homework5

Sunny Lee

3/9/2021

1)

```
library("ISLR")
library(tree)
```

```
## Warning: package 'tree' was built under R version 4.0.4
```

```
data("College")
```

2)

a)

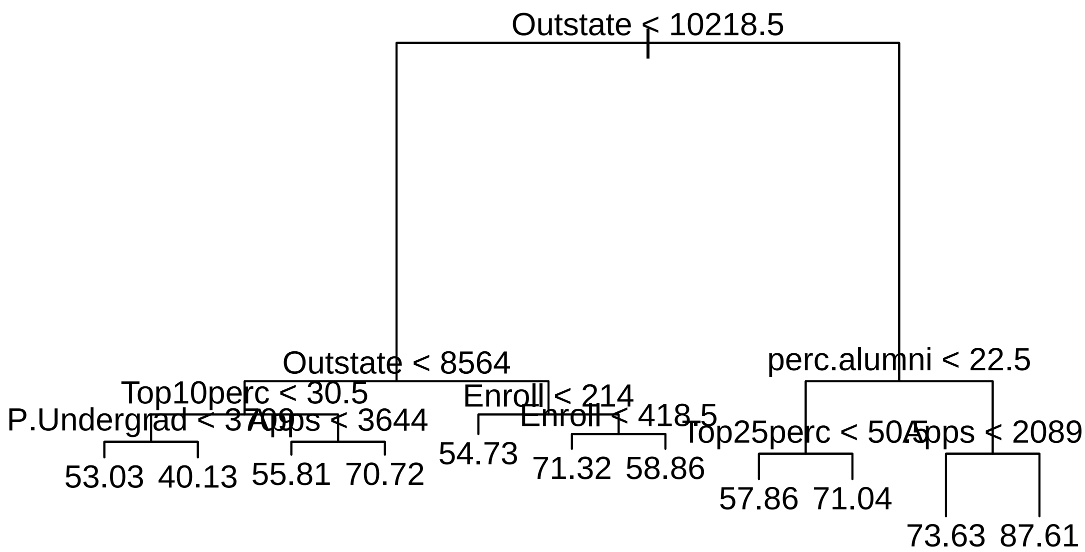
```
train_size <- floor(.75 * nrow(College))
set.seed(1)
idx <- sample(seq_len(nrow(College)), size = train_size)
train <- College[idx, ]
test <- College[-idx, ]
```

b)

```
tree <- tree(Grad.Rate~., College, subset = idx)
summary(tree)
```

```
##
## Regression tree:
## tree(formula = Grad.Rate ~ ., data = College, subset = idx)
## Variables actually used in tree construction:
## [1] "Outstate"    "Top10perc"   "P.Undergrad" "Apps"        "Enroll"
## [6] "perc.alumni" "Top25perc"
## Number of terminal nodes:  11
## Residual mean deviance:  148.6 = 84840 / 571
## Distribution of residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -43.03000  -6.73300  -0.02532   0.00000   7.38200  46.97000
```

```
plot(tree)
text(tree, pretty = 0)
```




```r
tree.pred <- predict(tree, test)
mean((tree.pred - test$Grad.Rate)^2)
```

