Predicting GOP Congressional Campaign Performance

May 16, 2015

1.0 Introduction to Congressional Races

Elections for the U.S. House of Representatives occur every first Tuesday in November of even numbered years. Predicting the precise outcome of any one of the 435 elections[1] for voting members requires considering hundreds or thousands of unique and mostly qualitative factors like candidate quality, voter attitudes, perceptions of important issues, and effectiveness of the campaign itself. Predicting the results accurately can win individuals acclaim[2] it is much easier to do so for national and large statewide campaigns than for every congressional race. While there is often a plethora of information and data available for Presidential races, and even Senate campaigns (robust and regular polling, news coverage of events and their consequences) there is often very little of that same information available to the public for congressional races.

For two key reasons, focusing on every congressional race is not a priority because of their predictability: both the power of incumbency and partisan leaning of the district play outsized roles in whether or not a congressional campaign is successful. People hate congress but love their congressman, which is why 90% of incumbent congressmen won re-election in 2012[3]. Even in wave elections like 2010, 85% of incumbents won re-election.

Secondly, the partisanship of the congressional district has a tendency to overweigh other factors: Democrats vote for Democrats, and Republicans vote for Republicans, and if there are enough of one party in a district, the other party's candidate is irrelevant. This feature, called the Partisan Voter Index (PVI), is a function of both the partisan voter registration of the district and the performance of presidential candidates, makes predicting the outcome of seats where there is no incumbent (open seats), also relatively easy when the PVI is sufficiently favorable to a party.

However, tendencies are not perfect predictors, and identifying other factors that can influence the outcome of an election can help outside groups make better decisions.

1.1 Why Better Information Matters

In 2014, non-campaign organizations (Super PACs, social welfare organizations, party committees, etc.) spent over $790 million on congressional races to influence outcomes.[4] That money comes from hundreds of thousands of individual donors who demand that their money is put to the most efficient use.

[1] Elections for Congressional representatives in Washington, DC and Puerto Rico are not considered consequential because those members do not have any voting power in Congress.
[2] "Finding Fame With A Prescient Call For Obama," *New York Times*, November 9, 2008
[3] "People Hate Congress. But Most Incumbents Get Re-Elected. What Gives?" *Washington Post*, May 9, 2013
[4] Outside Spending, Center for Responsive Politics

However, many of these decisions are currently made in the blind. Polling on congressional races is thin, and even then, making decisions about which races to poll and when are done on intuition, experience, generalizations, and perceptions often created through selected media consumption.

The blind is the hardest to consider. Federal election law requires regular filing of campaign finance reports, often an excellent indicator of support. However, those reports are generally required to be filed on the 15th of each month following the end of a quarter (Q1 filed on April 15, Q2 filed on July 15, Q3 filed on October 15). Because the 3rd quarter report is filed just weeks before the election, it is often the focus of media scrutiny, but useless in making meaningful decisions that late in the election cycle. Therefore, the 2nd quarter report, filed on July 15, is where we are looking to gain insight into how a campaign will perform in November.

1.2 Can Finance Data Be An Early Warning Indicator?

The question we'll be attempting to answer is what aspects of campaign finance data, available by July 15th of the election year, be used to help predict the performance of a Republican congressional candidate.

We will do this by using a combination of district and campaign finance feature to predict the percentage of vote received by the GOP candidate in the general election.[5]

2.0 Data Mining

When gathering data on congressional races, we needed three discrete sets of data that could be effectively merged together: census data on the characteristics of the district, the partisanship voting history of the district (including congressional election results), and campaign finance data for the race-specific features that we are going to use to attempt to predict election results.

U.S. Census Bureau: Census data was acquired using API access to the American Community Survey 3-year statistics for congressional districts. The chosen features included population, Caucasian population, median income, median home value, and median rent value.[6]

Cook Political Report: The Partisan Voter Index (PVI) is a score calculated by the respected Cook Political Report that creates an rating, expressed as D+x or R+x where x is the degree of partisanship.[7] Because we are looking at only the performance of GOP candidates, the PVI has been normalized to positive for GOP-leaning districts and negative for Democrat-leaning districts. See Figure 3.1 for distribution of PVI scores.

