Semester Project

Elizabeth McAvoy[1,2,3]

Abstract

Electoral forecasting is important for campaigns and political actors, as it influences campaign strategy and political strategy. This report analyzes U.S. voting behavior, including models to predict 2016 presidential election voting behavior. In presidential elections, political party appears to have the greatest effect on voting behavior. A large majority of Democrats and Republicans voted for their party's presidential nominee in the both the 2012 and 2016 presidential election. To further explore these findings, I analyzed whether voters consistently vote for the same party. Across four different positions, U.S. House of Representative for 2010 and 2012 and U.S. President for 2012 and 2016, 79% of respondents voted for their party's nominee in all four positions. 84% of Democratic respondents and 82% of Republican respondents voted for their party's nominee in all four positions. To model 2016 presidential voting behavior, I ran four different algorithms—K-nearest neighbors, Naive Bayes classifier, penalized multinomial regression, and random forest—using 10-fold cross-validation. Of the four algorithms, random forest performed the best at predicting 2016 presidential voting, receiving an accuracy of .9327 and AUC of .9921 on the test dataset. Additionally, random forest correctly predicted 97% and 98% of respondents who voted for Clinton and Trump, respectively.

Keywords

Electoral Forecasting — U.S. Presidential Elections — Political Parties

[1] *O'Neill School of Public and Environmental Affairs, Indiana University, Bloomington, IN, USA*
[2] *Philosophy, College of Arts and Sciences, Indiana University, Bloomington, IN, USA*
[3] *Political Science, College of Arts and Sciences, Indiana University, Bloomington, IN, USA*

Contents

1. Problem and Data Description

Campaigns and political actors rely on data for campaign and voting strategy. For example, campaigns must decide where to target their outreach to persuade the number of people necessary to win the election. Political actors, such as elected officials, need to know how their constituents feel on an issue in order to represent their interests in governing, Also, political actors must decide which candidate is likely to win an election and which elections to invest money into. In this report, I analyze voting behavior to predict how someone is going to vote given some information, such as certain demographic information.

The data come from the VOTER (Views of the Electorate Research) Survey, conducted by YouGov, an international, private survey firm. The data is publicly available on Kaggle.[1] YouGov surveyed 8,000 U.S. adults between November 29 and December 29, 2016 about their voting behavior and beliefs. The unique part of this survey is that these respondents were also interviewed by YouGov previously in 2011 and 2012, so this allows for time-series analysis, albeit limited. In total, every respondent was interviewed three times, first in December 2011, second after the 2012 general election, and finally after the 2016 general election. The surveys contain many questions, totaling to 668 variables in the dataset.

2. Data Preprocessing & Exploratory Data Analysis

2.1 Data Cleaning

Because the data were collected by surveys, I expect that there will be questions that were not answered, not answered correctly, or skipped (the respondent refused to answer or did not know how to answer). This section explains how these variables were cleaned. In the exploratory data analysis, nonrespondents are excluded because they are rare. However, in Section 3, an algorithm was implemented to predict these missing responses, which are used to build and test the models, because there are many missing values in the dataset. Only 157 observations have no missing values across all the variables.

2.1.1 Race

Respondents were asked about their race in the first survey in 2011, as well as in the third survey in 2016, In both surveys, respondents were asked to select their race as Asian, Black, Hispanic, Middle Eastern, Mixed, Native American, White, or Other. In 2011 and 2016, 169 of the 8,000 respondents selected Other as their race or did not answer the question. These 169 respondents were not all the same respondents in the two years; only 63 respondents selected Other or did not answer the question in both years. When respondents selected Other as their race, they were asked to explain their race. Most of these responses show that the respondent belongs in one of the other seven race categories, meaning that they incorrectly identified their race. For example, one respondent wrote that his race is ”White, Black, Hispanic,” which falls into the Mixed race category. Because race is time-invariant, meaning that it does not change over the course of someone’s life, I use the respondent’s response from the other survey to recode his race into one of the seven provided race categories. I was able to recode all but 63 respondents because they all selected Other or did not answer the question in both years.

2.1.2 Rest of the Survey

In total, the data file contains 668 variables. However, not all respondents answered every question. To fill in these missing values, I use MICE (Multivariate Imputation via Chained Equations).[2] MICE uses the observed values to predict missing values using regressions that includes all the columns as independent variables. Each variable has an imputation model depending on the type of variable. For continuous variables, linear regression is used to predict the missing values, while logistic regression is used to compute missing values for categorical variables. A major benefit of using MICE is that it creates multiple imputations for missing values, instead of only one imputation. This helps mitigate the uncertainty in missing values.

Before running MICE, I performed data cleaning. For example, all of the missing values were not read into R as NA because all the variables were read as strings. Thus, missing values were defined as ” (an empty string). I converted all those values to NA, as well as other no answers that I found in the data, such as did not say and prefer not to say. After converting these values to NA, I dropped columns that had 20% or more of missing values. This is because if a variable has too many missing values, there are not many observed values to use to predict these missing values, meaning that there is great uncertainty of whether the predictions are good representation of the missing data. In this process, I also dropped variables that are irrelevant, such as respondent ID and survey start and end time. After this process, the dataset was decreased from 668 variables to 437. All 8,000 respondents are still included in the data, only variables were reduced.

