Intermediate Tree 1

If you followed through the Basic Decision Tree exercise, this should be useful for you. This is like a continuation but we add so much more. We are working with a bigger and badder datasets. We will be also using techniques we learned from model evaluation and work with ROC, accuracy and other metrics.

Answers to the exercises are available here.

If you obtained a different (correct) answer than those listed on the solutions page, please feel free to post your answer as a comment on that page.

Exercise 1
read in the adult.csv file with header=FALSE. Store this in df. Use str() command to see the dataframe. Download the Data from here

Exercise 2
You are given the meta_data that goes with the CSV. You can download this here Use that to add the column names for your dataframe. Notice the df is ordered going from V1,V2,V3 _ _ and so on. As a side note, it is always best practice to use that to match and see if all the columns are read in correctly.

Exercise 3
Use the table command and print out the distribution of the class feature.

Exercise 4
Change the class column to binary.

Learn more about decision trees in the online courses

- Regression Machine Learning with R (it includes two lectures on definitions, characteristics, mathematical formulae, graphical representations, fitting, forecasting, and accuracy of decision trees)
- Machine Learning A-Z™: Hands-On Python & R In Data Science (including 3 lectures, ~45 mins, on decision tree regression in both R and Python)
- Data Science and Machine Learning Bootcamp with R (126 lectures and 17.5 hrs of video, including several lectures on decisions trees and random forests)

Exercise 5
Use the cor() command to see the corelation of all the numeric and integer columns including the class column. Remember that numbers close to 1 means high corelation and number close to 0 means low. This will give you a rough idea for feature selection

Exercise 6
Split the dataset into Train and Test sample. You may use sample.split() and use the ratio as 0.7 and set the seed to be 1000. Make sure to install and load caTools package.

Exercise 7
Check the number of rows of Train
Check the number of rows of Test

Exercise 8
We are ready to use decision tree in our dataset. Load the package “rpart” and “rpart.plot” If it is not installed, then use the install.packages() commmand.

Exercise 9
Use rpart to build the decision tree on the Train set. Include all features.Store this model in dec

Exercise 10
Use the prp() function to plot the decision tree. If you get

any error use this code before the prp() command

```
par(mar = rep(2, 4))
```