The Federalist Paper: An Enduring Mystery to True Authentication

HW 6: Decision Trees

Syracuse University: IST 707

Gavin Grosswald

Introduction

The Federalist Papers, a pivotal piece of American political history, present a true mystery: who authored the disputed essays, Alexander Hamilton or James Madison? Written under the pseudonym "Publius," these essays played a role in advocating for the ratification of the United States Constitution. Despite the contributions of Hamilton, Madison, and John Jay to this work, a subset of eleven essays remains ambiguously labeled "Hamilton or Madison." Data scientists across the globe have tried to solve this mystery using computational methods, building on the groundwork laid by statisticians Mosteller and Wallace in the 1960s. By analyzing the frequency distributions of common function words within the Federalist Papers, modern data mining techniques offer a fresh perspective on who the true Author may be.

Through studying the complete body of work and the linguistic makeup of the Federalist Papers, scientists aim to uncover unique stylistic nuances that can be attributed to the individual Authors. In turn, identifying the true voice of the mysterious 11 essays. Through deep examination and visualization of decision tree results, scientists seek to highlight the distinctive writing styles of Hamilton and Madison, while also exploring the possibility of joint authorship scenarios. As technology continues to enhance investigations and analysis, the application of data mining techniques will help to uncover the beautiful historical mystery of the ambiguous Federalist Papers.

Analysis, Models, and Results

About the Data

The dataset comprises the Federalist Papers, a collection of 85 essays advocating for the ratification of the United States Constitution. These essays were initially published between 1787 and 1788 under the pseudonym "Publius," but have since been attributed to Alexander Hamilton, James Madison, and John Jay. The dataset contains 74 essays with identified authors: 51 authored by Hamilton, 15 by Madison, 3 by Hamilton and Madison jointly, and 5 by John Jay, with the remaining 11 essays holding the ambiguous label "Hamilton or Madison." The dataset focuses on the occurrence percentages of function words within the essays, such as "upon" or "for," serving as indicators of writing style. This data forms the foundation for the application of decision tree analysis, offering a computational lens through which to view the mystery surrounding the disputed essays' true authors. Through analysis and visualization of these textual patterns, scientists aim to uncover subtle nuances that may shed light on the identities of the authors behind the Federalist Papers.

Reading / Cleaning the Data

The Federalist Papers csv file is read into R and converted to a dataframe called “FedPapers”. In order to best understand the story of the data, functions like dim(), str(), colSums(), summary(), and colnames() are queried. Initial observations are as follows:

- 85 observations of 72 variables
- Columns [,1:2] are <chr> and columns [,3:72] are <num>
- Columns [,1:2] relate to the author and essay file and columns [,3:72] represent feature words found among all 85 essays
- The values are the percentage of the word occurrence in the specific essay (i.e., if the word “upon” appeared 3 times, and the total number of words in the essay is 1,000, the value is 3/1,000 or 0.3%)

Data Analysis

A decision tree is a straightforward model that guides decision-making by segmenting data based on specific features. It functions like a flowchart, progressively refining data divisions until reaching predictive outcomes like "yes" or "no." Decision trees are commonly used for their simplicity and ability to support predictive analysis.

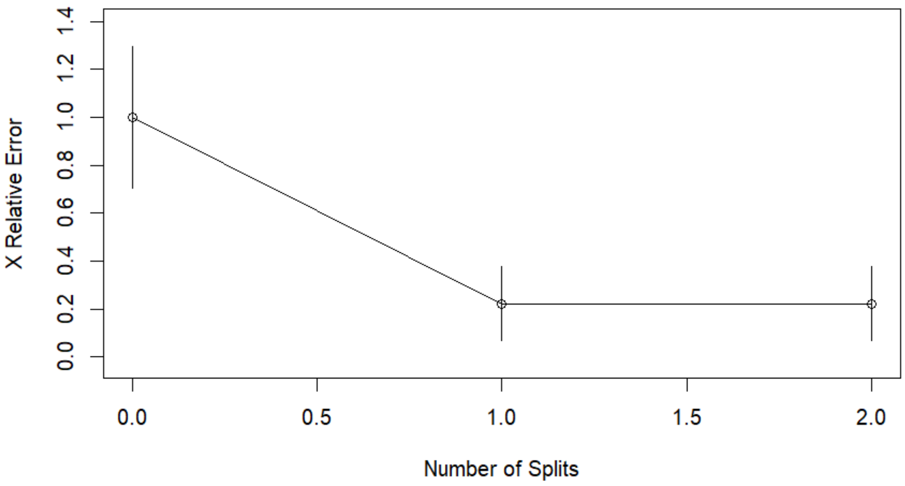
To prepare the data for decision tree analysis, the "author" and "filename" columns were removed, resulting in a fully numeric dataset. Similarly, crucial words for analysis were identified, prioritizing those with the highest variance among authors to reduce noise. Each author's dataset was scrubbed to select the top 40% of words for further analysis.

