## Assignment 1 Part 2 (total 250 points)

## Problem 5-8 are for Topic 3.1: Tree: Build and Prune

**Problem 5.** **(70 points total): Fit a classification tree**

Use **Carseats** dataset, follow Lab 8.3.1. With the dataset, we wish to predict the sales of car seats on the basis of various statistics associated with the sale data in the previous year. We choose Tree methods.

1. (5 points) Create a categorical variable “High” to replace “Sales”, and add the variable to the Carseats dataset. Sales>8 is considered as “yes” for “High”; Sales<=8 is considered as “no” for High
2. (10 points) Call library(tree) and use the tree() function. With the whole dataset as training data, fit a classification tree in order to predict “High” using all variables but “Sales.” What are the variables used as internal nodes in the tree? How many terminal nodes are there in the tree? What is the training error rate of the tree? What is the tree deviance? What is its entropy?

A computer screen shot of a blue screen

Description automatically generated

The variables used as internal nodes in the tree are shelveloc, price, income, compprice, population, advertising, age, us. There are 27 terminal nodes in the tree. The deviance is 0.4575. The error rate is 0.09. The entropy is related to the deviance.

1. (5 points) Plot the above tree and include node labels.

A diagram of a company

Description automatically generated

\*\*\* this tree is overfit. Population < 207.5 node leads to two terminal nodes with the same conclusion (‘yes’). Unnecessary added complexity. \*\*\*

1. (10 points) Read the tree and explain how it split.
   1. What is the most important factor that influences sale? What is the second most important factor, why?

The most important factor is shelf location (good vs bad/medium) and the second most important is the price. We read the tree from top to bottom and we can see the significance of each factor based on where it is in the tree.

* 1. How many observations of the data are in good shelve location and with price less than 135?

There are 68 observations of data in good shelf location and price less than 135.

* 1. Given an overall interpretation of the decision tree, e.g., how to follow the tree to determine if the car seat sale will be high or low.

Starting at the root node we look at whether or not the shelve location is bad/medium or good. If it’s bad/medium, we check if price is < 92.5 or >= 92.5. If it’s good, we check if price is < 135 or >= 135. We keep following the nodes all the way down to the terminal node where we get the final determination of whether or not the car seat sale will be high or low. When we get to each node, we are splitting the data and getting various subsets until we hit the terminal node and get a final answer.

1. (10 points) Split the data into training dataset and test dataset with set.seed(2).
   1. Train the tree with the training data.
   2. Read the tree and explain how it split, e.g., how to follow the tree to determine if the carseat sales will be high or low.

A diagram of a number of people

Description automatically generated

If price<96.5, we should go left and population < 414(left side), otherwise right shelve loc(right side)and we will continue in this way: If the statement is True, we will go left, otherwise to the right until we reach a terminal node.

* 1. Is there any difference between this tree and the tree we build from the whole dataset? Compare their performances in terms of training error rate and deviance.

A computer screen shot of a blue screen

Description automatically generated

The whole dataset built a tree with a deviance of 373 and an error rate of 0.09. The training tree had a deviance of 179 and an error rate of 0.115.

1. (10 points) Run the prediction with the test dataset with predict() function. Compare the test result with the observations with table() function and calculate the error rate of the test data.

A blue screen with white text

Description automatically generated

The test error rate is 0.23

1. (10 points) Now consider pruning the tree. Use misclassification rate to guide the cross-validation and pruning process. set.seed(7) to conduct cross validation for the pruning process.
   1. Print out the lists of the size of the subtrees, the corresponding number of misclassification errors, and α. Which size of tree should be chosen and what are its corresponding number of misclassification error and α

A screenshot of a computer

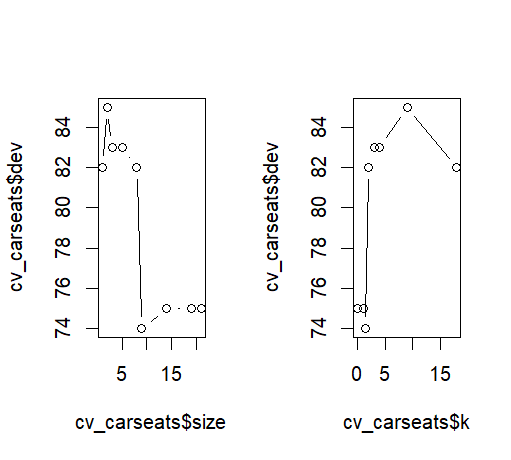
Description automatically generated

The best tree has a size of 9 with a misclassification error of 74 and α of 1.4.

