Ronia Hurwitz & Katherine Playfair

Applied Mathematics 50

April 2016

Predicting the 2016 Presidential Candidacy

Introduction

The 58th quadrennial presidential election in the United States will take place this fall on November 8, 2016. As of April, the race appears to be rather close, with five candidates left vying for the presidency. These candidates consist of three republicans: Ted Cruz, John Kasich, and Donald Trump; and two democrats: Hillary Clinton and Bernie Sanders. Many people are overwhelmingly invested in the outcome of this race, as it will dictate the future of our country’s political, economic, and social standing. In addition, if a democrat wins, as many reliable polls have predicted, a rare occurrence will result in which one political party will win more than two consecutive terms. This has only happened three times since 1948, most recently in 1988 when George H. W. Bush replaced the two-term republican president, Ronald Regan. Given that Barack Obama has now served America for his eight-year limit, another Democrat president is a statistically unlikely event. We set out to determine why the polls tell us differently. There are a number of models and algorithms currently in existence that attempt to predict the results of this election. They incorporate the political lean of each state and peoples’ preferences regarding a range of political issues. We would like to create a model that

1 This project was completed in April 2016 before Cruz and Kasich dropped out of the race. Further explanation of the impact of these occurrences can be found in the May Update.
2 http://elections.huffingtonpost.com/pollster
can accurately predict the presidency, but rather than using these measures to infer election results, we would like to forecast the winner using only relevant demographic information. We prove that we are able to successfully associate certain demographic data to infer how a person will vote.

In order to do so, we produced a model of our own to estimate the number of votes each presidential candidate would obtain if all five candidates were to run in the election. We pursued results through two applications of our model, one which utilized General Social Survey statistics on a representative population consisting of 2538 individuals and the other which used demographic statistics on each of the fifty states plus Washington, D.C. We obtained results for the percentage of votes each candidate would receive from these two sources and reran the model for the top democrat and republican nominee to simulate the presidential election with two opposing candidates. Next, we compared the results from the individual versus state data to one another to determine whether our model was consistent across various populations. Through analysis of our model’s results, we made conclusions about our confidence in the predictions obtained, and what this indicates for the upcoming presidential election in America. Finally, we analyzed the outcomes from two pollsters, PredictWise and Gallup, that have accurately predicted the presidency in previous elections to implement as confirmation of our model’s results.

Methods & Results

To produce a model that predicted how a person would vote according to demographic information alone, we needed to collect two major sets of data. First, we needed statistics
summarizing the demographics of our intended population. Second, we needed data that outlined the demographics of the ideal voter for each of our five candidates. We decided to make a model based on the assumption that a person will vote for the candidate whose relevant values are most in line with their own. Therefore, the voter will cast their ballot for the candidate with an “ideal voter” with whom they share the most demographic information. We then had to determine which demographic qualities we should incorporate into our model and how those qualities reflected which candidate a person would vote for. We decided to split this model into two areas of application. The first included General Social Survey (GSS) statistics detailing the preferences of 2538 individuals regarding a wide range of topics. The second application utilized state data obtained from the United States Census Bureau and other reliable surveyors. Due to discrepancies between the data that surveyed from each of these groups, we analyzed similarly indicative, yet different factors for each of these two populations.

For the GSS data, we chose to predict voter selection from the following seven demographic factors: socioeconomic index, whether they approve of abortion for any reason, the frequency at which they attend church, opinions on immigration, the role of women in politics, military stance, and whether guns should be permitted. After researching the campaign platform of each candidate, we deemed these seven preferences to be indicative of their support for a particular nominee. We obtained responses from all 2538 people regarding each of these opinions and normalized the data to be analyzed on a comparable scale. Next, we investigated the policies proposed by each candidate on these seven issues. We used this data

---

4 [http://gss.norc.org/Get-The-Data](http://gss.norc.org/Get-The-Data)
5 [https://www.census.gov/hhes/www/income/data/statemedian/](https://www.census.gov/hhes/www/income/data/statemedian/)
to compute the normalized values that a respondent would provide if they were the hypothetical ideal voter for a given candidate\(^6\). For example, the ideal voter for Clinton has a slightly above-average socioeconomic index, approves of abortion, attends church less than the average American, supports immigration, prefers to have equal representation of genders in politics, feels that our military involvement should neither increase nor decrease, and does not permit gun access to all citizens. The following relationships for each of the candidates’ preferences are outlined below.

