

**BUDT 758T**

**assignment #6: 100 PTS**

The goal of this homework is to introduce you to running and working with PCA and ensemble methods. You will gain experience with PCA, bagging, random forests, and boosting in R. You are required to complete this assignment in R—be sure to include the code you used and output any results you use!

**The Data**

There is a famous data set (contained in the ISLR package but also available on Canvas/ELMS as *College.csv*) that was collected on U.S. colleges in 1995 to show common statistics on over 700 American universities. The main goal was to find factors that influence the success of the school and determine whether it was worth it for parents to send their child to a given university. The main measure of this success was the graduation rate of the school, because parents and students wanted to be assured of a successful degree.

The variables collected in the data set are:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| *Private* | Is college a private college (Yes or No) |
| *Apps* | How many students applied for admission |
| *Accept* | How many students did the college accept |
| *Enroll* | How many students chose to enroll in the college |
| *Top10perc* | Percentage of new students enrolling who were in top 10% of their high school class |
| *Top25perc* | Percentage of new students enrolling who were in top 25% of their high school class |
| *F.Undergrad* | Number of full-time undergraduate students |
| *P.Undergrad* | Number of part-time undergraduate students |
| *Outstate* | Expected tuition cost for students who are out-of-state |
| *Room.Board* | Expected cost of room and board |
| *Books* | Expected cost of books |
| *Personal* | Expected cost for other personal expenses |
| *PhD* | Percentage of faculty who have a PhD |
| *Terminal* | Percentage of faculty with a terminal degree |
| *S.F.Ratio* | Student to faculty ratio |
| *perc.alumni* | Percentage of alumni who donate to the school |
| *Expend* | Instructional expenditure per student (how much does school spend per student on instruction) |
| *Grad.Rate* | Graduation rate of the school |

**Assignment**

Please answer all questions in the dedicated space and upload on Canvas. Please ensure that your numbering of questions matches those below. You must include your R code, but as with previous assignments, you are welcome to include your full code at the end of the assignment file rather than including it with the appropriate question. However, you should make sure any output that is requested or necessary to answer the question is including with the question. Any additional output you wish to provide may be included at the end of your assignment in an appendix, if you wish.

Remember: you are allowed to consult with others in the class on this assignment, but all submitted work must be your own (and don’t forget to include the names of anyone you consulted in the last question!).

1. **(0 points) Data Preparation:** 
   1. Read thedata set into R.
   2. Set the seed to 91101.
   3. Randomly partition the data set in the following order (note that if you do *not* follow this order, many of the questions in this assignment will not make sense to you!):
      1. Split 35% of the observations in the full data set to use as testing data. Using these observations, create a testing data set called *college\_test.*
      2. Save the remaining 65% of the data as *college\_rest*.

college <- read.csv("College.csv")

View(college)

set.seed(91101)

num\_obs=nrow(college)

test\_obs = sample(num\_obs, 0.35\*num\_obs)

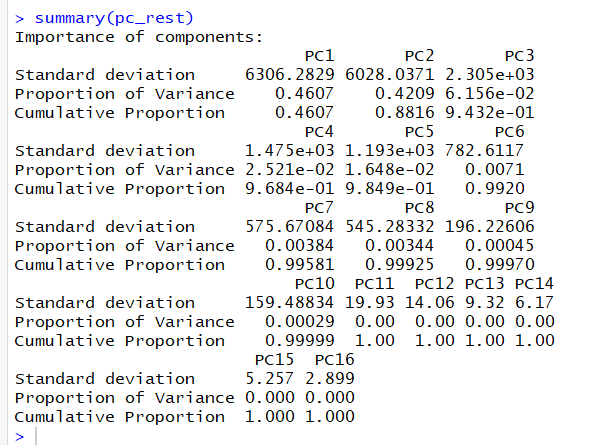
college\_test <- college[test\_obs,]

college\_rest <- college[-test\_obs,]

1. **(25 points) Principal Components Analysis**
   1. Using the *college\_rest* data, run a PCA on all numeric X variables (so you should exclude the name of the college, *Private,* and *Grad.Rate*, since graduation rate is the main goal of the analysis; it will be the Y variable).

pc\_rest = prcomp(college\_rest[,-c(1,2,19)])

summary(pc\_rest)

