General Assembly NYC

Data Science 24

Dylan Gessner

October 14, 2015

**Introduction**

In 2010, the United States Supreme Court ruled that the First Amendment of the U.S. Constitution prohibited the government from restricting independent political expenditures by non-profit organizations. This landmark case, *Citizens United v. Federal Election Commission*, forever changed the nature of federal elections as from then on any Political Action Committee (PAC) could make unrestricted numbers of donations to and on behalf of political candidates. Money influencing politics is not a new concept, but my question is: to what extent does PAC expenditure influence the outcome of a federal election?

**Data**

To explore my research question, I used data from the 2012 federal election cycle. The data comes from the Center for Responsive Politics’ Campaign Finance data tables. The two tables of interest were the Candidates (‘cands12.txt’) and PACs (‘pacs12.txt’) tables. Important features to note are “recip\_code” in the Candidates table, which includes the election outcome for the candidate in question, and “amount” in the PACs table, which is the amount donated or spent by a PAC. For the full definitions of all features included in each table, please refer to Appendix A.

**Analysis**

Due to the large number of features included in my models and the presence of some outliers, I chose to use Random Forest classification to determine the effects of my features on election outcome. In each run, I benchmarked the performance of my Random Forest models against Support Vector Machines (SVM) classification and that of a simple dummy classification algorithm. I measured the performance of my tests using cross validation and chose to measure success as prediction accuracy, since I have no preference towards precision or recall.

Model 1— My first Random Forest model was able to predict election outcome with 95.9% accuracy, compared to SVM accuracy of 94.8% and dummy classification accuracy of 82.8%. Included in the first model were the following features:

amount + pac\_id\_unique + pac\_id\_total + C(party) + C(race\_type) + C(state) + C(district) + C(no\_pacs\_bin) + C(incumbent)

After looking at the feature importance generated by my Random Forest model, I decided that I could potentially achieve higher accuracy by including the distid\_run feature instead of two separate features for State and District, which were features engineered off the distid\_run feature.

Model 1a – My second Random Forest model was able to predict election outcome with 96.3% accuracy, a marginal improvement over my first model. Since it appeared that controlling for State and District did not have a significant impact on my prediction accuracy, I did not include location variables in further iterations.

Model 2 – In my third run of Random Forest, I was able to improve my prediction accuracy to 96.7%. The prediction accuracy of SVM also improved to 96.3%, since SVM performs better when fewer features are included. Dummy classification also improved to 84.6%. Included in this model were the following features:

amount + pac\_id\_unique + pac\_id\_total + C(party) + C(race\_type) + C(incumbent)

Model 3– At this point I thought it might be useful to try clustering my data to see if I could use the groups to improve prediction accuracy. Although the K-Means clustering algorithm did yield some interesting candidate groups, the addition of those cluster groups to my algorithms did not improve prediction accuracy. The following features were included in the final classification tests:

amount + pac\_id\_unique + pac\_id\_total + C(party) + C(race\_type) + C(incumbent) + C(clusters)

The prediction accuracy of Random Forest in this run was 96.9%, compared to a prediction accuracy of 96.2% for SVM and 84.7% for the dummy classifier.

**Discussion**

It is more or less a bromide to say “money influences politics.” However, the results of my analysis show that an election outcome can be predicted with nearly 97% accuracy just based off political donations. Feature importance from my Random Forest models showed that over 76% of the variance in the data could be explained by the amount of money donated to a candidate, the total number of PACs donating to a candidate, and the number of times any PAC made a donation to a candidate. It is also interesting to note that political affiliation as a Democrat or Republican only accounted for about 1% of the variance, possibly indicating that as both parties are becoming more polarizing, membership in either do little to improve a candidates chances of success.

As mentioned above, clustering yielded eight descriptive groups of candidate types during the 2012 election cycle. They are:

1. House Incumbents
2. House Democrat Challengers
3. House Republican Challengers
4. House Independent Challengers
5. House 3rd Party Challengers
6. Romney vs. Obama
7. House Libertarian Challengers
8. Unknown Presidential Hopefuls

There are far more groups of candidates running for congressional seats, as there are more congressional races than any other type in any given election cycle. What is interesting is that the only groups that received meaningful amounts of money were House Incumbents and Romney vs. Obama. This makes sense, as presumably PACs that have helped candidates win a seat in Congress have vested interests in maintaining the status quo. Additionally, since the office of the President of the United States is the most powerful in the world, PACs are clearly eager to throw gobs of money at the leading candidates in that race with hopes to influence the outcome and have some pull if their supported candidate wins.

This realization leads to another interesting question—what is the effect of the *Citizens United* ruling over time? A time-series analysis of the effect PAC expenditure on election outcome is warranted and would likely yield interesting results, particularly in non-presidential election years. Also not included in this analysis is the effect of PAC ideology and expenditure type (coordinated or independent) on election outcome. This is certainly an area that could be immediately investigated, as that data is readily available through the Center for Responsive Politics.

*ToDo:*

* Make cleaner data dictionary, including engineered features
* Include ideology codes and expenditure type in data set and re-run models, if possible
* Make slides for presentation

**Appendix A**

Data Dictionaries:

Candidates table-- [http://www.opensecrets.org/resources/datadictionary/Data%20Dictionary%20Candidates%20Data.htm](http://www.opensecrets.org/resources/datadictionary/Data%20Dictionary%20Candidates%20Data.htm" \t "_blank)

PACs table-- [http://www.opensecrets.org/resources/datadictionary/Data%20Dictionary%20for%20PAC%20to%20Cands%20Data.htm](http://www.opensecrets.org/resources/datadictionary/Data%20Dictionary%20for%20PAC%20to%20Cands%20Data.htm" \t "_blank)