* 1. plot the error number as a function of both size and α of the best subtrees.

A graph of different sizes and lines

Description automatically generated with medium confidence



1. (10 points) Apply the prune.misclass() function to prune the tree to obtain the nine-node tree.
2. Print the pruned tree.

A diagram of a family tree

Description automatically generated

1. Run the prediction with the test dataset with predict() function

A blue background with white text

Description automatically generatedA blue screen with white text

Description automatically generated

1. Compare the performances of the tree in V and the pruned tree (on the test dataset), which one performs better? Why?

The misclassification error rate in the previous tree is lower than the pruned tree (0.115 vs 0.175) This is because the previous tree is overfit. The pruned tree has less complexity. However, the pruned tree has a smaller deviance. The pruned tree is better because it is simpler and is not overfit on the training data.

**Problem 6.** **(70 points total): Fit a classification tree**

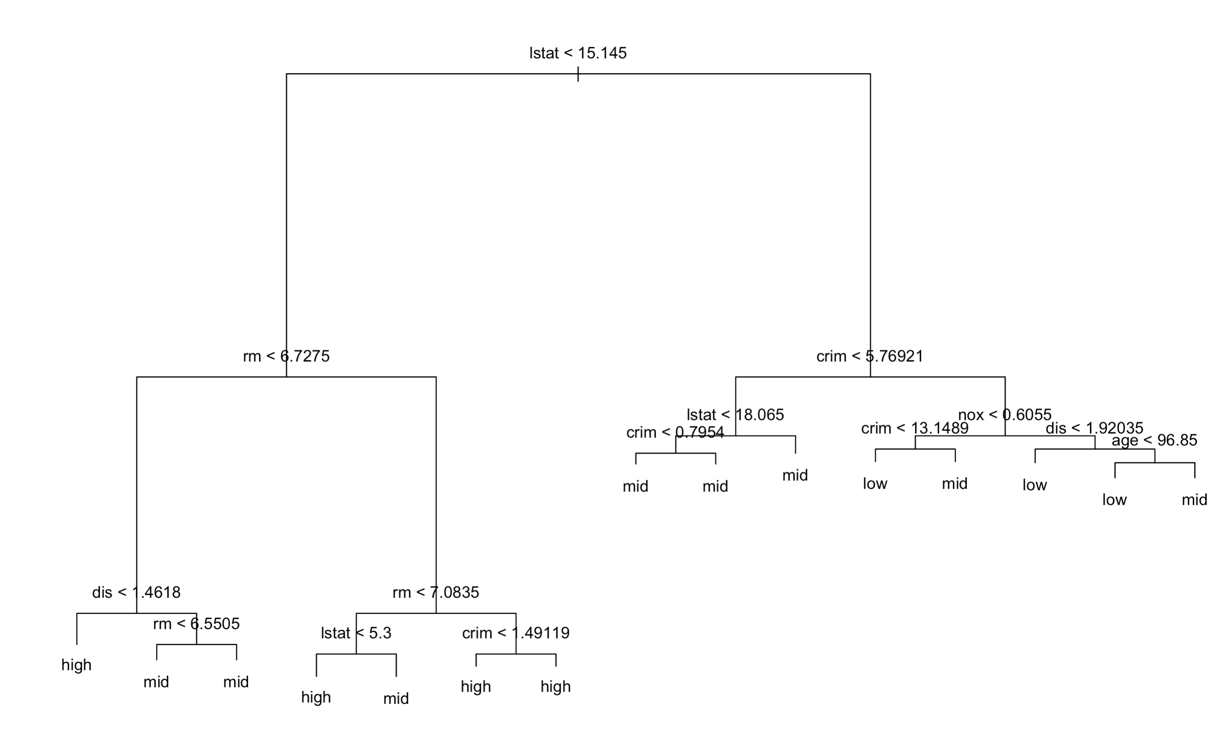
Use **Boston** dataset. With the dataset, we wish to predict the median house prices of greater Boston area on the basis of various statistics associated with the house price. We choose Tree methods. “medv” is the response variable.