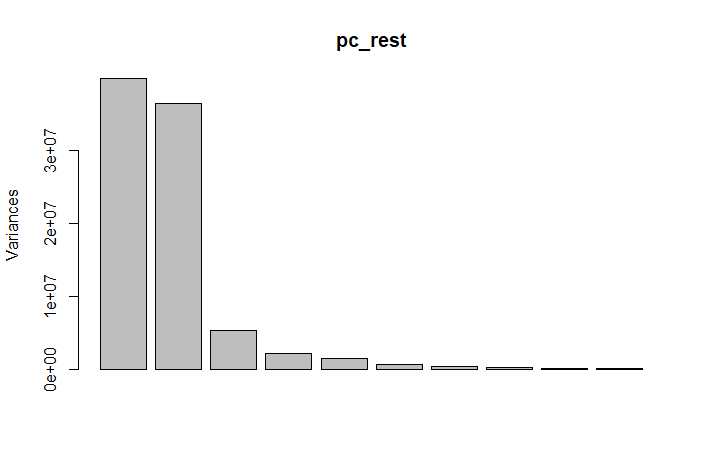


* 1. Plot the results of your PCA analysis. How many principal components would you suggest we use for this data?

head(pc\_rest$x)

PC8 = pc\_rest$x[,c(1:8)]

plot(pc\_rest)



I would consider 8 PC’s to capture all most all the variance.

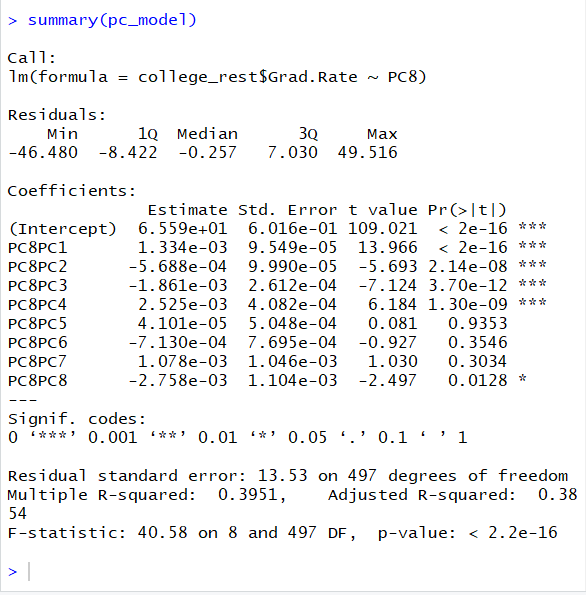
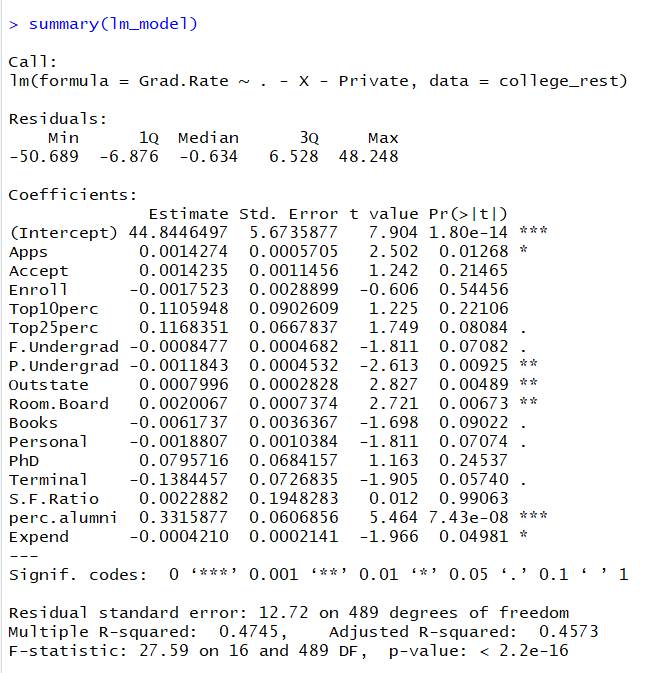
* 1. Run two linear regressions: (1) a standard linear regression using the same variables as you used for the PCA (that is, predict graduation rate using the numeric X variables from part (a)), and (2) a principal components linear regression using the number of principal components you chose in part (b).

lm\_model=lm(Grad.Rate~.-X-Private, data=college\_rest)

pc\_model=lm(college\_rest$Grad.Rate~PC8)

summary(lm\_model)

summary(pc\_model)



* + 1. Report either the adjusted or the AIC for both models.

Adjusted is 0.4573 for lm\_model (standard regression) and 0.38 for the PCA model.