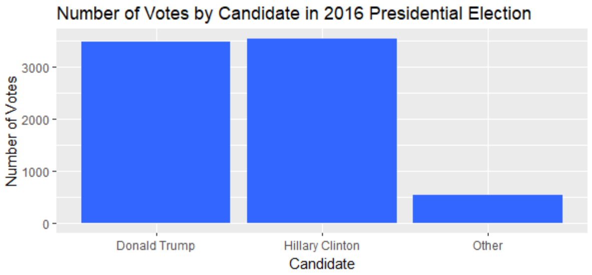
MICE was then run on this dataset of 8,000 observations and 437 variables to impute missing values. Imputing missing values was done because a dataset without missing values needs to be used in modeling to create better models. Missing values could not be ignored in the dataset because very few respondents answered all the questions. Thus, excluding incomplete rows would decrease the size of the data from 8,000 to 157 observations. Additionally, I ran FAMD (Factor Analysis of Mixed Data) to represent all 437 variables in a lower dimension. FAMD was chosen as the dimension reduction technique because the data contains both categorical and continuous variables, with most of the variables being categorical. Thus, PCA would fail because the variables are not linearly correlated.

2.2 Exploratory Data Analysis

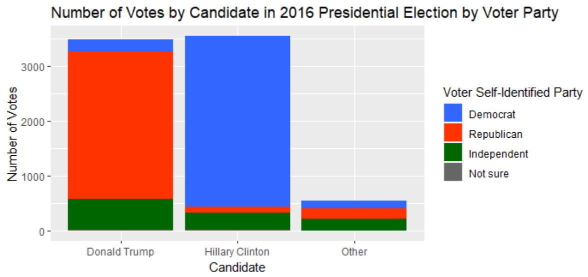
In this section, I analyze voting behavior in the 2012 and 2016 presidential elections. From the analysis, I conclude that political party seems to have the greatest effect on voting behavior. Therefore, I analyze whether voters constantly vote for the same political party across different elections.

2.2.1 Characteristics of Voters Based on Whom They Voted for in the 2016 Presidential Election

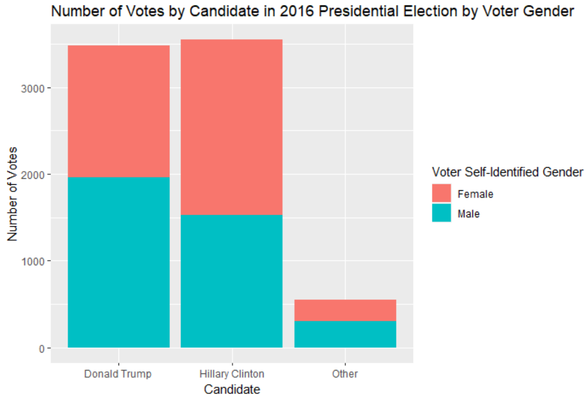
In the 2016 survey, respondents were asked whom they voted for in the 2016 General Election for president. 394 respondents did not respond, and 33 respondents did not vote for president in the election. The figure shows that slightly more respondents voted for Hillary Clinton than Donald Trump in the election, a difference of 66 respondents or less than 1% point. 47% of the respondents reported voting for Clinton, while 46% of the respondents reported voting for Trump. This is similar with the actual presidential results, where Clinton won 49% and Trump 46% of the total votes.[3]



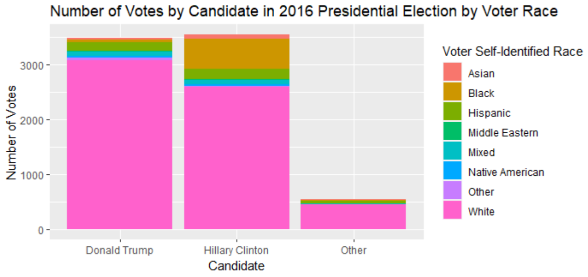
The survey also asked respondents to identify their party based on a 7 point party ID (pid7). The 7 categories are strong Democrat, not very strong Democrat, lean Democrat, independent, lean Republican, not very strong Republican, and strong Republican. There is an additional category of not sure. The figure below shows the number of votes by candidate in the 2016 presidential election by voter self-identified party. As expected, almost all the self-identified Republicans voted for the Republican nominee, Trump, and vice versa for self-identified Democrats. However, a larger percentage of self-identified Democrats voted for Clinton than the percentage of self-identified Republicans who voted for Trump, 87% and 77% respectively. This 10% point difference shows that Clinton had greater support within her party than Trump. Interestingly, more independents voted for Trump over Clinton, 211 (or 6% of Democrats) and 114 respondents (or 4% of Republicans) respectively.



