STA363Lab6SecA

Jean Ji, Phoebe Yan & Rickey Huang

4/6/2021

Question 1

Trees begin with a split on a single feature. Suppose we decided to consider splitting on whether or not the school is a private school. Explain in 1-2 sentences how you would use this feature to create one split, and how you would use the splitting rule to move rows into leaves.

We would use the splitting rule of "Private == Yes" to create one split which splits the rows into two leaves. The leaf 1 would store the rows whose variable *"Private"* is "Yes", and the leaf 2 would store the rows whose variable *"Private"* is "No".

Question 2

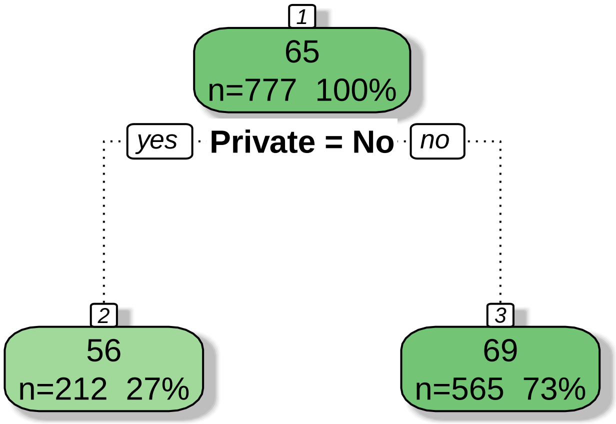
Without using rpart to build the tree, find the training RSS we would get if we split on whether or not a school is a private school. Show your code.

```r
# Store the number of rows in to n
n <- nrow(college)
# Store the exploratory and response variables
X <- college$Private
Y <- college$Grad.Rate
# Assign rows to leaves
leaf1 <- which(X=="Yes")
leaf2 <- c(1:n)[-leaf1]
# Compute the means and use them for prediction
mean1 <- mean(Y[leaf1])
mean2 <- mean(Y[leaf2])
preds <- rep(0, n)
preds[leaf1] <- mean1
preds[leaf2] <- mean2
# Calculate the training RSS
RSS <- sum((Y - preds)^2)
#Return the training RSS
RSS
```

```
## [1] 203101.6
```

Question 3

Now, using the rpart code, create a tree using only Private as a feature. Call tree Tree1. Show a visualization of your tree as your answer.



Tree1: Regression Tree Using only the Private feature

Question 4

Based on your tree, what percent of your training data comes from public schools?

Based on Tree1, 27% of the training data comes from public schools.

Question 5

Based on your tree, what graduation rate would you predict for a public school?

The predicting graduation rate for a public school is 56%.

Question 6

Fit a least squares linear regression model for graduation rate, using whether or not a school is a private school as a feature. Call this model LSLR1. Write out the fitted regression line.

Only using the feature *"Private"*, the fitted LSLR model has the regression line of $\widehat{Grad.Rate} = 56.042 + 12.956 PrivateYes$

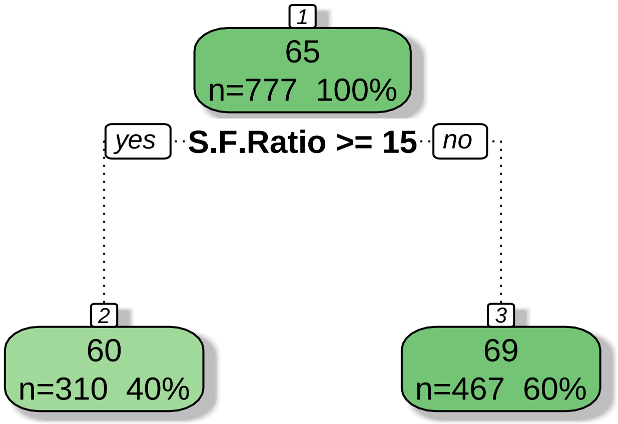
Question 7

Based on the LSLR model, what graduation rate would you predict for a public school? Keeping in mind that in the visualization our trees round to the nearest whole number, how do these predictions compare to those you made from the tree?

Based on the regression line of the LSLR model, the predicting graduation rate for a public school would be 56.042. The predictions from both the tree model and the LSLR model is almost the same, since the prediction from the tree model is also 56.

Question 8

Create a tree using only student faculty ratio as a feature. Use the maxdepth = 1 stopping criterion to make sure that for the moment, the tree only has one split. If you don't do this, the tree will keep growing, and for now, we only want one split. Call tree Tree2, and show a visualization of your tree as your answer.



Tree2: Regression tree using only the S.F.Ratio feature

Question 9

Based on your tree, what graduation rate would you predict for a school with a student faculty ratio of 10 (1 student to 10 faculty)?

Based on Tree2, the predicting graduation rate for a school with a student faculty ratio of 10 would be 69%.

Question 10

Fit a least squares linear regression model for graduation rate, using student faculty ratio as the only feature. Call this model LSLR2. Write out the fitted regression line.