* + 1. Which model appears to be better given the adjusted or AIC results? Does this surprise you?

The standard regression apprears to be better as it has a greater adjusted value.

No, it doesn’t surprise me because PCA captures the maximum variance but not all of it. The standard regression on the other hand has all the variance and has greater chance of getting a higher adjusted .

* + 1. Which of the two models is better for inference? Would your answer potentially change if the values of adjusted or AIC changed? Explain.

From adjusted we can say that the standard regression is better for inference here because it has a higher value.

This won’t potentially change even if the values changed because in the PCA model every PCA uses all the variables and we can’t find the acutal variables individually in the PCA model. Hence, it may give us inference only in PCA’s point of view but nothing in terms of the acutal variables.

In our standard regression, we have the p-values and coeffiecients for the variables used.

1. **(15 points) Bootstrap Sample (Single Tree)**
   1. Create a single bootstrapped sample from the *college\_rest* data.

num\_rest=nrow(college\_rest)

bootstrap\_sample=sample(seq(1,num\_rest),num\_rest,replace=T)

* 1. Using your single sample, run a single tree to predict *Grad.Rate* using all other variables except the name of the college*.* Is this a bagged tree or a random forest (of size 1)?

library(tree)

bag.tree=tree(Grad.Rate~.-X,data=college\_rest[bootstrap\_sample,])

summary(bag.tree)

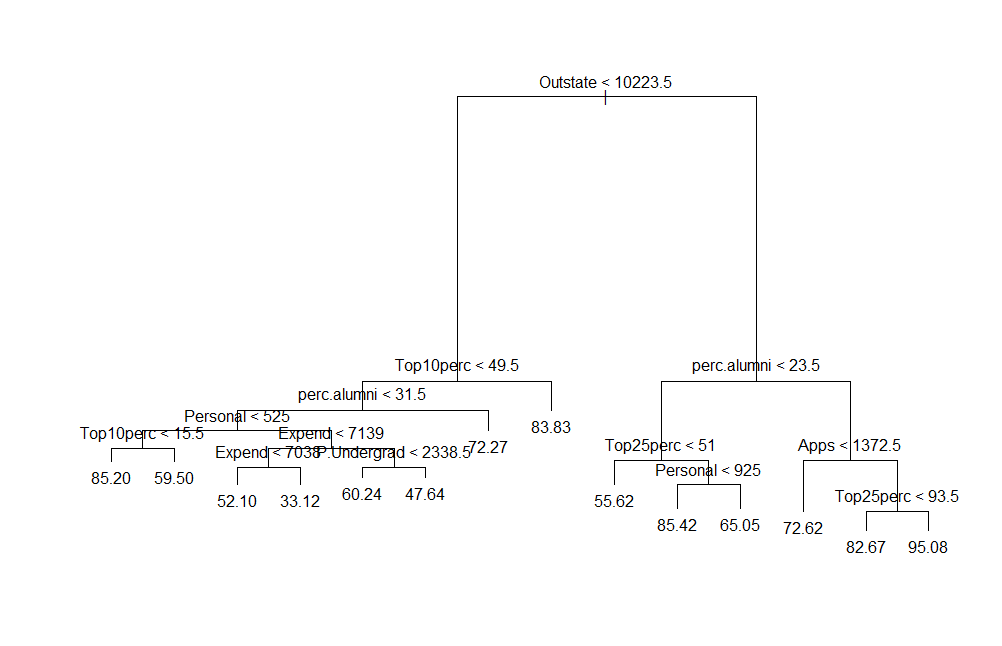
This is more likey a bagged tree but we can also say it’s a random Forest which uses all the variables instead of a subset for each split.

* 1. Plot your single tree. Does this appear to be a useful tree? Explain.

plot(bag.tree)

text(bag.tree,pretty=1)

This tree appears to be an useful tree because it predicts different mean values of Grad.Rate for different splits.



* 1. Use your single tree to predict *Grad.Rate* for the *college\_test* data. What is the RMSE for your predictions?

bag.tree\_preds <- predict(bag.tree,newdata=college\_test)

RMSE\_bag.tree\_test <- sqrt(mean((bag.tree\_preds-college\_test$Grad.Rate)^2))



1. **(15 points) Bagging: Use your *college\_rest* data to run a bagging procedure for 200 regression trees, again using *Grad.Rate* as the dependent variable and all other variables except the name of the college as independent variables.**
   1. Report the Variable Importance Plot for your bagging procedure. Given your single tree from Question 2, do the results surprise you? Why or why not?

