Matthew L. Pergolski
IST 707
2/15/2022

Introduction

The Federalist Papers, some would say, are the ultimate precursor to the United States Constitution. They contained many of the same values and initialized principles that would later define the birth of a new nation. Some historians indicate that these works define a larger importance to America than just about any documentation preceding or following it.

The authors of these significant papers to the United States' history are mostly known. In more detail, the main authors are definitively known; however, there lies a mystery in terms of who wrote (i.e., 'authored') each document. In terms of what is known by historians, 85 total papers lie in existence. Alexander Hamilton, close aide, and advisor to George Washington, took authorship over at least 51 essays, while two others also contributed. James Madison, who would go on to be the nation's 4th president, was responsible for authoring at least 15. Hamilton and Madison collaborated on an additional 3, and John Jay – who would go on to be the first Chief Justice of the US – wrote 5.

The overall mystery corresponds with the remaining 11 essays, where it was unclear as to who was the true author. Both Hamilton and Madison have allegedly accepted authorship for the documents – which has caused much debate and resulted in a 'disputed' perspective between the congressional (and historical) community.

The basis of the information in the following sections will attempt to utilize decision tree analysis to determine authors for the remaining 11 'disputed' essays.

Analysis and Models

About the Data / Exploring the Data

The decision tree analysis will ultimately consist of matching the writing styles of the two different authors and making a comparison to the 'disputed' essays/documents. This will be done by utilizing a data set with 'function' words used by the various authors. The actual values indicate a percentage of the how often the particular function word appeared in the paper. For example, Hamilton's 31st federalist paper utilized the word "upon" 3 times out of a total of 1,000 words, so the value for this respective observation is 0.3.

In looking at the data set as a whole, a total of 72 attributes exist with 85 observations (one observation for each federalist paper). The 72 attributes consist of various function words such as "should," "shall," "down," "into," etc.

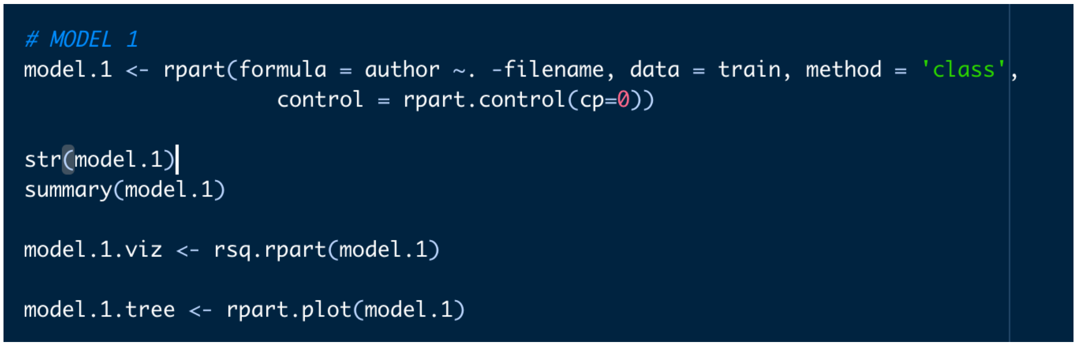
The first two attributes, however, are not function words. Instead, they are a categorization of author and the name of each file/paper/document corresponding to said author(s). These first two variables were imported as ‘char’ for character, while the rest are considered ‘num’ for numeric.

In terms of data cleaning and preparation, data sets were placed into subgroups – one group with the disputed essays and another with the remaining papers. A training and test data set was then established, utilizing the ‘createDataPartition’ function.

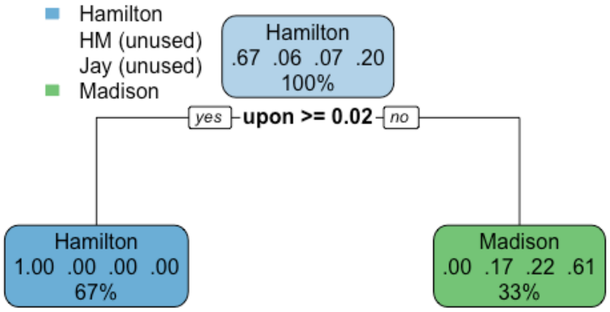
Results

All in all, various models were developed involving the decision tree analysis. For each node, three (3) overall observations can be made. First, the author headlines each respective node, followed by the probability and frequency percentage (i.e., rate). The models and visualizations show as follows:

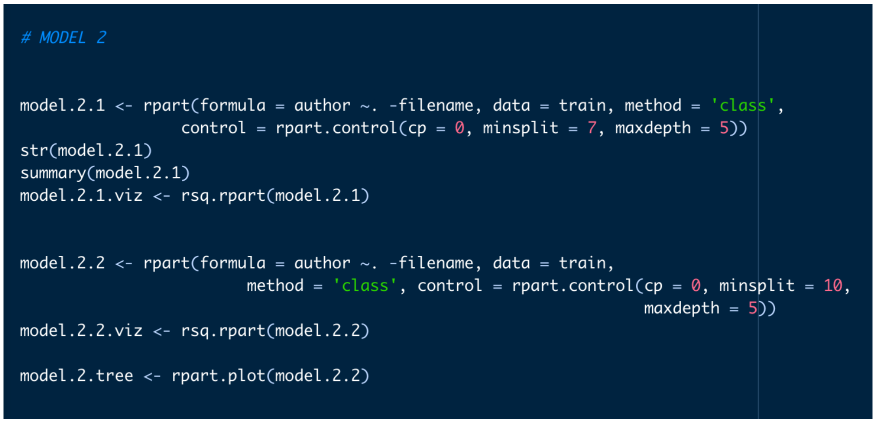
MODEL 1:



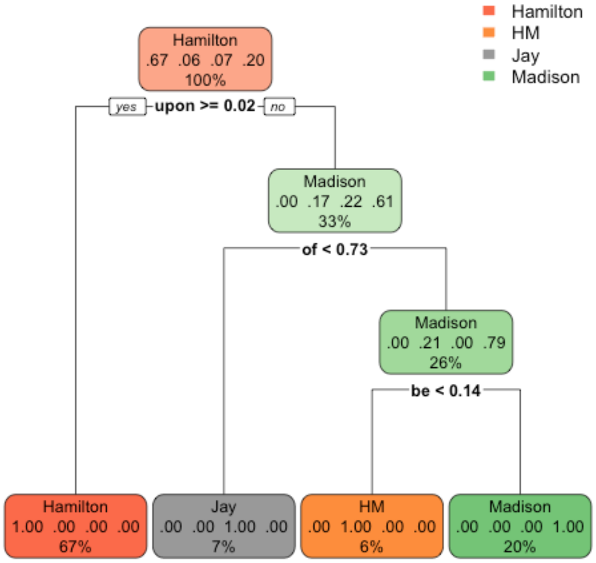
```
# MODEL 1
model.1 <- rpart(formula = author ~. -filename, data = train, method = 'class',
                    control = rpart.control(cp=0))

str(model.1)
summary(model.1)

model.1.viz <- rsq.rpart(model.1)

model.1.tree <- rpart.plot(model.1)
```



MODEL 2:

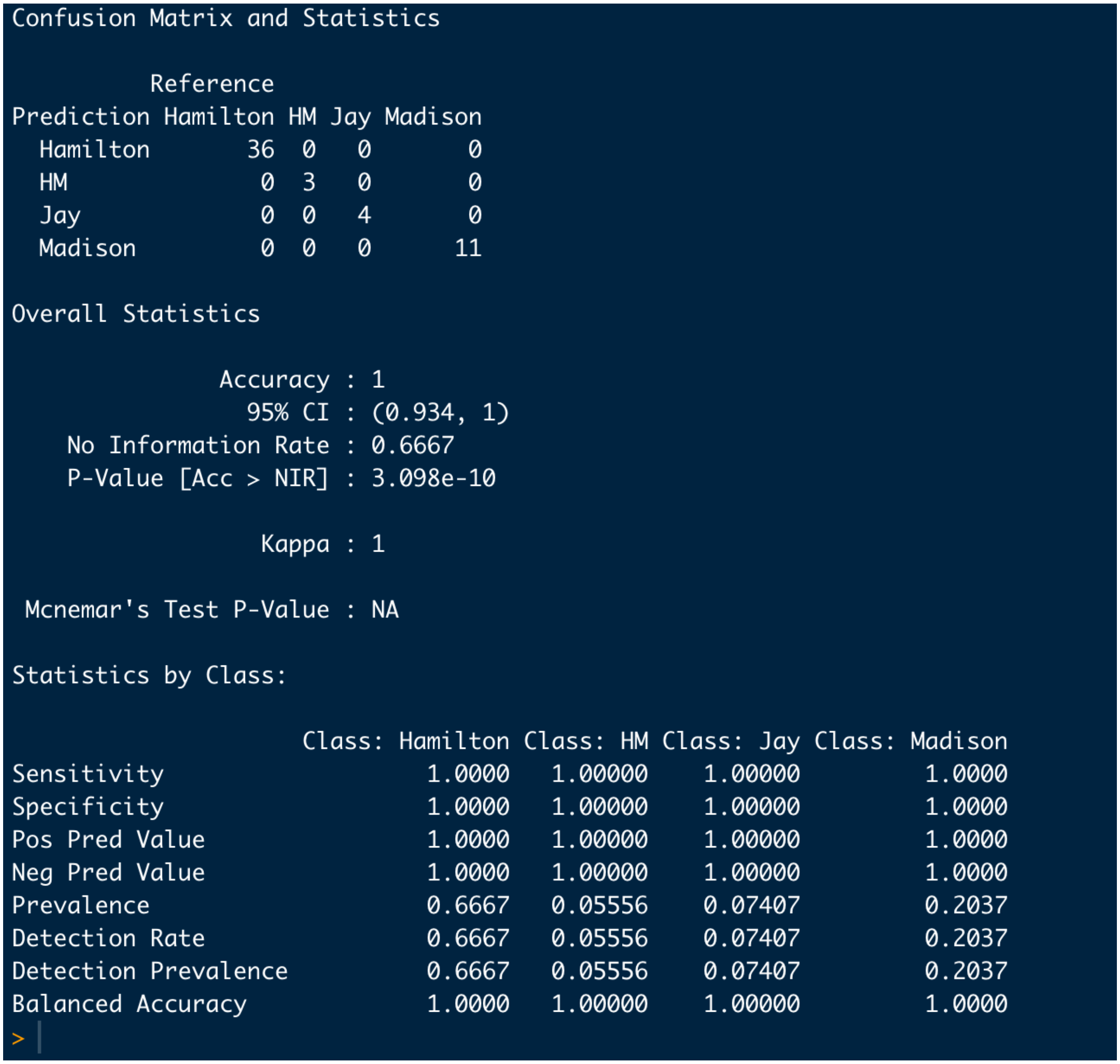


```
# MODEL 2

model.2.1 <- rpart(formula = author ~. -filename, data = train, method = 'class',
                 control = rpart.control(cp = 0, minsplit = 7, maxdepth = 5))
str(model.2.1)
summary(model.2.1)
model.2.1.viz <- rsq.rpart(model.2.1)


model.2.2 <- rpart(formula = author ~. -filename, data = train,
                        method = 'class', control = rpart.control(cp = 0, minsplit = 10,
                                                                   maxdepth = 5))
model.2.2.viz <- rsq.rpart(model.2.2)

model.2.tree <- rpart.plot(model.2.2)
```



Conclusions

Within the conclusion section, the prediction aspect of the analysis will be displayed within the confusion matrix visibility:



```
Confusion Matrix and Statistics

          Reference
Prediction Hamilton HM Jay Madison
  Hamilton       36  0   0       0
  HM              0  3   0       0
  Jay             0  0   4       0
  Madison         0  0   0      11

Overall Statistics

               Accuracy : 1
                 95% CI : (0.934, 1)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : 3.098e-10

                  Kappa : 1

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Hamilton Class: HM Class: Jay Class: Madison
Sensitivity                   1.0000   1.00000    1.00000         1.0000
Specificity                   1.0000   1.00000    1.00000         1.0000
Pos Pred Value                1.0000   1.00000    1.00000         1.0000
Neg Pred Value                1.0000   1.00000    1.00000         1.0000
Prevalence                    0.6667   0.05556    0.07407         0.2037
Detection Rate                0.6667   0.05556    0.07407         0.2037
Detection Prevalence          0.6667   0.05556    0.07407         0.2037
Balanced Accuracy             1.0000   1.00000    1.00000         1.0000
>
```

Based on the prediction model outlined in the confusion matrix, it appears as though James Madison was the most-likely author of the 11 disputed essays. Within the decision tree models themselves, we see Hamilton vs Madison appear the most. Since the overall data set is relatively small, the splitting of this set into both training and testing purposes could associate with the topic of overfitting within these models. The answer to limitations and quality concerns regarding a particular set is always the fact that more data would be ideal to work with. With that said, based on the following that was provided via class, James Madison appears to be the likely contact as alluded to within the models.