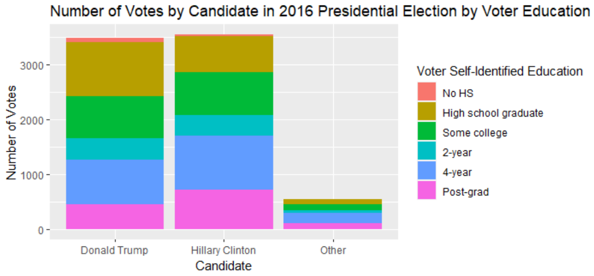
The survey also asked respondents to identify their gender, as either male or female. The figure below shows the number of votes by candidate in the 2016 presidential election by voter self-identified gender. Self-identified females voted for Clinton at a higher rate than Trump. 53% of females voted for Clinton, while only 40% voted for Trump, a difference of 13% points or 501 respondents. The opposite is true for self-identified males. 52% of males voted for Trump, while only 40% voted for Clinton, a difference of 12% points or 435 respondents. The difference in voting between males and females are similar, within 1 percentage point. However, there are more female voters than male voters (4,060 female voters and 3,940 male voters), so the difference in voters are different, giving Clinton more votes than Trump.



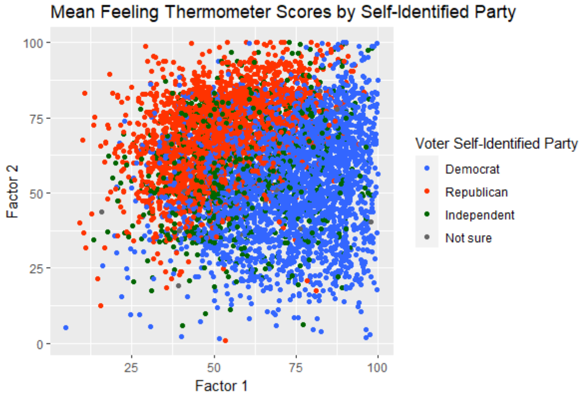
The figure below shows the number of votes by candidate in the 2016 presidential election by voter self-identified race. A higher percentage of voters who identified as Asian, Black, Hispanic, or Middle Eastern voted for Clinton than Trump. The opposite is true for voters who identified as Mixed, Native American, Other, or White. For White voters, 50% voted for Trump, while 42% voted for Clinton, a difference of 8% points or 479 respondents.

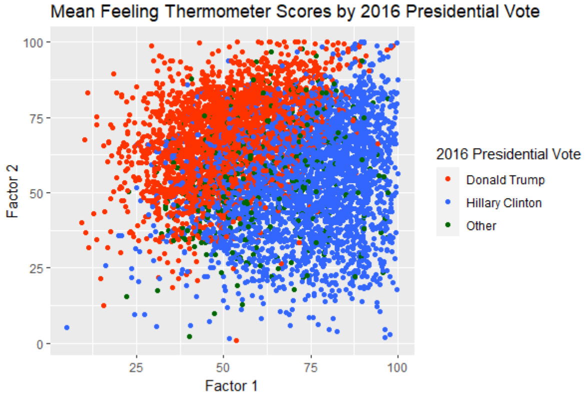


They survey also asked respondents to report their level of educational attainment on a 6 category scale. The 6 categories are did not attend high school, graduated from high school, attended college but did not graduate, graduated with a 2-year degree, graduated with a 4-year degree, and graduated with a postgraduate degree. The figure below shows the number of votes by candidate in the 2016 presidential election by voter self-identified educational attainment. More educated voters are more likely to vote for Clinton than Trump. Of college graduates, 51% voted for Clinton, compared to 41% for Trump, a difference of 10% points or 420 respondents. Alternatively, less educations voters are more likely to vote for Trump than Clinton. Of non-college graduates, 52% voted for Trump, compared to 42% for Clinton, a difference of 10% points or 354 respondents. Interestingly, the difference in percentages between the voting patterns of college graduates and non-college graduates are almost identical, both 10%. However, there are more college graduate voters than non-college graduate voters (4072 college graduate voters and 3501 non-college graduate voters), so the difference in voters are different, giving Clinton more votes than Trump.