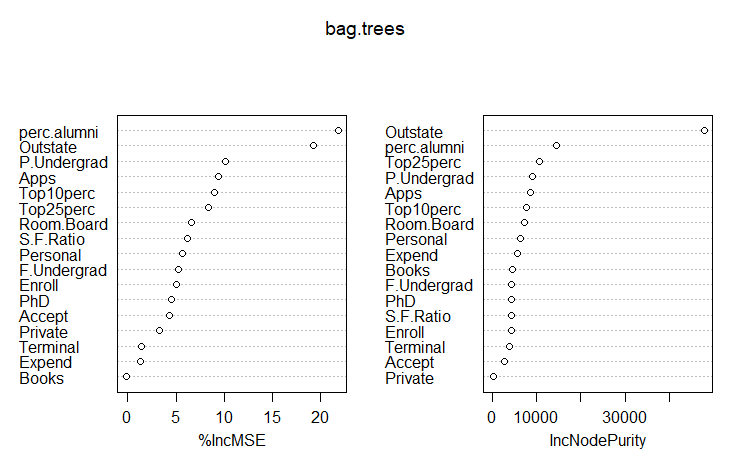
library(randomForest)

bag.trees=randomForest(Grad.Rate~.-X,data=college\_rest,ntree=200,mtry=17,importance=TRUE)

bag.trees

importance(bag.trees)

varImpPlot(bag.trees)



The results do not surprise me because the variables perc.alumni, Outstate, Apps, P.Undergrad, Top25perc and Top10perc seem to be important variables in the bagging model. Those variables also looked important in our single tree for earlier.

* 1. Use your bagging results to predict *Grad.Rate* on the training data. Report your training RMSE.

bagging\_preds\_rest=predict(bag.trees,newdata=college\_rest)

RMSE\_bag.trees\_rest <- sqrt(mean((bagging\_preds\_rest-college\_rest$Grad.Rate)^2))



* 1. Use your bagging results to predict *Grad.Rate* on the test data.
     1. Report your test data RMSE.

bagging\_preds\_test=predict(bag.trees,newdata=college\_test)

RMSE\_bag.trees\_test <- sqrt(mean((bagging\_preds\_test-college\_test$Grad.Rate)^2))



* + 1. Does there appear to be overfitting happening? Support your answer.

It looks like there is no overfitting here. Overfitting appears when the training RMSE is very low but there is high testing RMSE. Here, the testing RMSE is low as well. Also, bagging averages out the effect of all the trees resulting in low variance and low bias. Hence, we can say there is no overfitting here.

* + 1. Does bagging appear to show improvement over the single tree?

Yes, bagging appears to show an improvement because the it has testing RMSE decreased compared to a single tree.

1. **(15 points) Random Forest: Use your training data to run a random forest procedure for 200 regression trees using 4 random variables per split, again using *Grad.Rate* as the dependent variable and all other variables except the name of the college as independent variables.**
   1. Report the Variable Importance Plot for your random forest procedure. Does it appear that any of the variables in the data set is significantly affecting our results more than the others? Support your answer.

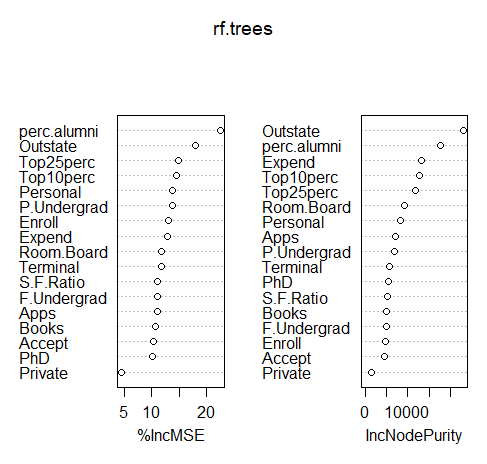
library(randomForest)

rf.trees=randomForest(Grad.Rate~.-X,data=college\_rest[bootstrap\_sample,],ntree=200,mtry=4,importance=TRUE)

rf.trees

importance(rf.trees)

varImpPlot(rf.trees)



Yes, perc.alumni and Outstate variables are significantly affecting our results more than the others. It means if these variables were not considered, the training RMSE would have been higher.

* 1. If we increase the number of variables per split from 4 to 15, would you expect the Variable Importance Plot to change? Support your answer.

