Modeling Voter Preferences in US General Elections

Enze Chen, Eric Gong, Sahitya Mantravadi
Institute for Computational and Mathematical Engineering
Stanford University
{enze, ericgong, smantra}@stanford.edu

Abstract

The ability to model voter preferences and predict election outcomes is an invaluable asset for politicians on the campaign trail. In this report, we present a series of models to predict voter percentages for Senate candidates in the US general election from historical voter data. By aggregating the data from Senate, House, and Presidential elections, we find an average absolute error of 6.9% when trying to predict the 2016 Senate elections. Our results suggest that decay weighted averaging provides the most robust predictions.

1 Introduction

The practice of campaigning is a distinctive feature of democratic governments, during which candidates up for election spend valuable time, resources, and energy into securing votes from their constituency. However, because resources and time are limited, candidates must also be strategic about when and where to campaign, and thus they should concentrate their efforts among populations where the race is tight. While this is an intuitive approach, the actual determination of tight races is highly non-trivial. There exists a fundamental uncertainty in how people will vote and how campaigning in a particular area will influence how people vote in that area. This uncertainty stems from a variety of sources that would be impossible to enumerate in full, thus creating the need for a machine learning model that can capture these sources of uncertainty and accurately predict how people will vote.

Consequently, we cast this problem as one of decision making under state uncertainty, where the true proportion of voters for a specific party (assuming a ground truth even exists) is unknown. If we view voter preferences as a continuous distribution that evolves over time, we can then consider polling and election data as samples (observations) that reveal information about the underlying state. Analyzing the voting data from previous years with the appropriate model would enable us to make predictions on voter preferences in successive years, refining our belief about the state and thereby allowing candidates to make informed campaign decision based on their likelihood of winning.

2 Related Work

Given the fundamental roles voting and elections play in shaping our nation's progress, it is unsurprising that predictive voter modeling is a widely studied phenomenon. There have been comparisons of various deterministic and probabilistic models for predicting voter preferences [1], and the relative success of each type of model was highly dependent on data availability and one's set of assumptions. More recently, the probabilistic models produced by FiveThirtyEight have attracted significant public attention [2], and others yet have attempted to integrate sentiment from social media sources to create more informed models [3, 4]. The growing body of literature both producing and critiquing such models [5, 6] highlight the urgency and complexity of this problem that deals with the inherently capricious nature of human preferences.

3 Data

For this report, we use the data sets provided by the MIT Election Data Science Lab [7]. The Election Data Science Lab provides data sets on federal, state, and local elections on a state, district, county, and precinct level. We focus on federal elections for this report. In particular, we utilize the data on US Presidential, Senate, and House elections. These data sets contain the candidate's name, political affiliation, and votes received for a year's election in a particular locale.

When creating our data pipeline, we decided to focus on the two main parties—Democrat and Republican. Based on this, we were able to calculate the vote percentage that each candidate received in that locale for a given election, normalized to the two parties. For example, in 1976, the Democratic Candidate for the Presidency, Jimmy Carter, received 65,9170 votes out of a total of 1,182,850 Democratic and Republican votes cast in the state of Alabama. From this we are able to determine that Carter received approximately 55.7% of the vote in Alabama during the 1976 US Presidential Election. With this additional step of calculation, we were able to gather all necessary data for this report. Moving forward, we aim to predict the Democratic vote percentage for the 2016 Senate elections, as the Republican vote percentage would be the complement of the value we predict.

4 Models and Methods

4.1 Baseline

We construct a baseline model to produce a performance benchmark against which we can compare our more intelligent model. This baseline model naively assumes that voter preferences have not changed between successive election cycles and thus directly uses the vote percentage from the most recent election as the predicted vote percentage for the next election cycle.

4.2 Averaging and Weight Decayed Averaging

To determine the political leanings of a given district or state, our model observes how this region voted in past elections. Two models we implemented were averaging and weight-decayed averaging. The implementation for the averaging model was simply averaging the voting percentages over all years to predict voting percentage for 2016 Senate elections. This weights all years equally. The political landscape does, however, change over time. Consequently, we want to place more weight on recent election data when compared to elections from four decades ago. Thus, we introduce a decaying weight coefficient to capture this intuition. As an election dates further and further back from the current year, we give less and less weight to the data from that election.