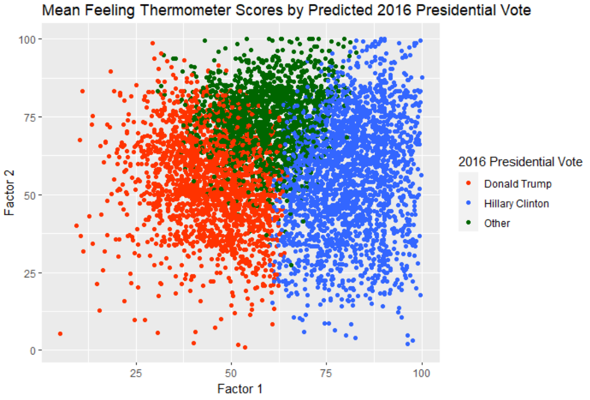


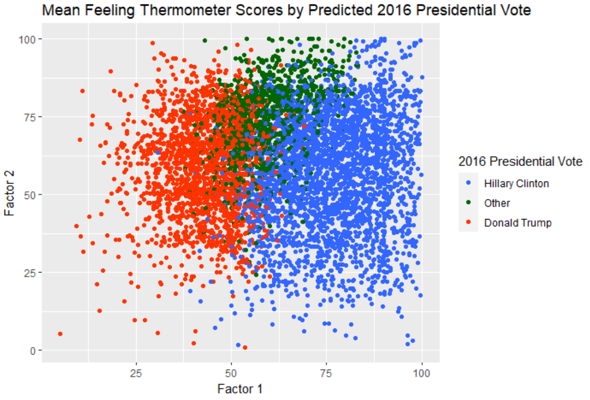
The survey also included 12 questions about how the respondents feel about certain groups of people. I run a factor analysis to see if these individual scales can be grouped in some way. The factor analysis reveals that the 13 questions can be grouped into 2 groups. The first group includes feelings about Blacks, Hispanics, Asians, Muslims, Jews, feminists, immigrants, Black Lives Matter (BLM), gays and lesbians, and labor unions. The second group includes feelings about Whites, Christians, and Wall Street bankers. Interestingly, these two groups follow closely what is to be expected, with minority groups, left-wing groups, and unions being included in group 1 and majority groups and Wall Street included in group 2. I create an average scale for each of the two groups. Because of the two groups, I expect Democrats to have high scores for group 1 and low scores for group 2 and vice versa for Republicans. Independents vary in ideology, from the Libertarian to the Green Party, so I expect Independents will be scattered throughout and not form a distinct group. The two figures below show the distribution of respondents based on these two scales by self-identified political party and whom they voted for in the 2016 presidential election. Both figures show the same trends, which are consistent with the predictions. As discussed previously, most voters vote for the candidate from their political party, so it is not surprising that the two figures are almost identical. However, there is no distinct split between any of the groups. Democrats and Republicans and Clinton and Trump voters overlap throughout. However, the density changes, with less overlapping between the two main parties and candidates occurring at the extreme ends of the spectrum than near the middle.





I also use k-means and hierarchical clustering to see how well whom the respondents voted for in the 2016 presidential election can be predicted from these 13 individual questions. The first figure shows the result from k-means, while the second figure shows the result from hierarchical clustering. The two figures show that k-means performs worse than hierarchical clustering at predicting whom the respondents voted for in the 2016 presidential election. A major reason why k-means performs poorly is because the size of the three clusters are different. 47% of the respondents voted for Clinton and 46% for Trump, with only 7% for all the other candidates. However, because of the high uncertainty in the data, both clustering techniques do not predict perfectly.

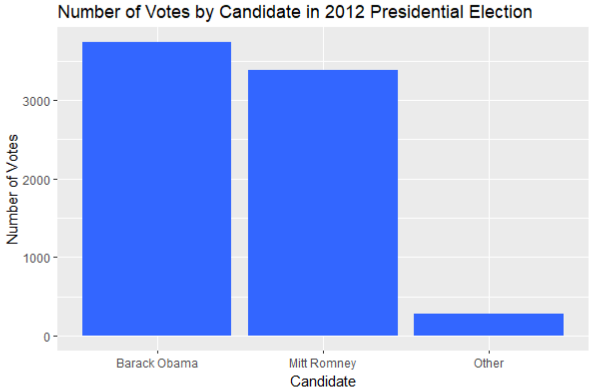




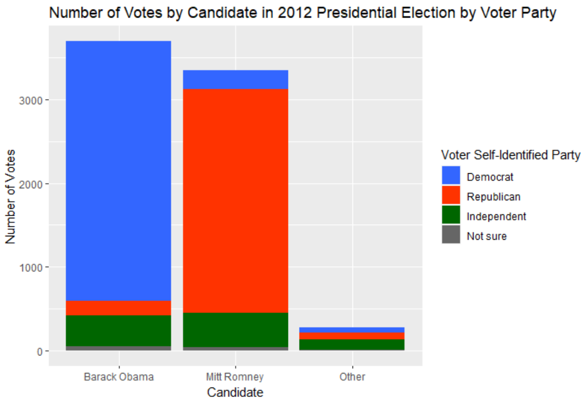
2.2.2 Characteristics of Voters Based on Whom They Voted for in the 2012 Presidential Election

Previously, I examined the characteristics of voters by whom they voted for in the 2016 presidential election. Now, I switch to the 2012 presidential election to see if the trends identified in the 2016 presidential election also exist in the 2012 presidential election.