I don’t expect the Variable Importance Plot to change because it could still be picking the most important ones only from the 15 and they are most likely to be the same variables again from the 4.

* 1. Use your random forest results to predict *Grad.Rate* for the test data. What is your RMSE here? Does random forest appear to be a better option than bagging for this problem?

rf\_preds\_test=predict(rf.trees,newdata=college\_test)

RMSE\_rf.trees\_test <- sqrt(mean((rf\_preds\_test-college\_test$Grad.Rate)^2))



No, random forest doesn’t appear to be a better option than bagging for this problem because it has a higher RMSE value on the testing data.

1. **(15 points) Boosting: Use your training data to run a boosting procedure using GBM for 200 regression trees, again using *Grad.Rate* as the dependent variable and all other variables except the name of the college as independent variables. Note that *Grad.Rate* is a continuous, numeric variable; your distribution should be normal (gaussian)!**
   1. Report the Relative Influence Plot for your boosting procedure.

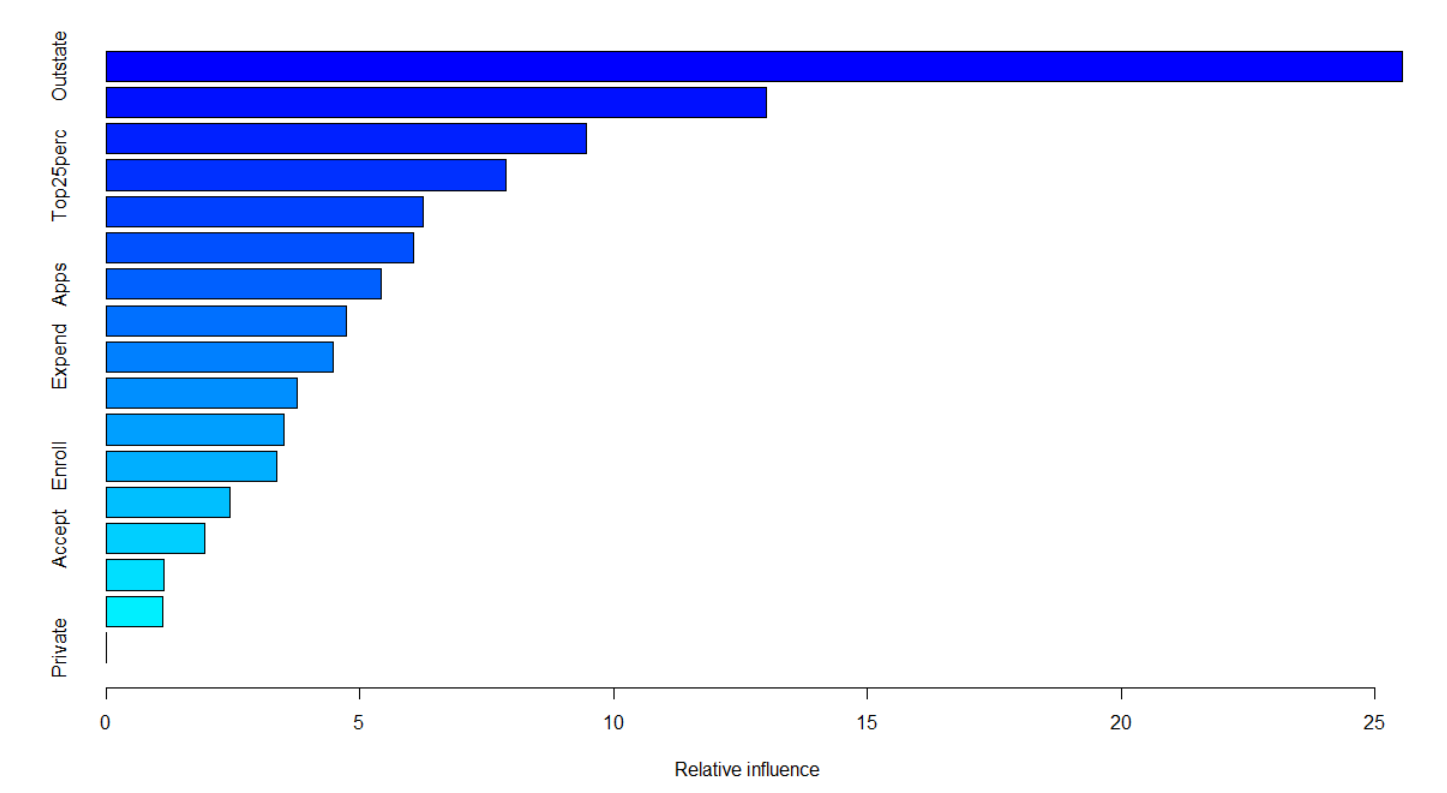
library(gbm)

college\_rest$Private = as.numeric(college\_rest$Private)

college\_test$Private = as.numeric(college\_test$Private)

boost.college=gbm(Grad.Rate~.-X,data=college\_rest,distribution="gaussian",n.trees=200)

summary(boost.college)