For US elections for a 40-year time frame, we define a weight w as

$$w_i = \gamma \cdot \frac{\text{Election year}_i - 1974}{2016 - 1974} = \gamma \cdot \frac{\text{Election year}_i - 1974}{42}, \quad \text{such that} \quad \sum w_i = 1$$

Then we solve for γ when we have elections every two years from 1976 to 2016:

$$\gamma \sum_{i=1}^{21} \frac{\text{Election year}_i - 1974}{42} = 1$$

$$\gamma \cdot 10 = 1$$

$$\gamma = \frac{1}{10}$$

We then compute a weighted average of voting percentages over the years using the weights w_i to construct sentiment for the state to predict voting percentage for 2016 senate elections.

4.3 Regression

In addition to the weight-decayed averaging model where the weights were chosen based on intuition and a single parameter γ, we also constructed a linear regression model to estimate the weights based on the data. Given a history of k election cycles, we use the voting percentages from the k cycles

$(p_1, \ldots, p_k)$ as features to predict the voting percentage in the $(k+1)$th cycle (p_{k+1}). This allows us to formulate the problem in the form of $\vec{y} = X\vec{w}$ where

$$\vec{y} = \begin{bmatrix} p_{k+1} \\ p_{k+2} \\ \vdots \\ p_n \end{bmatrix} \qquad X = \begin{bmatrix} p_1 & p_2 & \cdots & p_k \\ p_2 & p_3 & \cdots & p_{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n-k} & \cdots & \cdots & p_{n-1} \end{bmatrix}$$

and $\vec{w}$ are the weights to be learned by computing $\vec{w} = (X^T X)^{-1} X^T \vec{y}$. One advantage of this approach over the other methods is the flexibility of incorporating features beyond just voting percentages (e.g. incumbent status, presidential party affiliation, scandals) simply by appending extra columns to the X matrix. With the learned set of weights, we can then make a prediction on $\hat{p}_{n+1}$.

We experimented with using voting percentages from different types of elections (Presidential, Senate, and House). Presidential elections tend to be closer than Senate or House elections, but incorporating the voting percentages of all three types of elections for the prior three years allows the model to use more information than purely past Senate data. Then, we use the regression model to predict voting percentage for the 2016 Senate elections.

4.4 Error metric

There are many possible ways to quantify model performance, including absolute error of the predicted voting percentage $|\hat{p}_{n+1} - p_{n+1}|$, relative error (e.g. normalized by the average fluctuation), or binary 0/1 based on a threshold. Ultimately, we chose to use absolute error of the predicted voting percentage, i.e. the absolute value of the difference between true Democratic percentage vote and the predicted Democratic percentage vote. This gave the clearest translation to computing aggregated error metrics, i.e. average and standard deviation, for all states.

5 Results

When we run each of the three models, withholding 2016 data, and predict the 2016 Senate election outcomes, we compute prediction error for each model for each state. The three errors for select states are displayed in Figure 1.

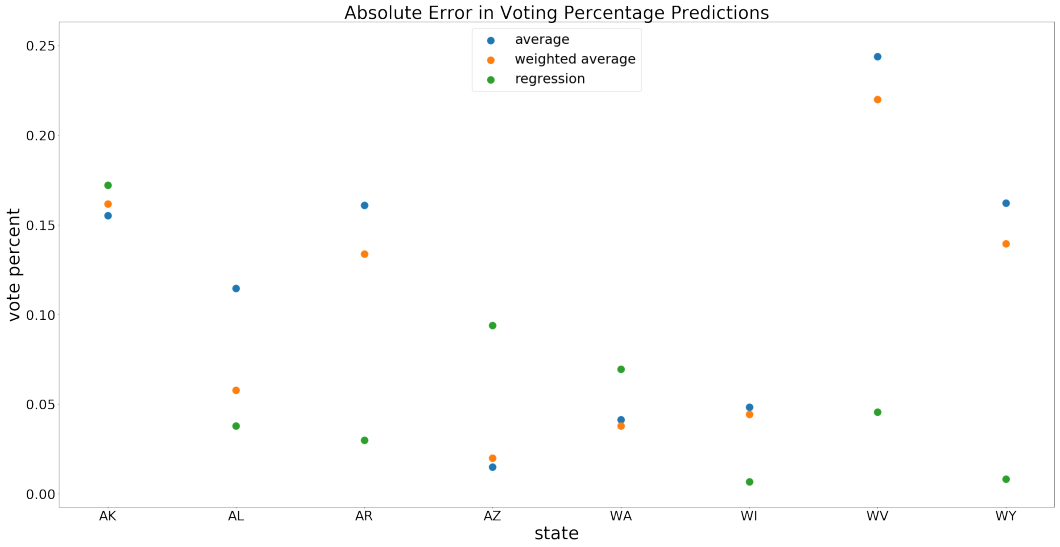

Figure 1: Errors (fractions of percents) for each method are fairly low across all models; error is especially low for the weighted average model and regression model.