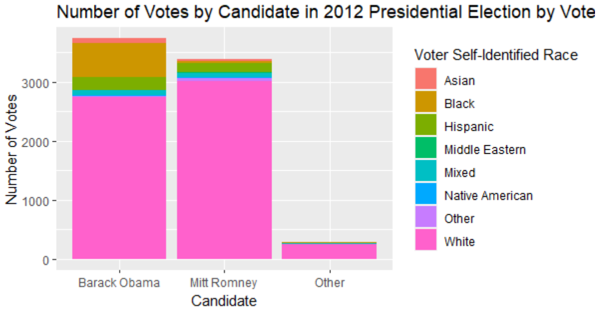
Respondents were asked whom they voted for in the 2012 presidential election. 568 respondents did not respond, 15 responded they did not vote, and 21 stated they were unsure of whom they voted for. The figure shows that more voters voted for Barack Obama than Mitt Romney in the election, a difference of 350 respondents of 5% points. 50% of the respondents reported voting for Obama, compared to 45% for Romney. This is similar with the actual presidential results, where Obama won 51% and Romney 49% of the total votes.[4] In 2012, the winner, determined by the electoral college, was Obama. He also won the popular vote, determined by a majority vote, and the simple majority, determined by receiving more than 50% of the votes. Alternatively, in 2016, Trump won the electoral college but did not receive the popular vote nor a simple majority. Clinton won the popular vote, and no candidate won a simple majority. The latter is due in part to the increase in votes to third-party candidates, such as Gary Johnson and Jill Stein, a difference of 271 respondents or 3% points.



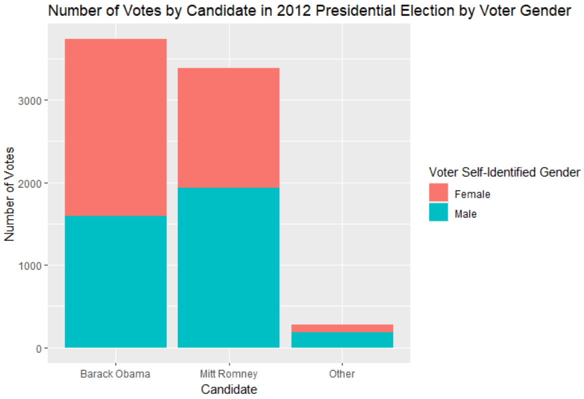
The survey also asked respondents to identify their party in 2010 based on a 7 point party ID (pid7). The figure below shows the number of votes by candidate in the 2012 presidential election by voter self-identified party. As expected, almost all the self-identified Democrats voted for the Democratic nominee, Obama, and vice versa for self-identified Republicans, 84% and 80% respectively. This difference in party support between the Democratic and Republican candidate is smaller than the difference in then 2016 presidential election. Clinton received a larger percentage of votes among Democrats than Obama, and Romney received a larger percentage of votes among Republicans than Trump. Also, slightly more Independents voted for Romney over Obama, a difference of 42 respondents or 1% point.



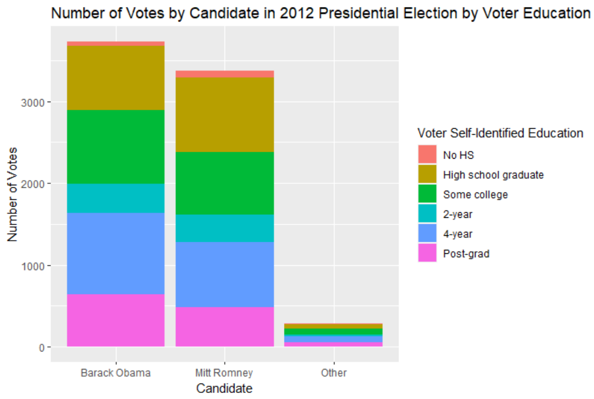
The figure below shows the number of votes by candidate in the 2012 presidential election by voter self-identified race. These patterns are similar to the 2016 election. The breakdown for which races are more likely to vote for which party's candidate is identical, except a majority of Mixed voters voted for Obama in 2012 and Trump 2016. 27% of voters who voted for Obama were racial minority (not White), compared to 11% for Romney. Interestingly, these percentages are similar to 2016, 27% for Clinton and 12% for Trump, which is surprising because Obama is the first African American nominated as the presidential nominee of a major political party. One explanation is that an increase in minority support occurred in 2008 because it was the first election with an African American as the presidential nominee of a major political party. Unfortunately, the respondents were not asked whom they voted for in the 2008 presidential election.



The figure below shows the number of votes by candidate in the 2012 presidential election by voter self-identified gender. Voters who identified as female voted for Obama at a higher rate than Romney. 58% of female respondents voted for Obama, while only 39% voted for Romney, a difference of 19% points or 693 respondents. The opposite is true for voters who identified as male. 52% of male respondents voted for Romney, while only 43% voted for Obama, a difference of 9% points or 343 respondents. Interestingly, these differences in voting by gender are consistent with the differences in voting in the 2016 presidential election. Most notably, Obama won 58% of the votes from females in 2012, while Clinton only won 53% in 2016. This is surprising because Clinton is the first female nominated as the presidential nominee of a major political party. Thus, if voters voted based on descriptive representation, an idea that people vote for the candidate who shares the same characteristics as them, a larger percentage of females would have voted for Clinton than Obama. However, this is not the case. Considering that a larger percentage of minorities did not vote for Obama in 2012 compared to Clinton in 2016 and that a larger percentage of females did not vote for Clinton in 2016 compared to Obama in 2012, these findings provide support that voters do not vote based solely on descriptive representation. Rather, political party appears to have a larger influence.



