General Assembly NYC
Data Science 24
Dylan Gessner
October 21, 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 chose the parameters of my models using grid search cross validation and I measured the performance of my tests using cross validation scores. I chose to measure success as prediction precision, since I want to predict the winner of an election as best I can.

Model 1— My first Random Forest model was able to predict election outcome with 79.5% precision, compared to SVM precision of 82.5% and dummy classification precision of 8.7%. 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 80.5% precision, 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 precision to 83.1%. The prediction accuracy of SVM also improved slightly to 83.8%, since SVM performs better when fewer features are included. Dummy classification also improved to 10%. 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 precision. 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 precision of Random Forest in this run was 83.7%, compared to a prediction precision of 82.7% for SVM and 8% for the dummy classifier.

Model 4 – I use logistic regression to test my prediction precision and yield the same result as when using Random Forest and SVM – 83%. When I remove the clusters from the feature set, I actually improve my prediction precision to its highest result yet – 84%. Since logistic regression provides feature coefficients, I use those coefficients to explain the meaning of my analysis in the discussion below.

Models 5-7 – I wanted to make sure that my model was not too noisy by including all three race-types (Presidential, Senatorial, and Congressional) in the same model at once. It's possible that campaign finance dynamics for one type of race is different than for another. I started by creating new features isolating the candidate on the race type, including only the most important features from Model 2 but replacing race_type with just the race in question, like so:

> amount + pac_id_unique + pac_id_total + C(party) + C(incumbent) + C(pres)

I found that when controlling for Presidential, Senatorial, and Congressional races, my model precision still remained in the 83-84% range. This reinforces the robust outcome of Model 4.

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 84% precision just based off campaign finance data. The coefficients from my logistic regression show that the number of PACs donating to a campaign and incumbency are hugely influential in predicting the result of an election. As the number of PACs that donate to a candidate increase by one standard deviation, the chances of that candidate winning the election increase threefold. Also, incumbents are more than twice as likely to be re-elected than their challengers are likely to take their seats. It is also interesting to note that in my Random Forest models, political affiliation as a Democrat or Republican were not important features, possibly indicating that as both parties are becoming more polarizing, membership in either do little to improve a candidates chances of success. Finally, by controlling for race type, I am able to conclude that the effects of campaign donations to a candidate are the same regardless of the race in which the candidate is running.

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

1. House Republican Challengers
2. Incumbent Congressmen and Senators
3. House Democrat Challengers
4. 3rd Party Longshots
5. Romney vs. Obama
6. Deez Nuts
7. Libertarian Challengers
8. Independent 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 the Incumbents, the House Democrat Challengers, 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, many Democrats were trying to win seats to maintain the Democratic majority in the House in 2012. The "Deez Nuts" cluster is a tongue-in-cheek reference to the satirical candidate created by Brady C. Olson in the most recent (2016) election cycle. This cluster is a grouping of "unknowns" that, regardless of election year, always garner some attention and can even manage to garner donations, however paltry those may be. Lastly, since the office of the President of the United States is the most powerful in the world, PACs donate huge sums of money to the leading candidates in the Presidential 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.

Appendix A

Data Dictionaries:

Candidates table--
http://www.opensecrets.org/resources/datadictionary/Data%20Dictionary%20Candidates%20Data.htm

PACs table--
http://www.opensecrets.org/resources/datadictionary/Data%20Dictionary%20for%20PAC%20to%20Cands%20Data.htm