The average error in voting percentage for the 2016 Senate elections predicted by each of the methods is displayed in Table 1. We can see that although the average absolute prediction error for the baseline

Method	Average Error	Standard Deviation of Error
Baseline	0.0731	0.1010
Averaging	0.0842	0.0594
Weighted Averaging	0.0692	0.0536
Regression	0.0713	0.0523

Table 1: The decay-weighted averaging method has the lowest average absolute prediction error, while the regression method has the lowest standard deviation of absolute prediction error.

method seems, at first glance, fairly low, the method has almost double the standard deviation of absolute prediction error as the other methods. It is understandable that the baseline method has high variance—in many cases, voting preferences in a state are fairly similar from one year to the next (which explains the method's fairly reasonable average prediction error), but since only one year of data is being used to make each prediction, standard deviation is high. Meanwhile, the averaging method has the highest average absolute prediction error but fairly low standard deviation compared to the baseline method. This is explained because averaging over many years decreases the standard deviation of our prediction error, but it is not reasonable to assume that recent years and far past years contribute to the next election cycle the same amount. Thus, the averaging method has high average prediction error. The weight-decayed averaging method and regression method both have low average prediction error and low standard deviation of prediction error. The slightly lower average error of the weight-decayed averaging method compared to that of the regression method is best explained by the fact that the weight-decayed averaging method uses all past data and weights accordingly, while the regression prediction is based only on the previous three years of data.

6 Discussion

We recognize that our model considers only historical statistical factors and does not account for political expectations. For example, particular states or districts are simply deemed unworthy of a candidate's time. In actuality, a candidate would most likely attract bad press by simply bypassing a number of states or districts.

At the same time, one must recognize that US politics is ripe with behavior that derives its justification based on tradition or convention. The act of campaigning has changed since its earliest ones seen at the onset of the American republic. With the prevalence of social media and a drastically increased digital presence among nearly all Americans, it may also be worth reexamining the role of physical political campaigning. Campaigning in the digital world, it must be noted, has nearly no marginal cost in time for a political candidate. With this in mind, one must also remember that Hillary Clinton's decision not to campaign in particular states in the weeks prior to the 2016 Election Day was the subject of much scrutiny.

We believe that it would be prudent for political candidates to consider a number of factors when deciding where and how to campaign. Our model can help provide valuable input for making this decision, but a political team must also consider other factors including personal characteristics (e.g. charisma) and biases (e.g. race/gender). All in all, a quantitative approach to decision-making in this scenario has its place, and this project demonstrates the importance of using data from past elections to predict the future.

7 Author Contributions

All authors contributed equally to the design of the experiments, implementation of the models, analysis of the results, and writing of the report.

Code for this report can be found at `https://github.com/sahityamantravadi/cs238-final-project/`

References

[1] Barry C. Burden. Deterministic and probabilistic voting models. *American Journal of Political Science*, 41(4):1150–1169, 1997.

[2] How FiveThirtyEight's house model works. https://fivethirtyeight.com/features/2018-house-forecast-methodology/, 2018.

[3] Nicholas Beauchamp. Predicting and interpolating state-level polls using Twitter textual data. *American Journal of Political Science*, 61(2):490–503, 2017.

[4] Andrea Ceron, Luigi Curini, Stefano M Iacus, and Giuseppe Porro. Every tweet counts? how sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to italy and france. *New Media & Society*, 16(2):340–358, 2014.

[5] D. Gayo-Avello. No, you cannot predict elections with Twitter. *IEEE Internet Computing*, 16(6):91–94, Nov 2012.

[6] P. T. Metaxas, E. Mustafaraj, and D. Gayo-Avello. How (not) to predict elections. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 165–171, Oct 2011.

[7] MIT election data and science lab. https://electionlab.mit.edu/data, 2017.