1. (5 points) Create a categorical variable “Price” to replace “medv.”, and add the variable to the Boston dataset. medv >30 is considered as “high” for “Price”; medv <=15 is considered as “low” for “Price”; 15< medv <=30 is considered as “mid” for “Price”
2. (10 points) Call library(tree) and use the tree() function. With the whole dataset as training data, fit a classification tree in order to predict “Price” using all variables but “medv.” What are the variables used as internal nodes in the tree? How many terminal nodes are there in the tree? What is the training error rate of the tree? What is the tree deviance? What is its entropy?

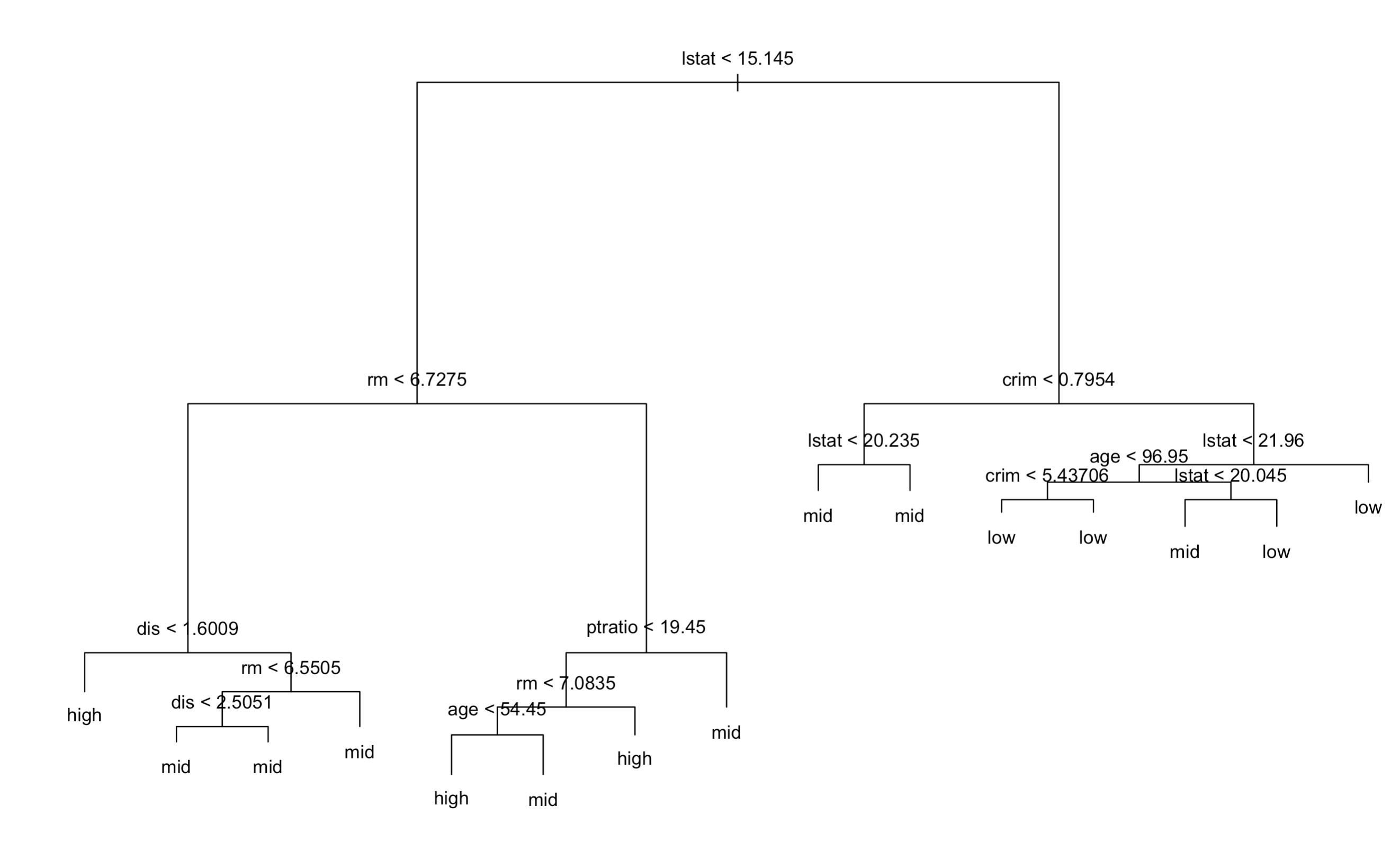
A blue screen with white text

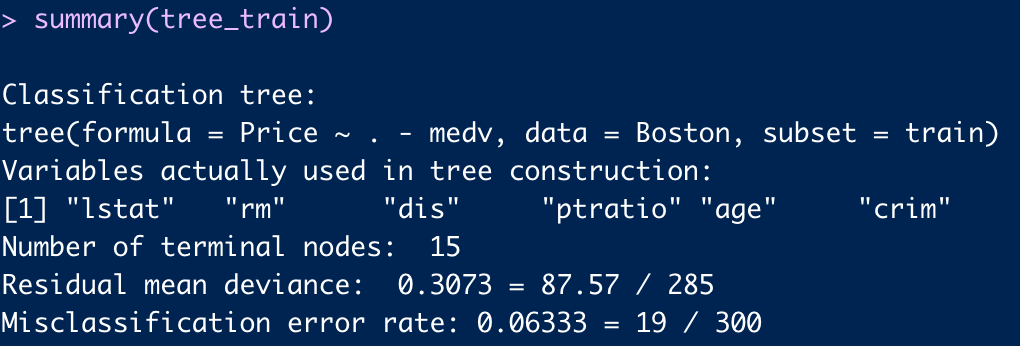
Description automatically generated

1. (5 points) Plot the above tree and include node labels.



1. (10 points) Read the tree and explain how it split.
   1. What is the most important factor that influences Price? What is the second most important factor, why?
      1. The most important factor that influences price is LSTAT, and the next one is RM. We can tell based on the tree that the largest branch is LSTAT and the second largest branch is RM(which tells us which variables are more influencial)
   2. Given an overall interpretation of the decision tree, e.g., how to follow the tree to determine if an area’s Price will be high, mid, or low.
2. (10 points) Split the data into training dataset and test dataset with set.seed(5).
3. Train the tree with the training data.





1. Read the tree and explain how it split, e.g., how to follow the tree to determine if an area’s median house price will be high, mid, or low.
   1. If you want to determine the median price, look down a branch, for example if lstat < 15.145, look at RM and if it is less than 6.7 look at dis. If dis is less than 1.6 than you have reached you leaf node of high. Follow this same process for different leaf-branch combinations till you determin mid and low.
2. Is there any difference between this tree and the tree we build from the whole dataset? Compare their performances in terms of training error rate and deviance.

The tree built from the whole dataset had a worse performance when evaluated by the training error rate and the deviance. The tree built on training data had an error rate if 0.05 and deviance of 0.28 while the tree built on the whole dataset had an error rate of 0.06 and a deviance of 0.3.

A blue screen with white text

Description automatically generated

A diagram of a tree

Description automatically generated

1. (10 points) Run the prediction with the test dataset with predict() function. Compare the test result with the observations with table() function and calculate the error rate of the test data.

