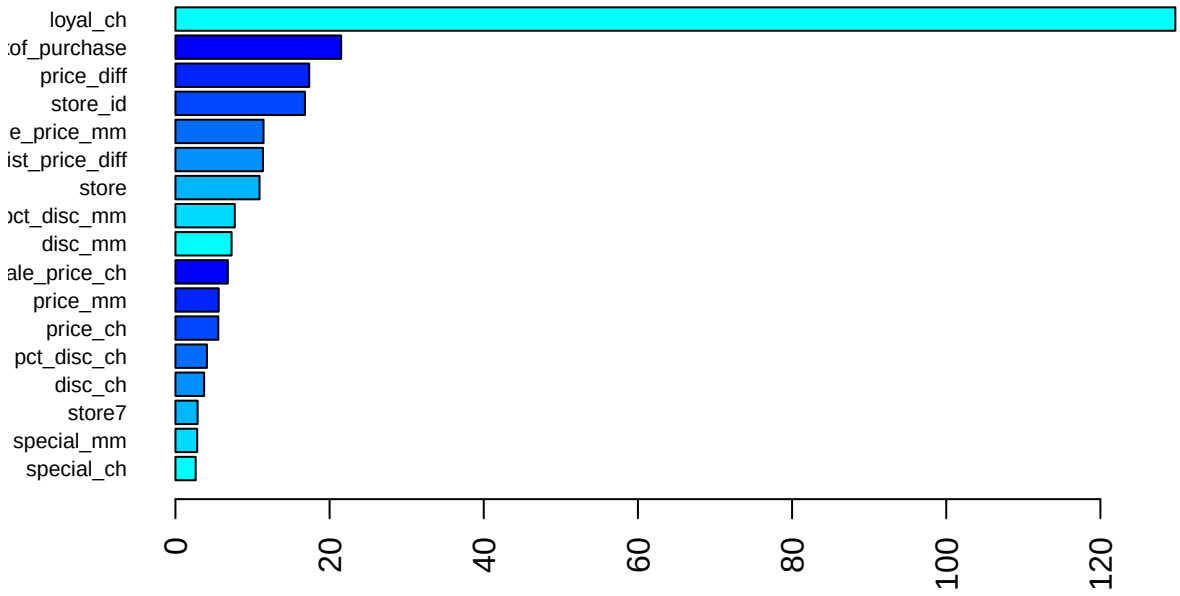
part b)

Perform random forests on the training set and report variable importance. What is the test error rate?




```
## Setting levels: control = CH, case = MM
```

```
## Setting direction: controls > cases
```