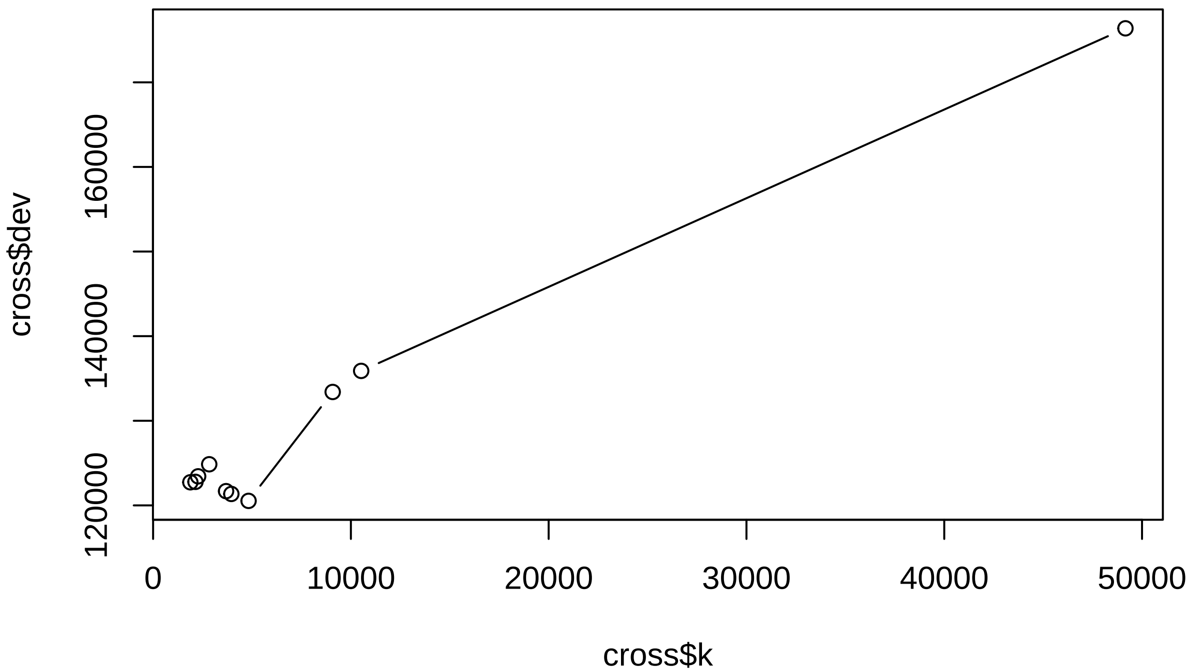
```
## [1] 213.0213
```

From the summary of the tree above, we find that the tree we have obtained has 11 terminal nodes, and has used Outstate, Top10perc, P.Undergrad, Apps, Enroll, perc.alumni and Top25perc as explanatory variables.

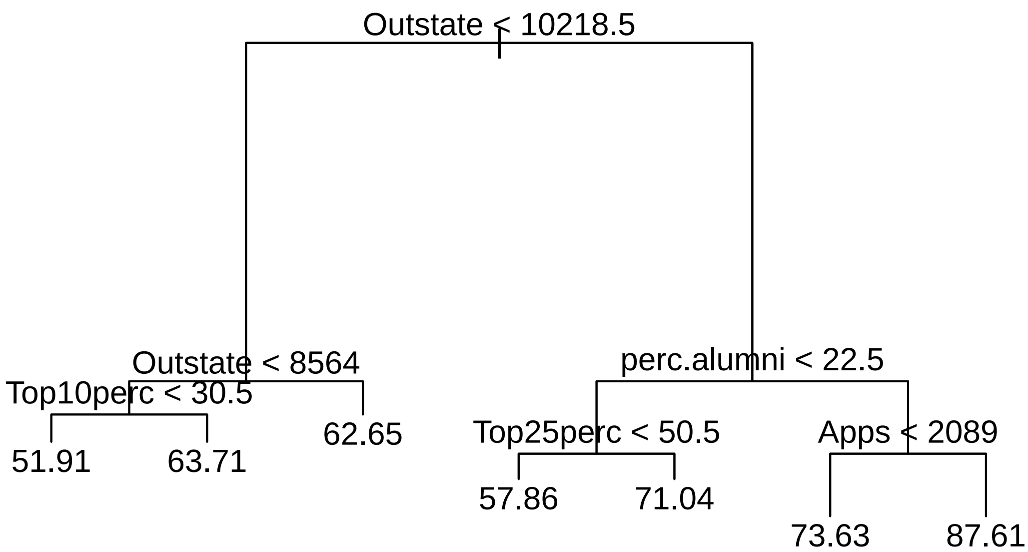
c)

```r
cross <- cv.tree(tree)
plot(cross$size, cross$dev, type = "b")
```


```
plot(cross$k, cross$dev, type = "b")
```




```
prune.tree <- prune.tree(tree, best = 7)
plot(prune.tree)
text(prune.tree, pretty = 0)
```




```
prune.tree.pred <- predict(prune.tree, test)
mean((prune.tree.pred - test$Grad.Rate)^2)
```

```
## [1] 189.5238
```

From the cross validation plot above, we see that our optimal tree size is 7 as the dev is at its lowest when the number of terminal nodes is 7. Thus, we prune our tree down to 7 terminal nodes and then calculate the MSE again with our new tree. After calculating the new MSE, we see that the pruned tree had a lower MSE than the original 11 terminal node tree.

d)