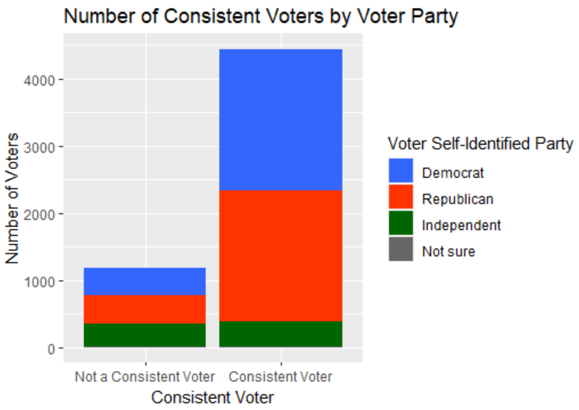
The figure below shows the number of votes by candidate in the 2012 presidential election by voter self-identified educational attainment. These patterns are similar to the 2016 election. More educated voters are more likely to vote for Obama than Romney. Of college graduates, 53% voted for Obama, compared to 43% for Romney, a difference of 10% points or 381 respondents. Alternatively, less educated voters are more likely to vote for Romney than Obama. Of non-college graduates, 48.5% voted for Romney, compared to 48.0% for Obama, a difference of .5% points or 18 respondents. This gap in non-college graduates is basically non-existent in the 2012 election. However, there was a 10% point gap in the 2016 election, meaning that Trump over performed and Clinton under performed in this demographic in the 2016 election, compared to the 2012 election.

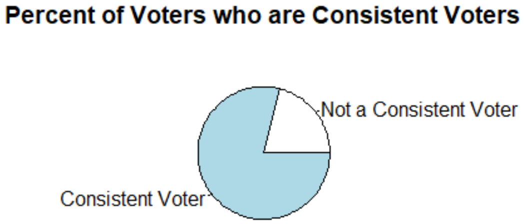


2.2.3 Do Voters Consistently Vote for the Same Party

Straight ticket voting is when voters in a general election vote for all candidates on the ballot belonging to a specific political party. This is common enough that seven states, including Indiana, include boxes on the ballot that indicate whether to select all Democratic and Republican candidates.[5] This way, the voter does not have to manually go through the entire ballot to vote for all candidates of a certain political party. This supports the claim that the candidate's political party is an important factor of whom the voters vote for. For example, Republican voters will always vote for the Republican candidate in the general election, no matter who the nominee is, and vice versa for Democratic voters. If this is true, voters should vote for the same party across different elections. To test this, I analyze the reported voting behavior for four different elected positions across the three time periods. The four positions are House of Representative for 2010 and 2012 and president for 2012 and 2016. Voting for Senator is excluded because not all states have a Senate election in every even year election.

The figure below shows the distribution of consistent voters by voter self-identified party. Consistent voters are identified as voting for the same political party across the four elections. Specifically, the voter voted for the Democratic nominee in all 4 elections, the Republican nominee in all 4 elections, or an independent candidate in all 4 elections. Most voters are consistent voters (79%), which provides support that party affiliation is the most important factor for a voter in deciding who to vote for in the general election.





The distribution of consistent voters among Democratic and Republican voters is comparable, 84% and 82% respectively. However, the distribution of consistent voters among voters who are Independent or not sure is substantially lower. Only 52% of Independent voters and 53% of unsure voters are consistent voters. Possible implications of this result is that political campaigns should target voters whom do not identify as Democratic nor Republican because they are substantially less likely to always vote for the same party in every election.

3. Algorithm and Methodology

As explained in the introduction, electoral forecasting is important for campaigns and political actors. In this section, I explain which algorithms I use to model voting behavior. Specifically, I use K-nearest neighbors, Naive Bayes classifier, penalized multinomial regression, and random forest to model 2016 presidential election voting. For all models, I ran 10-fold cross-validation and included all variables in the model. To model 2016 presidential election voting, the model will predict whether each respondent in the dataset voted for Clinton, Trump, or other. Because very few voters voted for a third-party candidate, all third-party candidates are grouped together in the other category. Additionally, predicting which third-party candidate a voter will vote for is unimportant for predicting the outcome of the election because third-party candidates have yet to even come close to winning the election. Only four third-party candidates in the entire U.S. history have received a single electoral college vote,[6] which is achieved by receiving the most votes in a state (a simple majority) or an elector not voting according to the state's popular vote (faithless elector). The most recent candidate was John Hospers in 1972, but he only received a single vote due to a faithless elector. The third-party candidate who won the most electoral votes was Wallace in 1968 at 46 votes or 9%, well below the majority needed to win. Because most voters vote for the candidate from their party, and do so consistently across all elections, I expect that this model will be similar to models predicting voting behavior for different elections and political party. Thus, the model might perform well at predicting future elections.

