Introduction:

As referenced in the previous assignment, prior to the age of social media and broadcast television, the strength of the written word and oratory prowess were the most compelling forms of influence. Usually the phrase “this country was founded on ‘blank’” is used in hyperbole, however in the instance of the Federalist Papers, it is meant quite literally. The Federalist papers were a collective of 85 essays written by statesmen Alexander Hamilton, James Madison, and John Jay. These essays were written under the joint pseudonym “Publius” with the impetus of persuading the people of New York state to ratify the proposed U.S Constitution. These enigmatic essays were published in various New York newspapers for a 1-year duration from 1787-1788 and can with bona fides, be describe as some of the most important documents in U.S history.

Analysis and Models:

About the Data:

The “FedPapersCorpus” data sets are a series of txt files containing each of the 85 federalist papers in their entirety. Within the file, the essay title, date of publication and author are displayed along with the full form of the essay. In total there are 74 essays with identified authors, 51 essays written by Hamilton, 15 by Madison, 3 by Hamilton and Madison, 5 by John Jay. The 11 remaining essays hold contention regarding authorship. It is unconfirmed whether they were written by Alexander Hamilton or James Madison.

The Fed Papers in Corpus format is loaded in and then cleaned and preprocessed. The first cleaning measures are data preparation and transformation. Punctuations, numbers, and spaces are removed from the data set. Rare words, categorized as words the occur less than 1% of the time are removed. Highly common words, categorized as words that appear in more than 50% of the document are removed. Words that are frequent but are of little value are referred to as stop words. Two layers of stop words are removed. The first, a vector of 37 words specific to the various essays. Following that, a vector of 174 stop words that are commonplace in the English language. A document term matrix is then created from the Fed Papers Corpus with the stopwords variable as a parameter, punctuations removed, numbers removed, all words converted to lowercase, stemming, separators removed, English language stop words removed, and finally rare and common words.

A picture containing text, receipt

Description automatically generated

Above is a word cloud output of the document term matrix based on a vectorized output of word frequency.

Model:

The R model of choice for this analysis is the rpart model. This model is used for classification and regression decision trees. In this instance, the rpart model will be used for classification. The data must first be bifurcated into two separate groups, a testing set and a training set. The data allocation is two-thirds for the training set and one-third for the testing data. This split allows the model more instances to learn with the data and be better suited for its predictive function. The selected root node and internal nodes comprise the splitting attributes. A root node splits the data in half. If the root nodes lead to unanimous decision, they are regarded as pure. If the internal node subsets give inconsistent answers, they will be continually split until they attain a full decision. Multiple trees can fit the same data. There are two methods of splitting data. A two-way split creating a narrow and deep decision tree or, a multi-way split that creates a broad but shallow decision tree. The methodology is dictated by the problem and the type of variables contained in the data set. Pruning is essential for simplifying the tree by keeping only the most important splits. Accuracy is intentionally reduced to keep model usability ubiquitous and prevent overfitting.

Rpart classification accuracy may be validated in several ways. Method 1, the trained model based on the training data is used to predict the testing data. Predicted values are compared against actual values in the test data to quantify model accuracy. Method 2, a cross-validation comparing the r-square against the number of splits. This tells us which split offers the most information. Method 3, the final check is to run another cross-validation, this time relative error against number of splits. This tells how the tree should be pruned split wise.

Results:

The data is vectorized to give more insight as to word use frequency. In order to further prep the data for optimal use in rpart, it is normalized. Normalizing has the dual benefit of minimizing outlier contamination and increasing processing ease. Post normalization the appropriate author names are added in addition to a “dist” label for the disputed essays. The data is now a data frame with each essay being a record. The data frame begins with the author’s name and is followed by the most frequently used words and individual normalized values in the columns after.

The first model used splits the data into two groups, one to train the model and the other to test the model. The training data is .60 (60%) and the testing data is the remaining .40 (40%). After the training data is used as a parameter to the rpart model, the model predicts the author of an essay in the test dataset which we compare against the actual author to evaluate efficacy.

A picture containing text, clock

Description automatically generatedTable

Description automatically generated

The images above show that the model predicted all 11 of the disputed essays were written by Madison. If the word “upon” was used more than 0.0035% of the time the model predicted the essay written by Hamilton. If not, then predicted it was written by Madison. Cross validation is performed model fit to the data and ensure no overfitting has occurred.

Chart, line chart

Description automatically generated

Chart, line chart, scatter chart

Description automatically generated

Because only one split is displayed, the above 2 validation images show an extremely linear relationship. The data from these images were the impetus to changes the number of splits in the decision trees for later updated models. After creating and testing multiple models with different criteria as well pruning, the predictions started to move from predominantly Madison to Hamilton for the disputed texts.

The best model based on the validation metrics had essays written by Jay, the disputed essays, and the co-authored essays by Hamilton and Madison included. The original model only included the essays written by Hamilton or Madison independently and the disputed essays.

A picture containing text, clock, watch, device

Description automatically generatedTable

Description automatically generated

Chart, line chart, scatter chart

Description automatically generated

Chart, line chart, scatter chart

Description automatically generated

The above figures are all from the best model. In comparison with the first model above, the cross-validations shows a better fit. The predictions are nearly the same as the original model except one is attributed to Hamilton. Of the 11 disputed essays, Madison is predicted to have predominantly written 10 and Hamilton, 1.

The rpart classification tree model had similar outcomes to the previous HAC assignment. The k-means model predicted that the majority of the disputed essays were written by Hamilton. The HAC model predicted that most of the disputed essays were written by Madison. The k-means predictions were the polar opposite. However, the HAC model’s predictions were much more analogous to the rpart model predictions.

Conclusions:

The mystery of the federalist papers is one that transcends time and technology alike. Multiple complex algorithms have been utilized in this issue to no avail with conflicting and unclear results. Departing from frustration born of curiosity, what’s most important is the historical impact this collection of essays has had on the United States of America. We can, with confidence, credit much of the ratification of our constitution to the ghost writing of these influential galvanizing men. Within the totality of this age-old mystery, that comforting fact alone, will have to be enough.