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Socioeconomic Index</th>
<th>Abortion Approval (yes=1, no=-1)</th>
<th>Church Attendance</th>
<th>Exclude Immigrants (agree=1, disagree=-1)</th>
<th>Women not in Politics (agree=1, disagree=-1)</th>
<th>Military</th>
<th>Gun Permits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinton</td>
<td>0.5</td>
<td>1</td>
<td>-1</td>
<td>-0.5</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>Cruz</td>
<td>1.5</td>
<td>-1</td>
<td>2</td>
<td>-0.5</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Kasich</td>
<td>2</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>0.5</td>
</tr>
<tr>
<td>Sanders</td>
<td>-1</td>
<td>1</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>Trump</td>
<td>-1.5</td>
<td>-1</td>
<td>1.5</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
</tr>
</tbody>
</table>

In order to determine which candidate each voter will select, we calculated the Euclidian distance between the values for the individual and the values for each candidate\(^7\). We stated that the voter will then choose the candidate that he or she is closest to when considering all seven factors. Therefore, we assigned one vote to a candidate each time they attained the minimum distance to a particular individual, calculating distance with the formula below.

\[
\text{Euclidean distance} = \sqrt{\sum_{i=1}^{v} \left(\frac{(p_{1i} - p_{2i})^2}{md_i}\right)}
\]

---

\(^6\) We made candidate predictions using various sites and sources, reported in the References section.

...where \( p \) is the position of the data point and \( m_d_i \) is the maximum squared discrepancy per variable \( i \) of \( v \) total variables.

After running this algorithm, we obtained a certain number of votes for each candidate, which summed to the total 2538 individual votes. This raw data can be observed below, along with a bar plot presenting the distribution of votes.

**INDIVIDUAL VOTES:**

- Trump: 456
- Clinton: 834
- Sanders: 799
- Cruz: 379
- Kasich: 70
- Total: 2538

**PERCENTAGES:**

- Trump: 17.967
- Clinton: 32.861
- Sanders: 31.481
- Cruz: 14.933
- Kasich: 2.758
- Total: 100.00
Seeing that Trump was the leading republican candidate and Clinton was the leading
democratic candidate, we decided to follow up with these results to determine a winner if
Trump and Clinton were the only two nominees in the running. This would simulate the the
presidential race that would occur in November once the primaries were complete. From this
analysis, we obtained the following outcome.

<table>
<thead>
<tr>
<th>INDIVIDUAL VOTES:</th>
<th>PERCENTAGES:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trump: 764</td>
<td>Trump: 30.102</td>
</tr>
<tr>
<td>Clinton: 1774</td>
<td>Clinton: 69.898</td>
</tr>
<tr>
<td>Total: 2538</td>
<td>Total: 100.00</td>
</tr>
</tbody>
</table>

We then repeated this process for the state data, aggregating votes via each state’s electoral
votes rather than on an individual level. Because the statistics made available to us differed, we
analyze slightly different demographic data to predict voter selection. We chose to examine the
following five demographic factors to indicate the political choice of each state: average income, percentage of the population that lives in an urban area, proportion of evangelicals, percentage of the population that has obtained a bachelor’s degree or higher, and average age. We obtained this data from all fifty states as well as the District of Columbia. Next, we researched the candidates’ policies regarding these five topics and predicted the responses of an ideal state for each candidate. The normalized values we estimated per state are presented in the chart below.

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Income</th>
<th>Urban Population</th>
<th>Evangelical</th>
<th>Bachelor's Degree</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinton</td>
<td>0</td>
<td>2</td>
<td>-1.5</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>Cruz</td>
<td>0.5</td>
<td>-1</td>
<td>2</td>
<td>1</td>
<td>1.5</td>
</tr>
<tr>
<td>Kasich</td>
<td>1</td>
<td>1.5</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Sanders</td>
<td>-0.5</td>
<td>0.5</td>
<td>-1.5</td>
<td>1.5</td>
<td>-1.5</td>
</tr>
<tr>
<td>Trump</td>
<td>-1</td>
<td>-2</td>
<td>0</td>
<td>-1.5</td>
<td>2</td>
</tr>
</tbody>
</table>

Using the same Euclidian distance formula as in the previous model, we forecasted how each state would vote and assigned their respective number of electoral votes to that candidate. Running the algorithm for each of the states, our program produced the following data, which is also presented as a bar plot to show the relative percentages of votes acquired by the five candidates.