Federal Election Commission: The Federal Election Commission has helpfully compiled recent historical summaries of campaign finance reports[8], including the report we are interested in (Q2 of election year). These reports, however, did not include reference to the specific congressional

[5] In political polling, the commonly accepted margin of error for a valuable poll is +/- 3.5%. Our models will work on a scale of 0-100, with a goal of RMSE of 5.0, conversationally equivalent to a +/- 5.0%.

[6] U.S. Census Bureau American Community Survey Data, http://www.census.gov/

[7] Cook Political Report explanation of how index is calculated, http://cookpolitical.com/

[8] Federal Election Commission Data Catalogue, http://www.fec.gov/

district. Consequently, the data had to be matched manually to races, which creates the risk for errors, although duplication and cross-verification of data has hopefully eliminated any error.

3.0 Data Features

When looking at which features are more or less indicative of voter behavior, we found that partisan performance strongly correlates to home value (Figures 3.2, 3.3), white/Caucasian population (Figure 3.4), and the population density of the district[9] (Figure 3.5).

Figure 3.1: *Distribution of Partisan Voter Index across all congressional districts.*

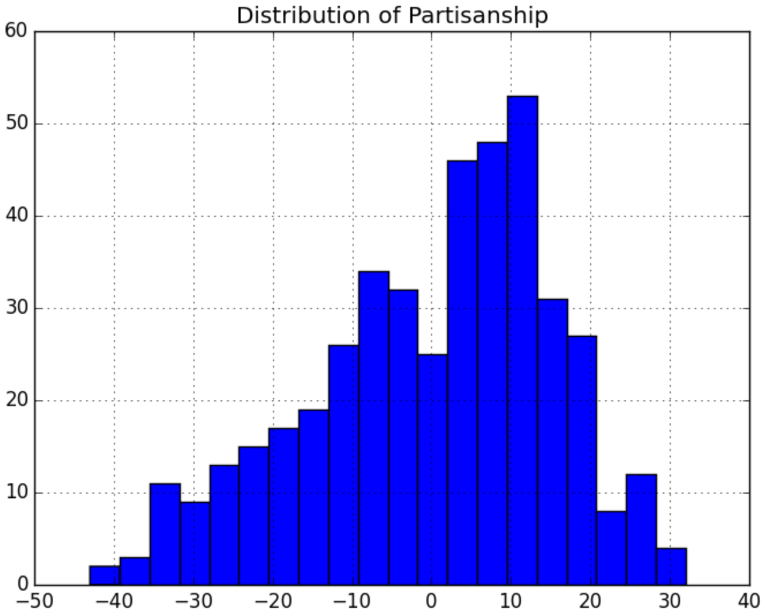


Figure 3.2 and 3.3: *Correlation between median home value and GOP Presidential candidate performance in 2008 and 2012*

[9] Similar studies have found a similar correlation between population density and the partisanship of the congressional district. See "If You Live Near Other People, You're Probably a Democrat. If Your Neighbors Are Distant, Republican" in *The Atlantic*.