3.1 K-Nearest Neighbors

K-nearest neighbors (KNN) is an example of a lazy learning algorithm. KNN finds stores all the data points in the training data and finds the K number of data points in the training data that are the closest to each data point in the test data. The test data is then classified based on this group of nearest data points. KNN can use different distance functions to determine which points are the closest, such as Euclidean or Supreme distance, and use different voting formulas to classify the test data, such as majority voting or weighted distance voting.

3.2 Naive Bayes Classifier

Bayesian classifier is an approach for modeling probabilistic relationships between the attribute set and class variable. This is useful in modeling because the relationship between the attribute set and the class variable is commonly nondeterministic. Naive Bayes classifier estimates the class-conditional probability by assuming that the attributes are conditionally independent, given the class label. The classifier creates and stores the posterior probability for each class using the training data. For each data point in the test data, the probability that the data point has a certain class label is calculated for each class label based on the already generated posterior probabilities. The point is then classified based on which class label has the highest probability.

3.3 Penalized Multinomial Regression

The class variable is categorical with three options, so I use a logistic regression as as regression model instead of a different type of regression, such as a linear regression; logistic regressions are appropriate for categorical variables, while linear regressions are appropriate for modeling continuous variables. Within logistic regressions, multinomial logistic regression was chosen because the variable has more than two categories (not appropriate for binary logistic regression) and is not ordered (not appropriate for an ordinal logistic regression). Additionally, penalized multinomial regression was chosen because of the high number of independent variables in the data. A penalized regression shrinks the coefficients of the less contributive variables toward zero.[7] A multinomial logistic regression is a supervised learning algorithm that uses the training data to create a regression line that models the class variable by using all the other variables as independent variables in the regression line. Class labels are predicted based on the regression line output for each data point in the test dataset.

3.4 Random Forest

Random forest is an ensemble method that combines the predictions made by multiple decision trees, where each tree is generated based on the values of an independent set of random vectors. These random vectors are generated from a fixed probability distribution. Random forest is more accurate than a single decision tree because it combines the predictions of a large number of small decision trees. This also makes random forest more robust and have a lower risk of overfitting because the averaging of uncorrelated trees lowers the overall variance and prediction error.[8] Overfitting the training data is a problem of using decision trees as models because decision trees that are grown very deep tend to learn patterns that are specific to the training dataset.[9]

4. Experiments and Results

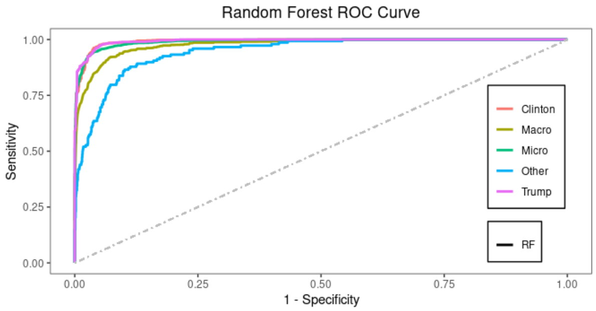
In this section, I present the results for the four different algorithms used to model 2016 presidential voting. Of the algorithms, random forest performed the best at predicting the class labels for the test dataset, receiving an accuracy of .9327 and an area under the curve (AUC) of .9921. Additionally, random forest and penalized multinomial regression were the only two algorithms that predicted cases of the voter voting for a third-party candidate. K-nearest neighbors and Naive Bayes classifier did not predict any of the test data as voting for a third-party candidate because voting for a third-party candidate is a rare occurrence; only 549 respondents or 7% voted for a third-party candidate in the 2016 presidential election. Not predicting voting for a third-party candidate is not grounds for immediately disregarding the model because correctly predicting voting for third-party candidates is not highly important. Third-party candidates rarely receive an electoral college vote, nevertheless come close enough to winning the election. Correctly predicting voting for the Democratic and Republican Party nominee is more important than correctly predicting voting for third-party candidates. Thus, when evaluating algorithms based on sensitivity for voting for Clinton and Trump, random forest is still the best algorithm. Random forest correctly predicted 97% and 98% of respondents who voted for Clinton and Trump, respectively. However, penalized multinomial regression is the worst algorithm by only predicting 94% of respondents who voted for Clinton and Trump correctly.

Presented below are the receive operator characteristic (ROC) curve and confusion matrix for each algorithm, in order of the algorithm's accuracy and AUC. The ROC curve shows the trade-off between sensitivity, or the true positive rate (TPR), and specificity, or 1-the false positive rate (1-FPR).

Algorithms that produce curves with high sensitivity and low specificity, as well as a high area under the curve (maximum value of 1), have a better performance. On these figures, I included a line for sensitivity=specificity, which represents the result if the class was randomly predicted and the baseline to see if the algorithm is successful. All of these models predict significantly better than random prediction.