* + 1. Is this plot consistent with the Variable Importance Plots from 3(a) and 4(a)? Support your answer.

Yes, this plot appears to be consistent with Variable Importance Plots from 3(a) and 4(a) because Outstate and perc.alumni seem to be the variables with most Relative influence. (These were the most important variables earlier)

However, we must not forget that Relative influence says how much influence a particular variable had in predicting misclassified points; variable importance is how often a variable was used in the model.

* + 1. Would you expect this plot to be consistent with the Variable Importance Plots from 3(a) and 4(a)? Support your answer.

No, Relative influence says how much influence a particular variable had in predicting misclassified points; variable importance is how often a variable was used in the model

* 1. Use your boosting results to predict Rating for the test data, making sure to use the full 200 trees. What is your RMSE here? Does boosting appear to be a better option than bagging and/or random forest for this problem?

boost\_preds\_test = predict(boost.college,newdata=college\_test,n.trees=200)

RMSE\_boost\_test <- sqrt(mean((boost\_preds\_test-college\_test$Grad.Rate)^2))



Yes, boosting appears to be a better option than bagging and/or random forest for this problem because it has the lowest RMSE on the testing data.

1. **(15 points) Repeat your ensemble methods from Questions 3 through 5 with 1000 trees instead of 200. (Do not re-report variable importance/relative influence plots.)**

bag.trees\_1000=randomForest(Grad.Rate~.-X,data=college\_rest,ntree=1000,mtry=17,importance=TRUE)

bagging\_preds\_test\_1000=predict(bag.trees\_1000,newdata=college\_test)

RMSE\_bag.trees\_test\_1000 <- sqrt(mean((bagging\_preds\_test\_1000-college\_test$Grad.Rate)^2))

rf.trees\_1000=randomForest(Grad.Rate~.-X,data=college\_rest,ntree=1000,mtry=4,importance=TRUE)

rf\_preds\_test\_1000=predict(rf.trees\_1000,newdata=college\_test)

RMSE\_rf.trees\_test\_1000 <- sqrt(mean((rf\_preds\_test\_1000-college\_test$Grad.Rate)^2))

college\_rest$Private = as.numeric(college\_rest$Private)

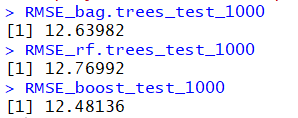
college\_test$Private = as.numeric(college\_test$Private)

boost.college\_1000=gbm(Grad.Rate~.-X,data=college\_rest,distribution="gaussian",n.trees=1000)

boost\_preds\_test\_1000 = predict(boost.college\_1000,newdata=college\_test,n.trees=1000)

RMSE\_boost\_test\_1000 <- sqrt(mean((boost\_preds\_test\_1000-college\_test$Grad.Rate)^2))

* 1. Calculate the new RMSE values on the test data with the three extended ensemble methods. Have the relative rankings of your ensemble methods changed? (That is, is there a new method you prefer based on these results?)



The relative rankings of the ensemble methods still remains the same. Boosting has the least RMSE on the testing data followed by bagging and random forest respectively.

* 1. Consider your results from part (a) and imagine you have a brand new data set of colleges and their 2018 data.
     1. Why might you prefer to use the ensemble methods with 200 trees instead of the ensemble methods with 1000 trees to predict for 2018?

Since from our experience above, since the relative ranking of the ensemble methods remain the same and also the testing RMSE’s haven’t changed drastically, I would use just 200 trees to save time and computational effort.

* + 1. Why might you prefer to use the ensemble methods with 1000 trees instead of the ensemble methods with 200 trees to predict for 2018?

Since, this is a new data set we can’t gurantee that the relative rankings of the ensemble methods will remain the same for 1000 trees as for 200 trees.

Also, we may get better testing RMSE incase of 1000 trees compared to just 200 trees.