A screenshot of a computer program

Description automatically generated

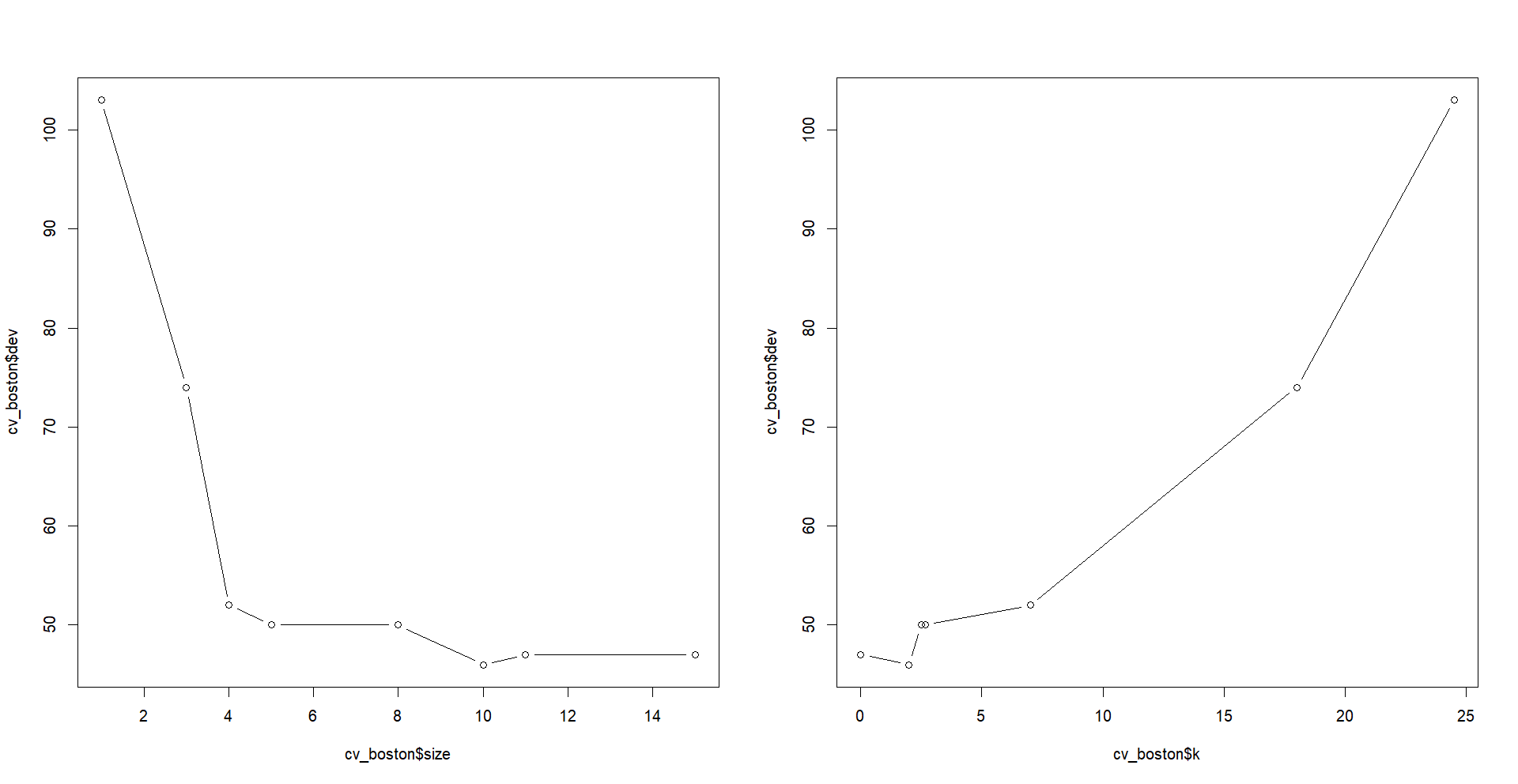
1. (10 points) Now consider pruning the tree. Use misclassification rate to guide the cross-validation and pruning process. set.seed(7) to conduct cross validation for the pruning process.
   1. Print out the lists of the size of the subtrees, the corresponding number of misclassification errors, and α. Which size of tree should be chosen and what are its corresponding number of misclassification error and α

