Nick Videtti

nvidetti@syr.edu

Assignment 5

Decision Trees

Supervised Learning/Testing

8/16/2022

| Hamilton or Madison? - Part Two | |
| --- | --- |
| **Introduction** | The Federalist Papers were written by Alexander Hamilton, James Madison, and John Jay, even sometimes by a mix of them. For the most part, the authors of each of these are known, but there is some dispute over who wrote some of them.  In total, there are 85 papers, with 11 of those having a disputed author. The vast majority were written by either Hamilton or Madison, and a smaller amount were written by either Jay or Hamilton and Madison together. The 11 disputed papers will be referred to as the following.    Data have been given that contain some words from each paper. Using those data, the author of the 11 disputed papers can be predicted. For both practical and technical reasons, it will be assumed that the disputed papers were either written by Hamilton or Madison individually, the practical reason being that only 8 of the 74 papers with known authors were not written by either Hamilton or Madison individually. |
| **Analysis** | The method used for this analysis will be decision trees. The process for using decision trees for predictions will make decisions based on whether or not a given word is in each of the papers. This will essentially convert each variable to a boolean type where the variable name is the word itself, a value of 1 will indicate that the word appears at least once in the paper, and a value of 0 indicates that the word does not appear at all in the paper. To make this a qualitative variable that truly behaves like a boolean, each of these variables will need to be converted to factor variables.  Next, the data set will need to be split into training and testing data. Rather than randomly subsetting a certain proportion, there are defined training and testing data that will be used for this problem. The training data will be the papers written by Hamilton individually or by Madison. The testing set will be the disputed papers. All other papers will be disregarded for the remainder of the problem. Note that this implies the assumption that the disputed papers are all written either by Hamilton or Madison, exclusively.  Once the data are split into testing and training sets, the columns that have not been taken into account yet will need to either be removed or modified. In the training data, there are filename and author columns. The author column will need to be converted to a factor and the filename column will need to be removed since it is not needed and not wanted for training the model. The testing data will need the author column removed entirely and will need the filename column reassigned to the row indices.  Finally it will be time to create the decision tree. This will allow making predictions as to which papers were written by whom. The prediction will then be done and results will be reported. |
| **results** | The decision tree that was generated is as follows.    This result was surprising, but based on the data, it appears that the author of each of the disputed papers can be predicted with a relatively high accuracy based on whether they contain the word “upon”. The results of the tree tell that there is a 77% chance of a paper being written by Hamilton regardless of whether that paper contains the word “upon”. However, when a paper does contain the word “upon”, there is a 94% chance it is written by Hamilton. Even more interesting, if a paper does not contain the word “upon”, it has a 100% chance of being written by Madison. It is important to distinguish that these results are based on the data provided and the Decision Tree model that was created using those data, as it is not intuitive for any result here to have 100% certainty.  As far as the predictions, the Decision Tree was used to predict the author of each disputed paper. The results were as follows.    All but one of the papers were predicted to have been written by Madison. Due to the nature of the model used for these predictions, it can also be stated that only one of these papers does not contain the word “upon”. |
| **conclusions** | It seems that many of the papers that are disputed were likely written by James Madison, at least according to the technical findings. This is surprising because Alexander Hamilton seemed to have written most of the Federalist Papers based on the papers with known authors, but the margin between the number Hamilton wrote and the number Madison wrote is perhaps much smaller than what is agreed upon.  It was predicted that Hamilton wrote one of the disputed papers and Madison wrote the rest. The predictions per paper were as follows.    The most interesting finding was that these predictions can be made with high accuracy simply by indicating whether the paper contains the word “upon”. In fact these predictions were made on that basis alone. |