**ELECTORAL VOTES:**
- Trump: 138
- Clinton: 185
- Sanders: 91
- Cruz: 56

**PERCENTAGES:**
- Trump: 25.651
- Clinton: 34.387
- Sanders: 16.914
- Cruz: 10.409

---

8 We made candidate predictions using various sites and sources, reported in the References section.
Once again, we see that Trump was the leading republican candidate and Clinton was the leading democratic candidate, so we repeatedly decided to follow up with these results to determine a winner if Trump and Clinton were the only two nominees in the running. From this analysis of the states, we obtained the following outcome.

<table>
<thead>
<tr>
<th>INDIVIDUAL VOTES:</th>
<th>PERCENTAGES:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trump: 205</td>
<td>Trump: 38.104</td>
</tr>
<tr>
<td>Clinton: 333</td>
<td>Clinton: 61.896</td>
</tr>
<tr>
<td>Total: 538</td>
<td>Total: 100.00</td>
</tr>
</tbody>
</table>
After producing a model and applying our algorithms to these two sets of data, we sought to further investigate the legitimacy of these results. To gain a better understanding of the effect of the electoral voting system, we created a plot of the United States, showing the distribution of electoral votes by state. This graph allowed us to determine the effect of aggregating votes through individuals, as we do in our first application of the model, versus determining the winning candidate of a state and subsequently aggregating electoral votes, similar to the process of our second application.
From this plot, we recognize that although the number of electoral votes is somewhat correlated to the population of each state, the ratio is not directly one-to-one. Thus we should expect to obtain different results when applying our model to individuals compared to state data. The fact that our models did not produce the same rankings of candidates for the GSS data and the state demographics may be a result of this discrepancy in determining a winner.

Next, we used statistics from reputable pollsters, PredictWise and Gallup, to create prediction models to either confirm or reject the model that we produced. PredictWise posts the probability in each state that a democratic candidate will win their delegated number of electoral votes\(^9\). Using these probabilities, we plotted a map of the United States, showing the likelihood that a democratic candidate will win in each state.

\(^9\) http://predictwise.com/politics/2016-president-winner
At first glance, it appears that a majority of the states are colored in red, representing a win by a republican candidate. Because this plot does not take into account the number of electoral votes, which we plotted above, we decided to run simulations that would determine an outcome taking into account both the political lean and electoral votes per state. We ran 10,000 simulations of an election that reflects the political bias that we obtained from PredictWise data. We then produced a histogram that plotted the frequency of a democratic candidate receiving enough electoral votes to win the presidency over these 10,000 simulations.
From these simulations, we found that the most frequent outcome to occur over these simulations is for the democratic candidate to receive somewhere between 275 to 325 electoral votes. In order to win an election, a candidate must receive a total of 269 electoral votes, as represented by the black line above. Therefore, we find that a democratic candidate has an 80.59% chance of winning this election, given the political lean determined by PredictWise for each state.

In order to provide a last method of validation for our model, we ran similar simulations on statistics obtained from the pollster, Gallup. These statistics also log the political lean of each state, using different slightly different data from PredictWise to obtain such numbers\textsuperscript{10}.

\textsuperscript{10}http://www.gallup.com/poll/188969/red-states-outnumber-blue-first-time-gallup-tracking.aspx?g_source=heavily%20democratic&g_medium=search&g_campaign=tiles
This additional verification process provided us with a thorough method to detecting the likely outcome when predicting the presidency according to the political stance of our country. We similarly graphed a map of the United States, presenting the probability of a democratic win, but this time we incorporated uncertainty of the model’s accuracy. Due to Gallup’s inability to poll every voter in America, they suggest that there is a sampling error of approximately 3% for each state. Therefore, we use the following Cumulative Distribution Function (CDF) to account for the uncertainty that may exist in each states’ record of political lean\textsuperscript{11}.

\[
CDF(z) = \frac{1}{2} \left( 1 + \text{erf} \left( \frac{z - \mu}{\sqrt{2} \sigma} \right) \right)
\]

...where \text{erf} represents the Guass error function, \(z\) is the threshold value, \(\mu\) is the mean, and \(\sigma\) is the standard deviation of the distribution.