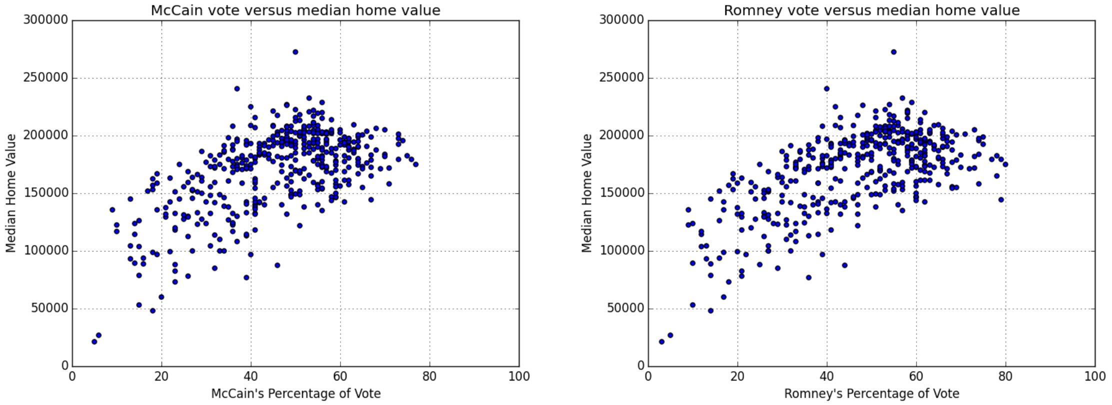
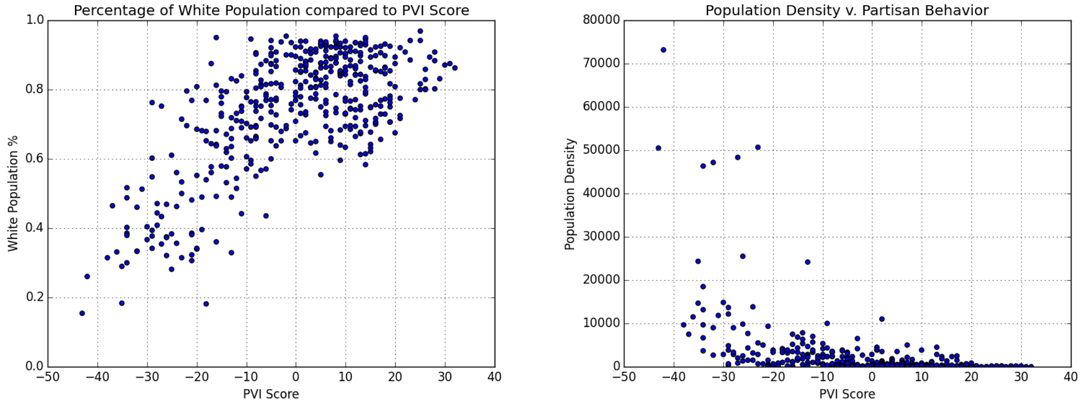


Figure 3.4 and 3.5: *Correlation between white population and population density to Partisan Voter Index score.*



Because of the high correlation between these characteristics and PVI, the PVI score is a dominant feature in predicting voter behavior, while features like past election results, population density, and district wealth characteristics, become much less significant.

After merging all of the data together, we were left with some difficult problems to solve. Although there are 870 races that occurred in 2012 and 2014, not all of them would be reasonable samples. The first step was to remove all the races where there was no Democrat v. Republican general election, due to lack of one side fielding a candidate. Thankfully, this is a common characteristic of districts so heavily partisan that the primary serves as the general election. Second, because we are looking at financial statements, campaigns where one candidate raised less than $1,000 by June 30 were considered too unviable to be considered. Combined, those account for 375 congressional races over the last 2 cycles.

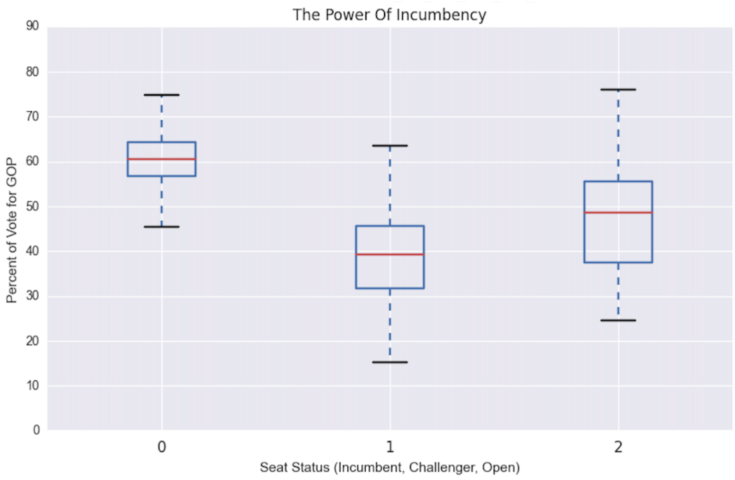
We are left, however, with a fairly robust set of 495 observations where 318 observations (64%) fall into the swing-district category, defined in Section 4.2.