A blue screen with white text

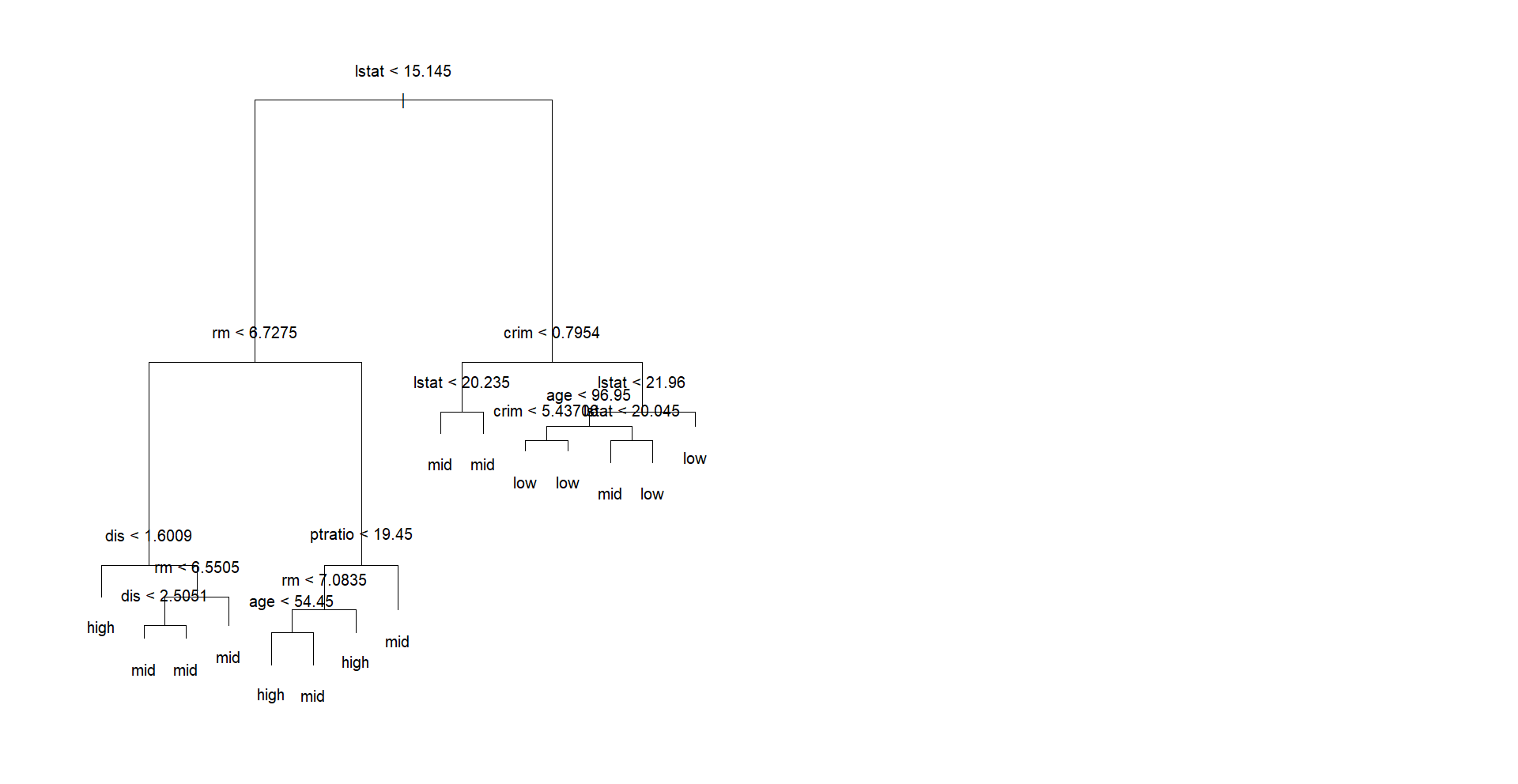
Description automatically generated

A tree of size 14 should be chosen because it has the lowest misclassification of 34 and corresponding α of -inf.

