Analyzing Voters

Group Members: Arshiya Sabzevari, Colton Ragland,

Noumik Thadani, Trevor Huis in ‘t Veld

Problem Statement

Using political data taken before and after the election, we are building a model to predict how Americans will vote. The data consists of voter demographics and their political stances during the elections season. The problem is extremely relevant as innovations in predicting and swaying voter actions changed the 2016 election and will impact all future elections. The use of machine learning in politics will continue to expand as candidates search for advantages and social media extracts more information from users every day.

Media companies, political parties, and tech companies are all interested in what we are trying to predict. Media companies would love to have an advantage in covering elections and predicting outcomes to earn a reputation as the most trusted source of news. Political parties will want this data to more effectively sway voters and utilize campaign funds more efficiently. The successful Obama and Trump campaigns famously had significant advantages in campaign analytics.

Tech companies and their users should become more knowledgeable about the way these machine learning algorithms can affect the users. Companies like Facebook obviously benefit from the advertising dollars that are spent by politicians, but users could also become wary of using their website if they are being data mined. The features we are using for a voter are different than the features of Facebook or Twitter. However, users should be aware of how predictable their actions can be from broad demographics and the data social media companies use is far more predictive and comprehensive.

The information is also interesting and thought-provoking for citizens concerned about the state of America. Race, sex, and religion are all hot-button issues and voters are interested in how those affiliations define our political leanings. Predicting voter decisions will be a major part of 21st-century democracy and a focus of the machine learning field.

Data

Our dataset is the Cooperative Congressional Election Study, a survey of thousands of people from around the United States centered on political stances and administered by YouGov. It has over 400 features and exactly 60,000 records. A majority of these features were answers to questions, some more relevant than others. These included basic demographics, specific questions about ancestry, detailed and overall opinion on major issues, and party identification.

The survey consists of a pre-election wave and a post-election wave. The pre-election wave is conducted from late September to late October. In the pre-election wave, respondents answer two-thirds of the questionnaire about general political attitudes, various demographic factors, assessment of roll-call voting choices, political information, and vote intentions. In the post-election wave, respondents answer the other third of the questionnaire, mostly consisting of items related to the election that just occurred. The post-election wave is administered in November.

Method

We were faced with an unmanageable amount of data: the majority of our computers could not load all of the records and features. Because of the lack of required computing power and the fact that many of the features were useless to what we sought to predict major political stances, we decided to drastically reduce the dimensionality. This was easy for many of the features, given that many questions were related to each other or simply elaborated on an already-addressed subject.

For the majority of the data set, we went through the columns, feature-by-feature, evaluating the worth of the information. Eventually, we settled on a specific set of features that were deemed sufficient for doing our analytics, and we deleted the rest of the features. Moreover, for the sake of storage, we deleted all records that were not registered to vote, reducing the data set to 31,945 records.

To predict the actions of voters we decided to build a neural network model. We believed a neural network’s ability to apply weights to the different features would serve this data well.

Challenges

The size of the dataset was a major challenge. Determining which features to use in order to elicit meaningful results was the most difficult part of our analysis. We eventually resigned to only use ~10% of our dataset to train the model. We were challenged by the computing power of our personal computers as even running tiny portions of the data would take over 20 minutes. A more powerful computer would have allowed us to tweak the model more efficiently and we did not anticipate this challenge. The dataset's 400 features were often very similar such as multiple questions on a voter's opinion on climate change. If a voter answered one question in a certain way, then they almost always agreed on the rest of the questions.

Results

The result of our Neural Network model gave us a 66.87% accuracy in predicting which party (Democratic, Republican, Independent, or Other) a voter would associate themselves with. As with all models, we were hoping for higher accuracy in our final results, but with greater computing power we could have incorporated more of the features we had to eliminate. In the end, we have a model in which you can input a voter's demographics and stances on core issues in order to predict their party.

Next Steps

The model created can be improved to increase the accuracy of the results. By using or conducting surveys and polls in a way tailored to input data into this model, we can feed more relevant data into the model. In addition, using a computer with fast processing power would allow us to input more features into the model to predict a likely party, which would result in greater accuracy.

Many political organizations, such as the Democratic National Committee and Republican National Committee, have data teams that rely on prediction models to forecast probable outcomes in elections. These forecasts are then used to allocate funding and focus grassroots political action in the appropriate areas.

As the accuracy of the model improves, the records could also be expanded to include non-voters. The model could then be used to predict likely party as well as the likelihood that someone will vote.

Sources

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910/DVN/ZSBZ7K