This paper will only, for sake of brevity, discuss the important features. For an explanation of the all of the individual data features that were considered, see the data dictionary.[10]

4.0 Visualizing The Indicators

4.1 The Power of Incumbency

It can not be over-stated how powerful incumbency is in determining the outcome of an election. We categorized the status of each GOP general election candidate as Incumbent (1), Challenger where the incumbent was the Democrat (2), or an open seat where the previous incumbent was not participating in the general election (3).

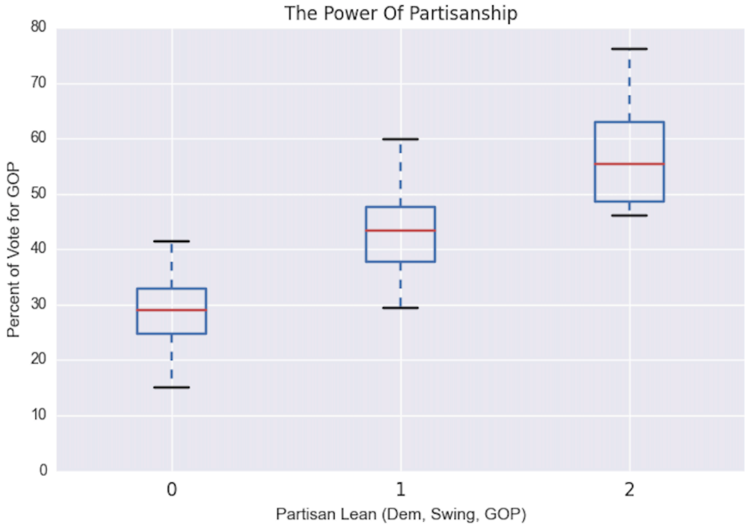


4.2 The Power of Partisanship

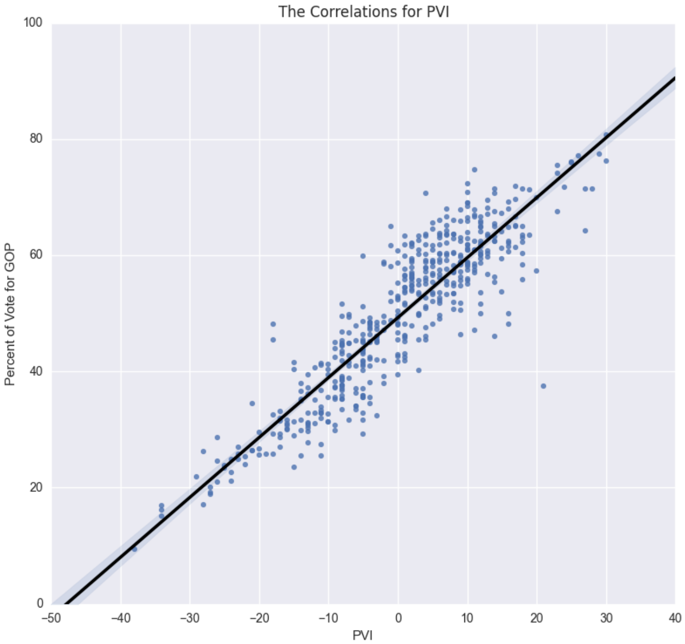
Campaigns are usually decided long before the election takes place. As you can see with the below chart, in districts where the PVI score is below -10 (described as a heavily Democratic district and labeled as 0 in the below chart), GOP campaigns max out at just above 40% of the vote. By comparison, districts that are heavily Republican (PVI > 10) almost guarantee the Republican candidate's election regardless of any other feature, with the minimum percent 48%.

