# DAT 520 Problem Set 4 Introductory Decision Trees

This module we leave the probability exercises and move into hands-on applications of the prior module’s lessons. There are three exercises, each using a different software for building trees. The goal this module is to get familiar with the software as it will be used in later modules and is an option for you to use in your final project. Work through these examples with a focus on understanding how to use each of the methods.

Problem 1: Cartfit Decision Tree: To complete this problem, you will need to read and follow an example. In following the example, you can simply cut and paste the commands, however the intended goals of this problem is to learn and understand how and why this decision is working and used. The steps for this problem are as follows:

1. Read Chapter 8 of the Larose text, “What Is a Decision Tree?”
2. Complete the R Zone exercise found on pages 180 and 181 of the reading.
3. Copy and Paste the R Zone decision tree you generate in R into a Word document.
4. Include an explanation of what this decision tree is telling you

The steps to complete this assignment are outlined here. These are provided to guide you through the problem. Note that in practice, you would be expected to know how to perform your own analysis, so making notes as you perform the exercise will be helpful in your use of these methods.

1. Open R.
2. Open the notepad file Module4\_Problem1.txt (found in DAT 520 Data Files folder).
   1. This text file contains the text commands listed in the RZone example. The commands have the extraneous characters removed from PDF copy pasting as well as points to the proper csv file for this assignment. If you choose to perform the assignment independent of these instructions, please see the problem notes below for additional details.
3. Each step begins with a # which is a comment identifier in R. Note: the first couple of instructions is to identify the file path and to then read in the file. You can copy all lines including the # if you want the R command execution will ignore the # line. Paste it into your R console and hit return. The command set should complete fine. If you encounter an error, please review the problem notes found below. Continue working through the file until all commands have been run from the example assignment and you have produced your tree.
4. Right-click on the tree image produced and copy it as a bitmap (image) into Word.
5. Write your explanation of what this tree is telling you.
6. Save the file. You will continue to use this file to add the results of Problem 2 and Problem 3 as well.

## **Problem 1 Notes:**

* If you are running R locally on your desktop, you can download the clemtraining data zip file and extract it to your local computer. Use the R file.choose() command to navigate to your file in R and R will display the exact path you need for the file=

The data set for this exercise is called Clem3Training.zip and unpacks to become Clem3Training.csv. It is provided in the DAT 520 Data Files folder on the Start Here page of your learning environment.

# Get the proper file path for CLEM3Training.csv

file.choose()

When you complete the R commands, make sure to save your new file for use in Problem 2

#write file for problem 2

write.csv(dat, file=“c:/dat520/CLEM3TrainingP2.csv”)

* Word is on the Virtual Lab and you can right-click on an image in R, copy it as a bitmap, and paste it into Word. Then save that document to your SNHU drive or use Dropbox. Alternatively, when you are finished with the assignment you could post it directly from the Virtual Lab into your learning environment.
* **Note**: There is a possible problem with copying the commands right out of the PDF of the Larose chapter. Some people have this problem and some do not. You may see an error like this:

Error: unexpected input in “cartfit <- rpart(income �”

It is the tilde character that R chokes on if you are getting that error.

To fix this error, use this:

cartfit <- rpart(income ~ age.z + education.num.z + capital.gain.z + capital.loss.z + hours.per.week.z + race + sex + workclass + marital.status, data = dat, method = “class”)

Copy and pasting that command into R from here works. It is the ~ character causing the problem. This [thread](http://r.789695.n4.nabble.com/unexpected-input-in-rpart-td3168363.html) explains a little more about the issue with R tildes.

* You can open your learning environment from the Virtual Lab to download the .zip file and unpack it, so that you can import Clem3Training.csv into R using the read.csv command in the VDI directly.
* Instructions for R are found in the Decision Tree reading beginning on page 180 in the R Zone step-by-step directions.

**Note that you may have to install the correct R modules in order to do the exercise**, like you had to do with “expm” to complete the exponentiation exercise. Get good at installing and invoking packages in R. It is a common task that needs to become second nature. Additionally, here areR directions for inputting data, in case you get stuck trying to get the data set into R to play with it: [How to Input Data Into R](https://stats.idre.ucla.edu/r/faq/how-to-input-data-into-r/). Keep this info handy for future use. Test it out on some data that you have, so that you can easily get data into R anytime you need to.

Problem 2: Rattle Decision Tree: To complete this problem, you will continue to use the same dataset from Problem 1, but use a different method to produce a tree. Before simply running the command in R for rattle to generate the tree, you will need to first perform an exercise in rattle to understand what it is and how it works. The steps for this problem are as follows:

1. Read and refer to the Data Mining with Rattle and R document in this module’s readings.
   * Read Chapter 11 p.205 sections 11.1, 11.2, 11.3
   * Work through the Example in section 11.4 p215 to get familiar with rattle
2. Complete the exercise using Rattle for the CLEM3Training.csv file
3. Provide an explanation of what the Rattle decision tree is telling you that is DIFFERENT than the Cartfit Decision Tree in Problem 1

The steps to complete this assignment are outlined here. These are provided to guide you through the problem. Note that in practice, you would be expected to know how to perform your own analysis so making notes as you perform the exercise will be helpful in your use of these methods.

1. Open R.
2. Enter the command library(rattle) and hit return.
3. Enter the command rattle() and hit return.
   1. This will pop open a graphical interface (GUI) for rattle.
4. Click on the Data tab.
5. Click on the Folder Icon next to the FileName.
6. Navigate to Local Drive (C:), double click dat520 folder and click on CLEM3TrainingP2.csv (the file created at the end of Problem 1) and click **Open**.
7. Click Execute which will load the data into Rattle.
8. Change “income” variable to **target** variable radio button
9. Change all other variables to **input** radio button
10. Click “Execute” to reload the changes to the dataset
11. Once loaded, select the Model tab.
12. On the Model panel, ensure Tree is the selected model and click **Execute**.
13. After it completes executing, on the right, there is a **RULES** and **DRAW** button displayed.
14. Click on **RULES**, wait for the process to complete, then click on **DRAW**.
    1. Note, that you may receive a prompt when using draw to load rpart.plot package and RColorBrewer package. Select yes to load if prompted. If you are prompted to select a CRAN site, please choose a site such as OH for Ohio or your state. The Draw image should then display. If it does note, the R console will have the log of the package load and you can use this information further troubleshoot any error with your instructor.
15. Go to RStudio to view the decision tree in the Plot pane.
16. Right-click on the tree image produced and copy it as a bitmap (image) into Word.
17. In Word, write your explanation of what the Rattle decision tree is telling you that is DIFFERENT than the Cartfit Decision Tree in Problem 1.
18. Save the file (you will continue to use this file to add the results of problem 3 as well).

If there are questions, please post them early in the module (sooner than Sunday night!) in the provided discussion for this problem set.