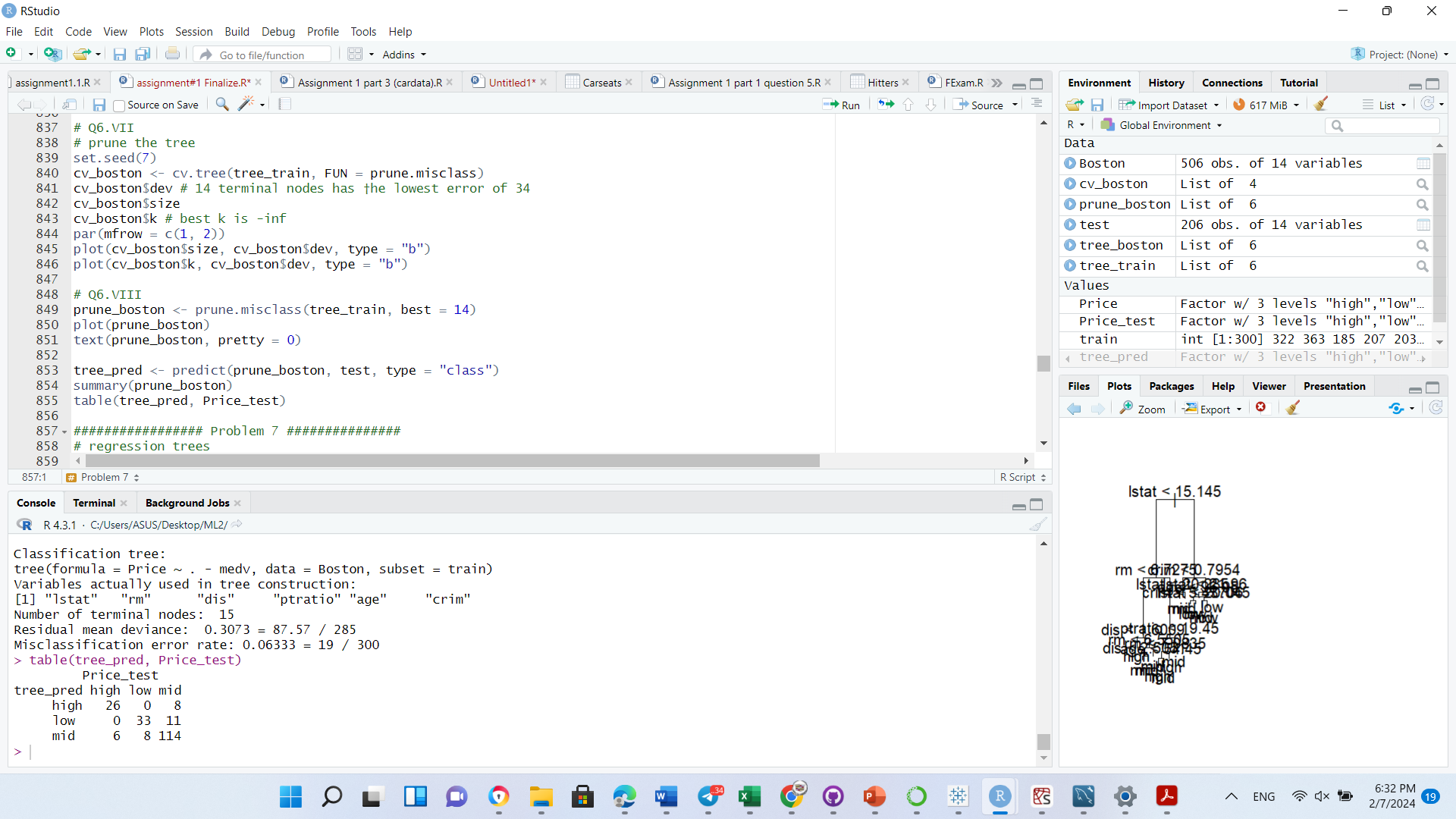
* 1. plot the error number as a function of both size and α of the best subtrees.



1. (10 points) Apply the prune.misclass() function to prune the best subtree Print the pruned best subtree.



1. Run the prediction with the test dataset with predict() function



1. Compare the performances of the tree in V and the pruned tree (on the test dataset), which one performs better? Why?

The trees are the same from V and the pruned tree. The previous tree already had 15 nodes which according to cross-validation is the best size. This is because the optimal size was determined by the original tree and confirmed with cross validation.

**Problem 7.** **(50 points total): Fit a regression tree**

Use **Boston** dataset, follow Lab 8.3.2. With the dataset, we wish to predict the median house prices of greater Boston area on the basis of various statistics associated with the house price. We choose Tree methods. “medv” is the response variable.

1. (10 points) Call library(tree) and use the tree() function. With the whole dataset as training data, fit a regression tree in order to predict “medv” using all variables. What are the variables used as internal nodes in the tree? How many terminal nodes are there in the tree? What is the tree deviance? Print out the tree.

A blue screen with white text

Description automatically generated

The variables used as the internal nodes in the tree are rm, lstat, dis, crim, and pratio. There are 9 terminal nodes in the tree. The deviance is 13.55.