It is the middle group of congressional districts, between -10 and 10 PVI, where the most variation occurs and the least amount of predictability. However, 207 of 435 districts exist within this range and are a significant enough sample size to provide insight.

[10] Data Dictionary: https://github.com/

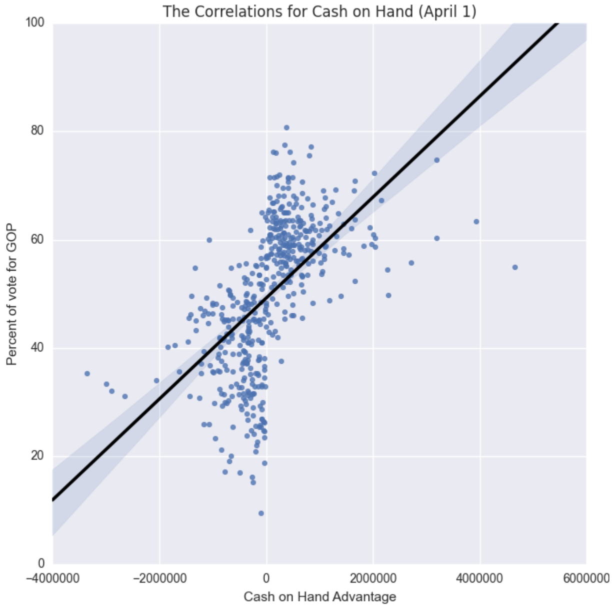


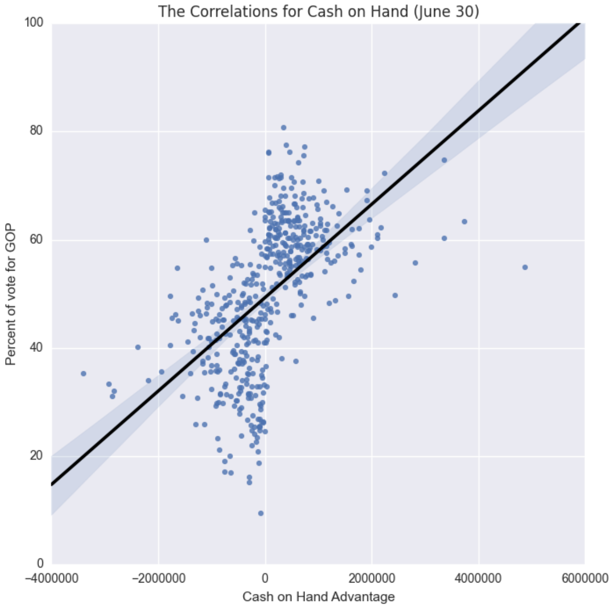
Another way of visualizing the data shows just how well PVI correlates to actual campaign performance.



4.3 The Campaign Finance Data

Although the individual cash on hand totals do not correlate strongly to the performance of the GOP candidate, the delta between two candidates does correlate. One of the most relevant insights gained from this project is that it is the delta of cash on hand at the beginning of the period that is more relevant than cash on hand at the end of the period in predicting the vote share in the general election seven months later.


The Correlations for Cash on Hand (April 1)


The Correlations for Cash on Hand (June 30)

5.0 Modeling The Races

To gain the insights we desired, we used three regression models to predict the percentage of the vote the GOP candidate should receive given a set of features. Because, as discussed above, PVI and incumbency are such strong indicators, we include them in every model. Incumbency was given a score of 1 for races with a GOP incumbent, and 0 for all other races.

5.1 Linear Regression

The first model we approached was simple linear regression. Because this model is highly interpretable, we hoped it would provide some insight into which features (outside of incumbency and PVI) would play a role in the eventual election outcome.