\textsuperscript{11} http://mathworld.wolfram.com/NormalDistributionFunction.html
From this data, we are able to observe that the results are much less obvious. With the inclusion of uncertainty, we find that many more states fall in the middle of the spectrum, rather than at the extreme values of 0.0 or 1.0, as we saw in the PredictWise model without this sampling error. Using these values that incorporate, we again run the data for 10,000 simulations of the election to determine whether we still see consistencies between these models and the model we produced at the start.
We found that even when incorporating the sampling error of 3% per state, there was still a 97.73% chance of a democratic president according to the political lean of each state obtained by Gallup. We then extended our comprehensive analysis of this data by analyzing the effects of bias on the predictions. We incorporated potential bias that could have arisen due to a number of causes: the data collected is unrepresentative of the population due to response bias or other flawed polling methodology, particular party affiliation does not indicate a vote for that party’s candidate, the trend in voter affiliation is not consistent leading up to the date of the election, etc. Because we were unable to calculate exact measures for these biases, we ran simulations with estimated levels of 1% and 5% bias. This meant that we shifted the democratic lean of each state by the amount of estimated bias, decreasing the likelihood of confidently determining a presidential winner.
The left and right graphs represent 10,000 simulations run from the Gallup data with 1% and 5% bias, respectively. When we estimated the data to only involve 1% bias, we see that the chance of a democratic president drops from 97.73% to 84.03%. This difference of almost 14% is significant, but ultimately does not significantly influence the outcome of the election. With 5% bias, we see a strikingly dramatic decrease in likelihood to 0.23%. The chances of a democratic president decreased exponentially when we took into consideration these potential biases within the model.

**Discussion**

From both applications of our model, we found that Trump and Clinton were the respective republican and democratic candidates with the greatest proportion of votes. This consistency helps to build an argument in support of our model’s accuracy. Since both applications pointed to these candidates, and more specifically, Clinton, leading the race, we can place more confidence in the precision of our process. We also have support for our model’s results in the verification we received from PredictWise and Gallup analysis. Given that both sources predicted a democratic president, we can validate our model’s prediction of Clinton’s victory.
occurring as well. Through this thorough process of investigation, there is great justification to support the theory that a democratic president will win in the upcoming election.

While the distribution of votes were, for the most part, in accordance between the two applications, there were inconsistencies amongst the exact candidate rankings. We can attribute these differences to a number of limitations of our model. First, it is possible that the GSS survey results were not representative of the American population. It is also possible that the specific factors we analyzed are not directly indicative of political preference. Lastly, the preferences of the voter may not have equal weight on their voter choice. These limitations were not accounted for in our model, so our predictions should be approached with apprehension before blindly accepting them. In addition, deeper analysis of the Gallup data led us to believe that our results may invoke false confidence due to uncertainty and biases. While we recognize that we cannot place 100% confidence in our model, we feel that is provides a reliable method in predicting the winner of the 2016 presidential election.

May Update

Because this model was created in April 2016, we incorporated both Cruz and Kasich in our algorithms. Both candidates ultimately dropped out of the election in early May of 2016. This further supports our work because both applications of our model predicted that these two candidates would receive the least number of votes, justifying their decision to remove themselves from the race. If we were to further extrapolate future events of the election, we expect to see that either Sanders will drop from the democratic primaries, or he will maintain his position in the race but Clinton will acquire more votes to win the democratic primary. We
predict that Trump and Clinton will then face off in the presidential election, and Clinton will be victorious.
References

http://archive.star.txstate.edu/node/3851.html
http://elections.huffingtonpost.com/pollster
http://gss.norc.org/Get-The-Data
http://predictwise.com/politics/2016-president-winner
http://www.270towin.com/
http://www.gallup.com/poll/156437/heavily-democratic-states-concentrated-east.aspx#2
http://www.gallup.com/poll/188969/red-states-outnumber-blue-first-time-gallup-tracking.aspx?g_source=heavily%20democratic&g_medium=search&g_campaign=tiles
http://www.huffingtonpost.com/entry/john-kasich-is-seemingly-baffled-by-young-women-who-get-politics_us_570c07fde4b0836057a2153c
http://www.ontheissues.org/Immigration.htm
http://www.pbarrett.net/techpapers/euclid.pdf
http://www.pewforum.org/2016/01/27/faith-and-the-2016-campaign/
http://www.theatlantic.com/politics/archive/2016/03/who-are-donald-trumps-supporters-really/471714/
http://www.thetrace.org/2016/01/presidential-candidates-views-guns-2016/
https://github.com/cs109/content/blob/master/HW2_solutions.ipynb
https://newrepublic.com/article/131762/hillary-voter
https://www.census.gov/hhes/www/income/data/statemedian/