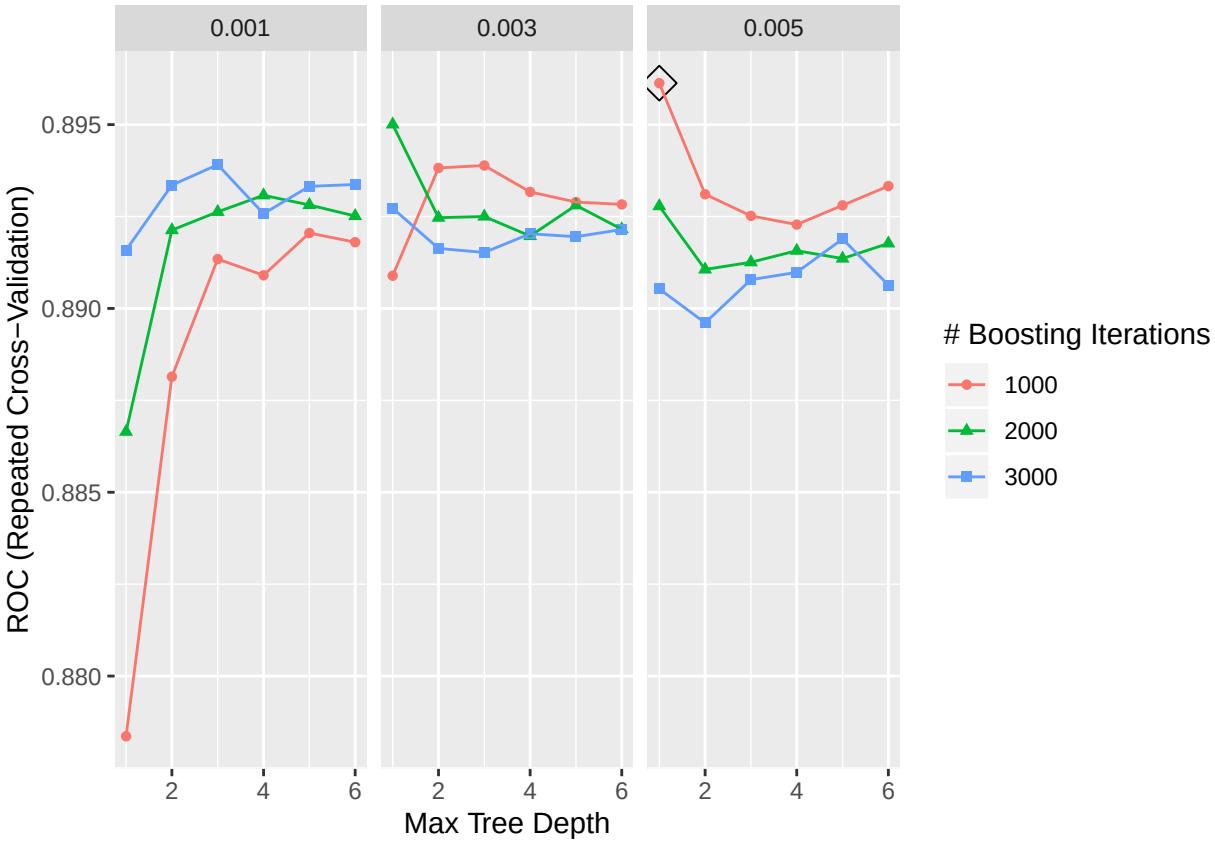


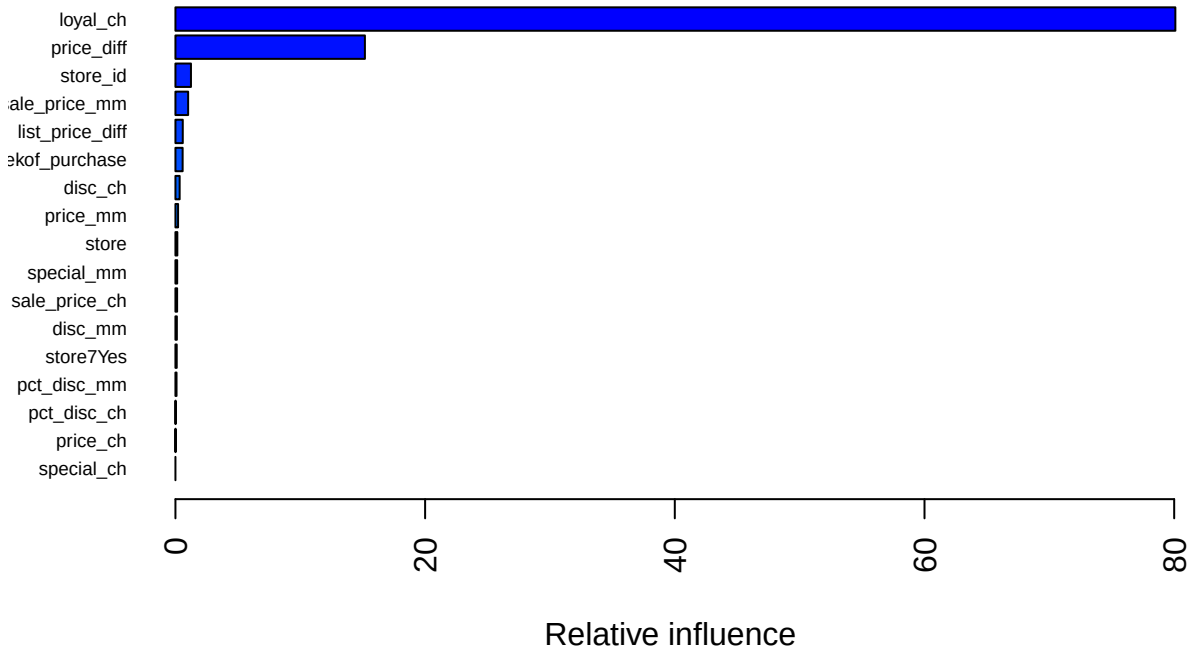
Based on the plot of importance, the variable LoyalCH has the greatest importance (>120) on the model and its importance is much greater than others. On the opposite, SpecialMM and Special CH has the smallest importance to the model.

The auc obtained by random forest is 0.8756133.

part c)

Perform boosting on the training set and report variable importance. What's the test error rate?


0.001
0.003
0.005
ROC (Repeated Cross−Validation)
0.895
0.890
0.885
0.880
2
4
6
Max Tree Depth
Boosting Iterations
1000
2000
3000




```
##                              var     rel.inf
## loyal_ch               loyal_ch 80.09395109
## price_diff           price_diff 15.17538970
## store_id               store_id  1.24280752
## sale_price_mm     sale_price_mm  1.01627682
## list_price_diff list_price_diff  0.58651089
## weekof_purchase weekof_purchase  0.57684961
## disc_ch                 disc_ch  0.33720707
## price_mm               price_mm  0.21102268
## store                     store  0.15256815
## special_mm           special_mm  0.13806839
## sale_price_ch     sale_price_ch  0.13559091
## disc_mm                 disc_mm  0.11272613
## store7Yes             store7Yes  0.10252702
## pct_disc_mm         pct_disc_mm  0.08805140
## pct_disc_ch         pct_disc_ch  0.01714380
## price_ch               price_ch  0.01330882
## special_ch           special_ch  0.00000000

## Setting levels: control = CH, case = MM

## Setting direction: controls > cases
```

The LoyalCH has the greatest relative influence on the model and it's much greater than others. Besides LoyalCH and price_diff, the other variables' importance are approximate to 0.

The AUC obtained by boosting is 0.9058874.