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.0.4
```

```
## randomForest 4.6-14
```

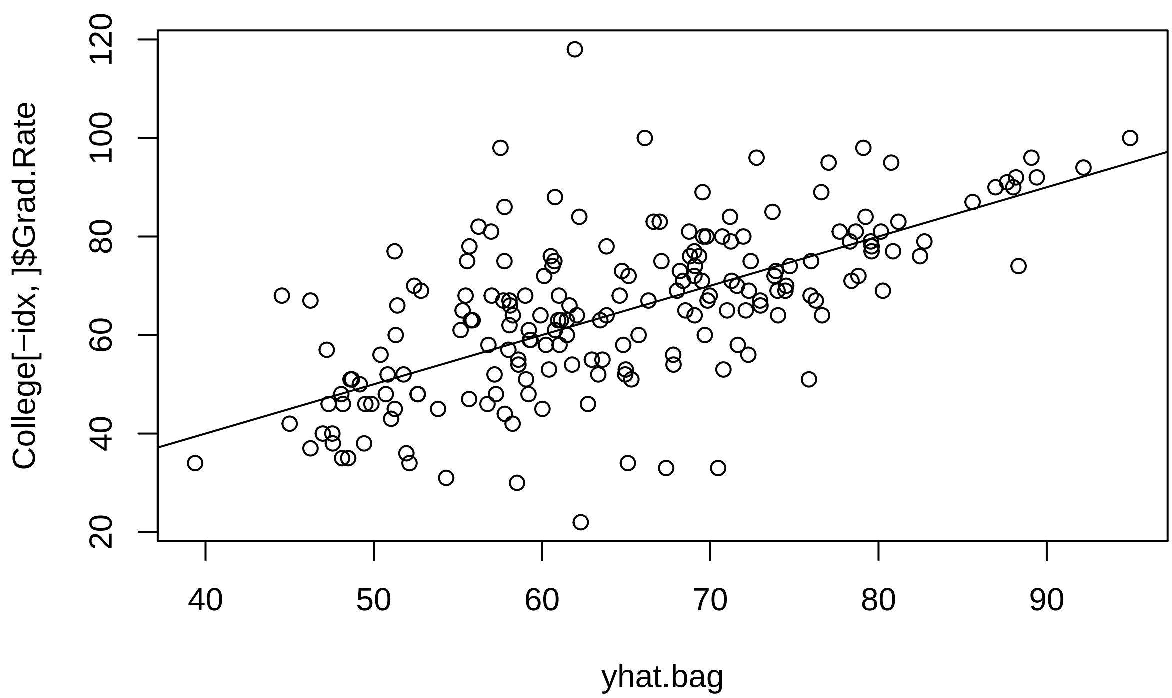
```
## Type rfNews() to see new features/changes/bug fixes.
```

```
set.seed(1)
bag.college <- randomForest(Grad.Rate~., College, subset = idx, importance = TRUE, mtry = 17)
bag.college
```

```
##
## Call:
##  randomForest(formula = Grad.Rate ~ ., data = College, importance = TRUE,      mtry = 17, subset = i
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 17
```

```
##
##           Mean of squared residuals: 172.2645
##                     % Var explained: 42.78
```

```
yhat.bag <- predict(bag.college, newdata = College[-idx, ])
plot(yhat.bag, College[-idx, ]$Grad.Rate)
abline(0, 1)
```




```
mean((yhat.bag-College[-idx, ]$Grad.Rate)^2)
```

```
## [1] 167.1067
```

```
importance(bag.college)
```

```
##                %IncMSE IncNodePurity
## Private       3.723884      282.5198
## Apps         26.375943    10199.7348
## Accept        8.267985     3928.2601
## Enroll       13.163469     5737.4136
## Top10perc    12.858764     8422.8778
## Top25perc    14.221070     9797.1491
## F.Undergrad   5.866050     5549.9020
## P.Undergrad 16.310103     8754.9022
## Outstate     34.817214    60528.0338
## Room.Board    8.182890     6558.1404
## Books        -3.640854     4689.4988
## Personal      6.340703     6826.5140
## PhD           7.866133     4590.8111
```

```
## Terminal     7.243832     4774.1160
## S.F.Ratio   10.011419     5522.4573
## perc.alumni 35.172778    17240.6581
## Expend       6.272947     7589.8903
```

After fitting our data with the bagging technique, we get an MSE which is lower than both of our previous regression trees: 167.1067. Using the importance() function, we can conclude the three most important variables in our random forest are perc.alumni, Outstate and Apps