By looking at the two core features, PVI (gopvi) and Incumbency (inc_gop) only, we get a fairly satisfactory RMSE of 5.26368908322. This is fairly close to our ideal RMSE of 5.0. This also shows just how critical incumbency can be to winning, as the coefficient for being the incumbent is -7.6 – being the incumbent correlates to a 7.6-point increase over the intercept (45.8695), all other (financial) factors being equal. This is the power of incumbency.[11]

Next, we looked at two other critical features – cash on hand advantage at the beginning of the period (gop_cash_adv_beg) and at the end of the period (gop_cash_adv_end).

Interestingly, while both cash on hand at the beginning of the period (RMSE 5.11) and at the end of the period (RMSE 5.10) create slightly more accurate models, the two factors combined decrease the accuracy slightly to RMSE 5.12.

[11] The coef for incumbency does drop to 5.5 when cash on hand advantage for end of the period is taken into account.

No other combination of features performs better than PVI, Incumbency, and Cash on Hand Advantage at the End of the Period.

5.2 Decision Tree

Although not terribly useful in prediction, a decision tree is probably the most easily interpretable model, and can definitely provide some insight.

A combination of features produced an RMSE of 5.51. While this is not very useful relative to the much simpler linear regression model above, it did confirm several insights above, which you can see by looking at the relative importance of the features used to get the minimum error.

```
Feature Importance
gopvi             0.888646
gop_cash_adv_beg  0.080406
tot_rec_gop       0.023629
inc_gop           0.004956
gop_fund_adv      0.002363
```

While this model does not put a high priority on incumbency, it doesn't need to. Most districts are already represented by candidates of the favored party, so the district PVI is much more revealing. Additionally, in this model, and as we'll see in the random forest model, it is the beginning cash on hand delta that proves more insightful. In fact, the only way the model ever gets an RMSE below 5.6 is by excluding the cash on hand at the end of the period, a feature that linear regression found important, the decision tree found problematic.[12]

Neither of these models, however, give us what we really want.

5.3 Random Forest

The random forest is a costly model, but because we are working with only 495 observations, it is possible to 1,000 trees, making for a highly accurate model. Using the same features that our best performing decision tree identified, a random forest does even better.

We found that the random forest model can produce a RMSE score of just 5.0159, which is the closest to acceptable of any model and any feature selection. When we eliminated the fundraising advantage, our RMSE score did not change, and for simplicity of the model, we felt comfortable eliminating that as a relevant feature.

The feature importance breakdown is as follows:

```
Feature Importance
gopvi             0.675347
gop_cash_adv_beg  0.227494
tot_rec_gop       0.077886
inc_gop           0.019273
```

[12] View the complete decision tree: https://raw.githubusercontent.com/

Our goal was to identify the features that most inform the outcome of an election (percentage of vote for the Republican candidate on a 0-100 scale) and we believe the key features, after PVI and incumbency status, are the cash advantage at the beginning of the period, the total amount raised by the end of the period.

6.0 Predicting Performance for 2016

Thanks to the above models, we have some significant insight that can be used to help predict election outcomes in 2016 and, more importantly, identify under-performing and over-performing campaigns.

The first is that incumbents rarely require assistance. The power of incumbency correlates strongly to a 5- to 8-point increase over the mean, which alone is often enough to push a candidate over the 50% threshold (or just getting 1 vote more than anyone else). When an incumbent is in a swing-district, however, two features become exceptionally important.

The first is how much money the campaign has on hand when entering the 2nd quarter. The reason this is more important than the cash on hand at the end of the period is because of party primaries, many of which occur during the 2nd or 3rd quarter, and generally drain cash but do not diminish a candidate's standing if they are successful.

Campaign often put a lot of emphasis on 2nd quarter fundraising for the same reason we are looking at those reports – it's the last metric they can't fake until mid-October.

Therefore, when getting ready for 2016, we can gain some insight on April 15th by looking at a campaign's cash on hand advantage over their Democratic opponent. On July 15th, when looking at campaign finance reports, the total raised should be the only metric taken seriously.