Forrest Fallon  
Homework 5  
IST 707  
11/7/21

This week’s assignment requires us to use decision trees as a means of description and prediction for the Federalist Papers dataset. We will first clean the data, then create decision tree models, and finally we will create prediction models for the data sets based off of the trees!

First up, cleaning. Below are the steps we take in order to install the R packages necessary, read the data file, and create test data sets in order to avoid confusing the decision tree parameters.

Text

Description automatically generated

Take note that we removed the Filename column from the dataset as this provided minimal information necessary to our desired outcome for this exercise. Also worth noting that we use the subset command in order to create a dataset free from the “dispt” papers which we will come back to later.

Now that our data is in good shape, we can begin creating the decision trees. We will start with default values to make sure the plot will function properly:



Diagram, text

Description automatically generated

Success! A tree was made, and the program is able to read and separate our dataset based off of how often each author is associated with which words. “Upon” seems to have the most instances of divisiveness within the data, and seems to be our biggest separator between Hamilton and Madison authorship. Let’s adjust our decision tree parameters and try to get more authors within the data involved.

Text

Description automatically generated

A picture containing text, clock

Description automatically generated

Getting better! Jay and HM are now being deciphered as well, this tree is actually a very solid representation of the data and how it splits between authors. For the sake of investigation we will try and change the parameters a third time:

Text

Description automatically generated  
Diagram

Description automatically generated

As we can see here, changing the minsplits to a high of 15 eliminated HM from the tree, and began assuming that all things not written by Madison after the first split were then written by Jay. We know this to be false, we will continue using the 2nd tree for further analysis.

Let’s now check the cp plot of tree2 in order to know for sure if we are avoiding as much error as possible (plotcp(tree2)):  
Chart, line chart

Description automatically generated

Good news! Based on the graph above, our tree2 plot can remain as is. Even if we changed co from 0 to 0.16, the amount of change in error would be minimal. That may be necessary with larger tests than this, however, the nature of this dataset prevents any significant change from taking place if that cp change were to happen. Let’s take a quick look at the summary of the tree2 plot:

Text

Description automatically generated with medium confidence

Excellent, our summary command has solidified what we assumed based off of the cp plot. As splits go up, relative error goes down and xerror stabilizes. A good rule-of-thumb is to make sure that rel error + xstd < xerror, seeing as though our bottom two splits easily satisfy this rule, we can proceed with confidence in this model.

Next up in our assignment is the using the decision tree model to predict against our dispt papers data. The following code will help us accomplish that:

test\_pred <- data.frame(predict(tree2, newdata = dispt\_data))

Which grants the following:  
Calendar

Description automatically generated

As seen above, the program feels very strongly about the authorship of the disputed papers! It tells us that ~91% of the time, the disputed paper is written by Madison, and the other 8% of the time it is likely written by either Hamilton or Madison. This makes a lot of sense seeing as though the data we are presented with is overloaded with Hamilton entries, followed by Madison. With a data set such as this, running decision tree models followed by prediction models based on the decision tree, we are setting ourselves up for a fancy way of receiving the results of a count(\*) function on our data. In other words, more data does not always equal more accurate predictions.

Decision tree models can be very useful with the right kinds of datasets; datasets with many different outcomes, variables, and proper representation for each outcome. This exercise has illustrated the need for diversified data in relation to having the program return probabilities to us, that is if we want to receive probabilities that tell us anything other than what we can already deduce from looking at the raw data.