A diagram of a tree

Description automatically generated

1. (15 points) separate the data into training subset and test subset with setseed(1). fit a regression tree in order to predict “medv” using all variables with the training data. What are the variables used as internal nodes in the tree? How many terminal nodes are there in the tree? What is the tree deviance? Print out the tree. Are there any difference between this tree and the tree from the whole dataset?

A blue screen with white text

Description automatically generated

The variables used as the internal nodes in the tree are rm, lstat, crim, and age. There are 7 terminal nodes in the tree. The deviance is 10.38.

A diagram of a tree

Description automatically generated

The tree using the training data is different from the previous tree because it has less terminal nodes and a smaller deviance.

1. (10 points) Now consider pruning the tree. use the cv.tree() function to see whether pruning the tree will improve performance. Plot the deviance as a function of tree size. Does pruning improve performance?
   1. Not by much based on MSE

A graph with a line

Description automatically generated

1. (5 point) Prune the tree with tree size=5, plot the tree.

A diagram of a tree

Description automatically generatedc

1. (10 points) Predict for the test data with the full tree and the pruned tree respectively. Calculated the test error for the full tree and the pruned tree. Which one performance better?

The full tree had an MSE of 35.28 and the pruned tree had an MSE of 35.90. The full tree did slightly better but not really enough to warrant using the full tree over the less complex, pruned tree.

**Problem 8.** **(60 points total): Fit a regression tree**

Use **Carseats** dataset. With the dataset, we wish to predict the sales of car seats on the basis of various statistics associated with the sale data in the previous year. We choose Tree methods.

1. (10 points) Call library(tree) and use the tree() function. With the whole dataset as training data, fit a regression tree in order to predict “Sales” using all variables. What are the variables used as internal nodes in the tree? How many terminal nodes are there in the tree? What is the tree deviance? Print out the tree.

A blue screen with white text

Description automatically generated

The variables used as the internal nodes in the tree are shelveloc, price, age, income, population, advertising. There are 17 terminal nodes in the tree. The deviance is 2.878.

A diagram of a company's structure

Description automatically generated with medium confidence

1. (15 points) separate the data into training subset and test subset with setseed(1). fit a regression tree in order to predict “Sale” using all variables with the training data. What are the variables used as internal nodes in the tree? How many terminal nodes are there in the tree? What is the tree deviance? Print out the tree. Are there any difference between this tree and the tree from the whole dataset?

A blue screen with white text

Description automatically generated

The variables used as the internal nodes in the tree are shelveloc, price, age, advertising, compprice, us. There are 18 terminal nodes in the tree. The deviance is 2.167.

A diagram of a company

Description automatically generated with medium confidence

The tree using training data has one more terminal node compared to the tree using the whole dataset and also has a lower deviance.

1. (10 points) Now consider pruning the tree. use the cv.tree() function to see whether pruning the tree will improve performance. Plot the deviance as a function of tree size. Does pruning improve performance?

I do not think pruning improved the performance and we can see the proof for that with the increase in the amount of deviance.

A graph with lines and numbers

Description automatically generated

1. (5 point) Prune the tree with tree size=5, plot the tree. A diagram of a company's sales flow

   Description automatically generated with medium confidence
2. (10 points) Predict for the test data with the full tree and the pruned tree respectively. Calculated the test error for the full tree and the pruned tree. Which one performance better?

The MSE for the full tree is. 3.15 and the MSE for the pruned tree is 4.76. The full tree performed slightly better but the pruned tree is much less complex. The full tree probably has a lower MSE due to being overfitted on the training data so overall the pruned tree is better to use for predictions on new data.

1. (10 points) Compare the regression tree you fit in this problem 8(II), 8(VI), and the classification tree you fit in problem 5(V), what are the differences? Which one you would prefer? Why?

The classification tree from 5(V) and the regression tree from 8(II) are both overfitted on the training data which would not be good for making predictions on new data. However, the pruned regression tree only has 5 terminal nodes and is much easier to read and understand. The deviances for both full trees are lower than the pruned tree deviance. I would prefer the pruned regression tree due to its simplicity and the fact that it still performs well and is not overfit.