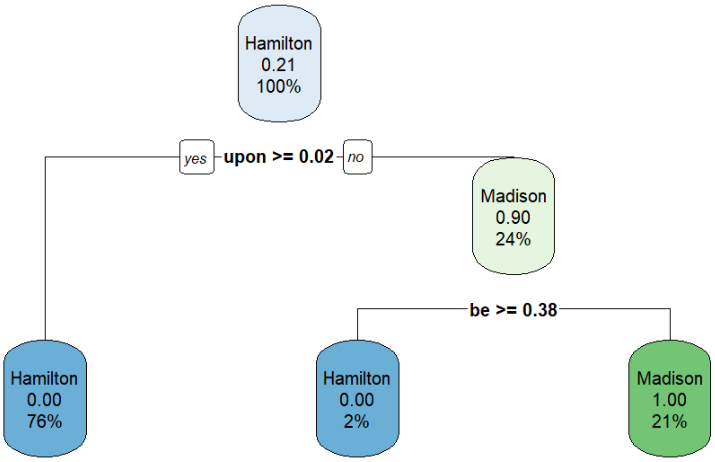
Subsequently, the data was filtered to include only papers by Hamilton and Madison, the key authors under debate for the disputed essays. A training and test set was created by randomly selecting 65% of both Hamilton's and Madison's essays, facilitating decision tree training.

*Please refer to the associated R Script for specifics.

Data tree analysis can be seen below for both unpruned and fully pruned datasets. Both a decision tree and the prediction of the Author for each of the 11 disputed essays can be found:

Unpruned:



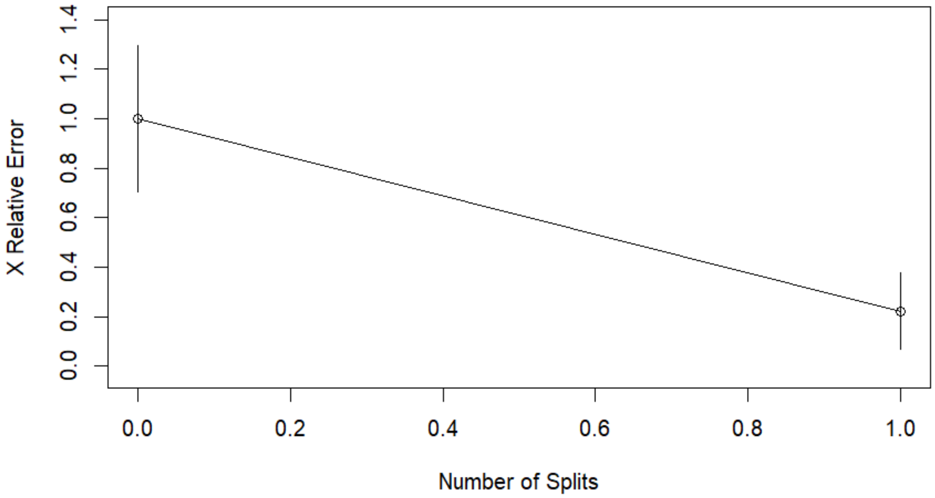


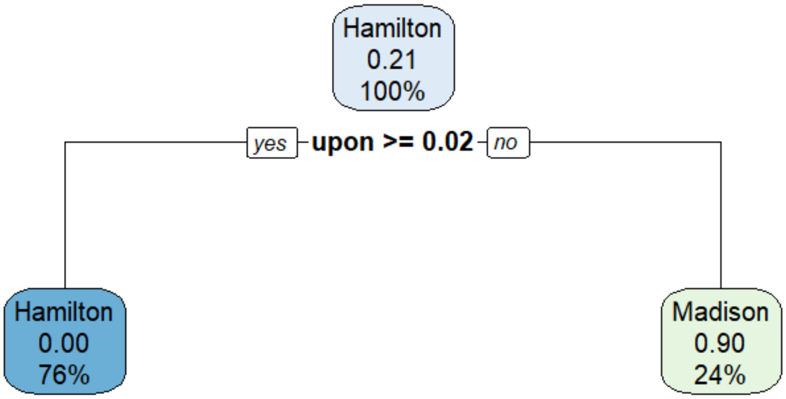
	Hamilton	Madison
Hamilton	18	0
Madison	1	5

1	2	3	4	5	6	7	8	9	10	11
H	H	H	M	M	M	H	M	M	H	M

Pruned:

The tree was recreated using a pruned dataset. This is automatically done using rpart() which involves splitting the dataset into smaller subsets based on the values of input variables.





	Hamilton	Madison
Hamilton	18	0
Madison	0	6

1	2	3	4	5	6	7	8	9	10	11
M	M	M	M	M	M	M	M	M	M	M

Results

Both unpruned and pruned decision trees underwent testing and training, using a 65% split of the dataset. The initial unpruned model attributed 5 of the 11 disputed essays to Hamilton and the remaining 6 to Madison. However, upon pruning and re-evaluating the data, the revised model conclusively assigned authorship of all 11 essays to Madison.

Conclusion

The search to uncover the authors of the disputed Federalist Paper's essay is a mystery that continues to leave data scientists perplexed. Scientists around the world continue to leverage advanced data mining techniques to analyze language patterns in hopes of distinguishing between Alexander Hamilton's and James Madison's unique writing styles. However, despite these efforts, the mystery behind the eleven ambiguous essays remains, highlighting the complexity of language and the need for continued study. With the constant advancement and investment in the data science field, analysis techniques will continue to improve, ultimately helping to solve this mystery once and for all.