Only using the *"S.F.Ratio"*, the LSLR model fitted has the regression line of $\widehat{Grad.Rate} = 84.2168 - 1.3310S.F.Ratio$

Question 11

Based on your LSLR model, what graduation rate would you predict for a school with a student faculty ratio of 10 (1 student to 10 faculty)? How does this compare to what you get from a tree?

Based on the LSLR model we just built, the graduation rate we would predict for a school with a student faculty ratio of 10 is 70.9068. This result is different from the prediction we get from the tree model which is 69.

Question 12

Find the test MSE for your tree and for your LSLR model with student faculty ratio as a feature. Based on test metrics, which model would you choose and why?

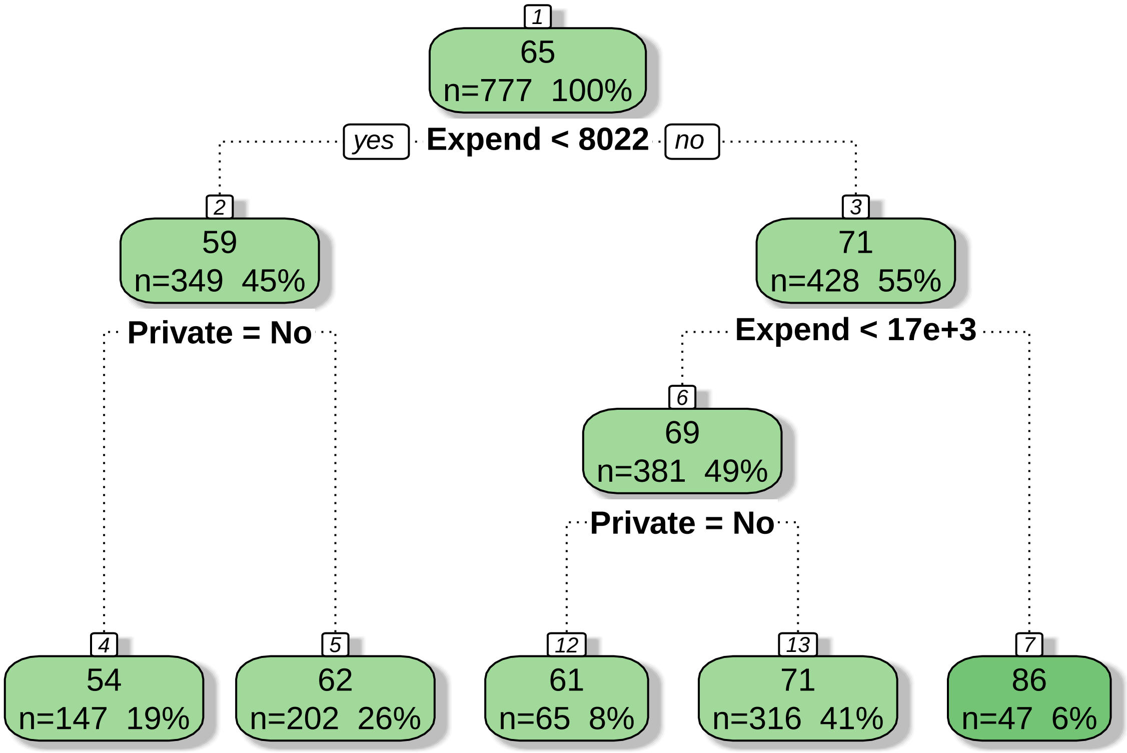
Using the 10-fold cross validation technique, we compute the test MSE for our LSLR model which is 268.1957.

Since the test MSE for the tree model can be computed by using $Xerror \times RNE$, test MSE for our tree model is $294.69 \times 0.9632 = 283.8454$.

Comparing the test MSE of two models, we would choose the LSLR model, since it has a lower test MSE, which means the prediction for the test data are closer to the real data for the LSLR model than the tree model.

Question 13

Create a tree using student faculty ratio, whether or the not the school is a private school, and expenses on each student as features. Call the tree Tree3, and show a visualization of your tree as your answer.



Tree3: Regression Tree using three variables

Question 14

Type the code ?rpart.control into a chunk, and hit play, and then put a # in front of the code. What will pop up is the R help page. This page shows all of the stopping criteria you can choose to use when growing a tree. It also shows (in the code at the top) the default stopping criteria that R uses if we don't specify our own. What is the default number of rows that have to be in a leaf in order for it to split?

The default number of rows that have to be in a leaf in order for it to split is 20.

Question 15

Which feature was able to give us the largest reduction in training RSS in one split?

Based on Tree3, we know that Expend was able to give the largest reduction in training RSS in one split.

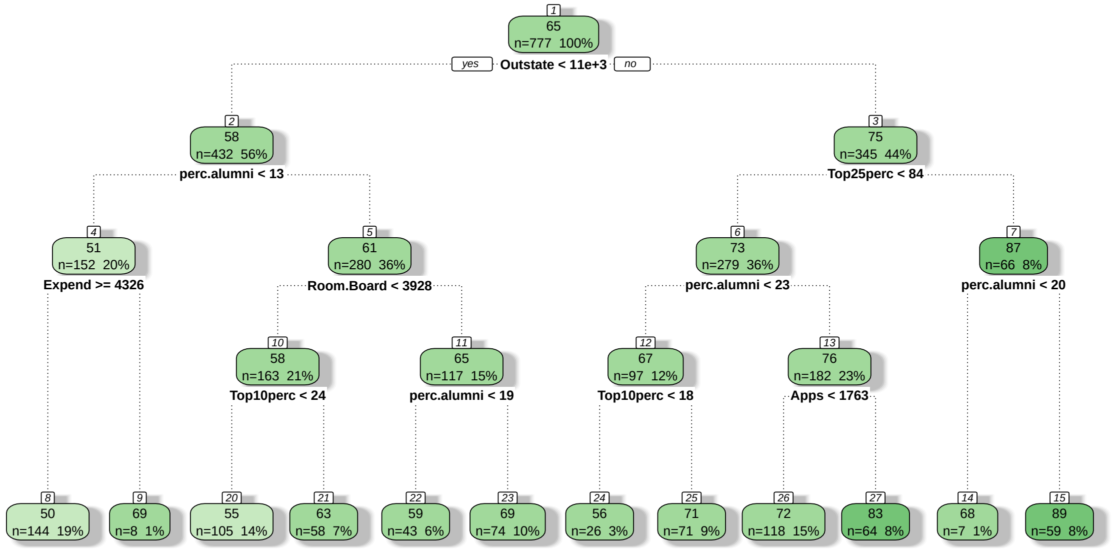
Question 16

Based on our tree, what is the predicted graduation of a public school that spends about 12,000 US dollars on each student in terms of school expenses, with a student faculty ratio of 20 (meaning 1:20)?

The predicted graduation rate would be 61%.

Question 17

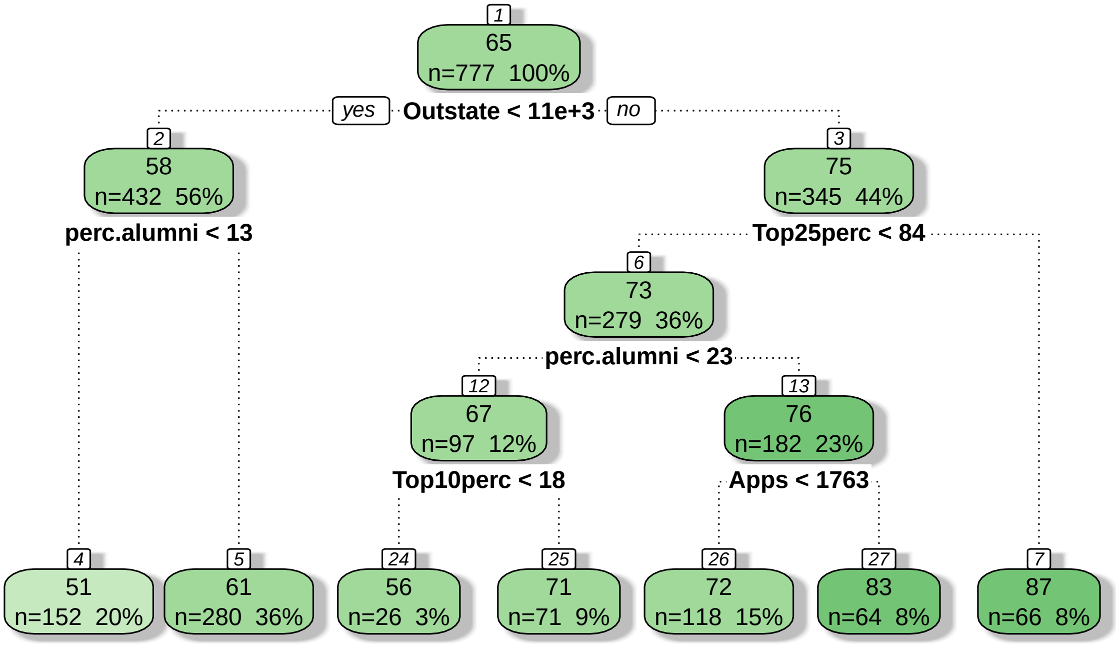
Create a tree using all of the features (except university name). Call the tree TreeAll, and show a visualization of your tree as your answer.



Tree4: Regression Tree using all variables

Question 18

Prune your tree. Call the final tree TreeFinal, and show a visualization of your tree as your answer.



Tree5: Pruned Regression Tree using all variables

Question 19

With your pruned tree, how many leaves did you remove from the original tree? Hint: It is okay in practice if the answer is 0, it just means that the stopping rules already gave us a tree that predicted (relatively!) well.

Comparing the pruned tree with the original tree with all feature, we removed 5 leaves from the original tree.

Question 20

What is the predicted test RMSE of your final pruned tree?

The predicted test RMSE of the final pruned tree is 14.11611.