4.1 Random Forest

It is no surprise that random forest is a good algorithm for modeling 2016 presidential voting behavior because random forest is good at handling heterogeneous feature types and high dimensionality, as well as good for classification problems.[9] Of the 4 algorithms, random forest had the highest accuracy and AUC, .9327 and .9921 respectively.



Random Forest Confusion Matrix

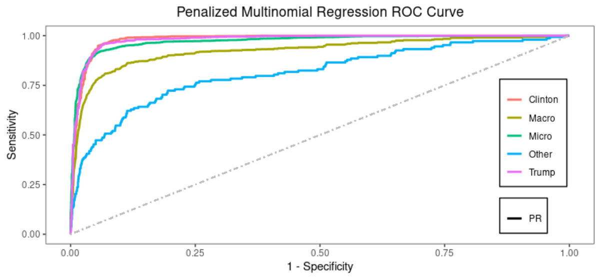
Prediction	Clinton	Other	Trump
Clinton	895	39	13
Other	10	61	8
Trump	15	48	886

Random Forest Statistics

	Clinton	Other	Trump
Sensitivity	.9728	.4122	.9768
Specificity	.9507	.9902	.9410
Positive Predictive Value	.9451	.7722	.9336
Negative Predictive Value	.9757	.95411	.9795
Prevalence	.4658	.0749	.4592
Detection Rate	.4532	.0309	.4486
Detection Prevalence	.4795	.0400	.4805
Balanced Accuracy	.9618	.7012	.9589

4.2 Penalized Multinomial Regression

Penalized multinomial regression had the second highest accuracy and AUC, .9013 and .9806 respectively, out of the four algorithms.



Penalized Multinomial Regression Confusion Matrix

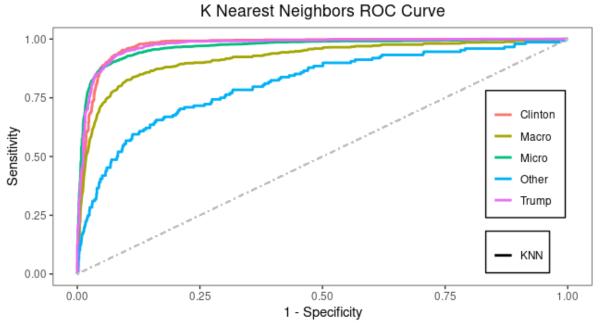
Prediction	Clinton	Other	Trump
Clinton	866	43	17
Other	37	60	36
Trump	17	45	854

Penalized Multinomial Regression Statistics

	Clinton	Other	Trump
Sensitivity	.9413	.4054	.9416
Specificity	.9431	.9600	.9419
Positive Predictive Value	.9352	.4511	.9323
Negative Predictive Value	.9485	.9522	.9500
Prevalence	.4658	.0749	.4592
Detection Rate	.4385	.0304	.4324
Detection Prevalence	.4689	.0673	.4638
Balanced Accuracy	.9422	.6827	.9418

4.3 K-Nearest Neighbors

K-nearest neighbors had the third highest accuracy and AUC, .8851 and .9722 respectively, out of the four algorithms.



K-Nearest Neighbors Confusion Matrix

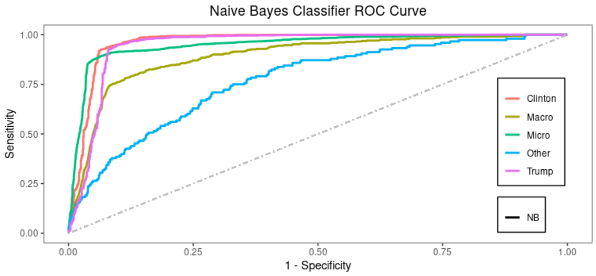
Prediction	Clinton	Other	Trump
Clinton	882	68	41
Other	0	0	0
Trump	38	80	866

K-Nearest Neighbors Statistics

	Clinton	Other	Trump
Sensitivity	.9587	.0000	.9548
Specificity	.8967	1.0000	.8895
Positive Predictive Value	.8900	NA	.8801
Negative Predictive Value	.9614	.9251	.9586
Prevalence	.4658	.0749	.4592
Detection Rate	.4466	.0000	.4385
Detection Prevalence	.5018	.0000	.4982
Balanced Accuracy	.9277	.5000	.9222

4.4 Naive Bayes Classifier

Naive Bayes classifier had the worst accuracy and AUC, .8825 and .9617 respectively, out of the four algorithms.



Naive Bayes Classifier Confusion Matrix

Prediction	Clinton	Other	Trump
Clinton	874	67	38
Other	0	0	0
Trump	46	81	869

Naive Bayes Classifier Statistics

	Clinton	Other	Trump
Sensitivity	.9500	.0000	.9581
Specificity	.9005	1.0000	.8811
Positive Predictive Value	.8927	NA	.8725
Negative Predictive Value	.9538	.9251	.9612
Prevalence	.4658	.0749	.4592
Detection Rate	.4425	.0000	.4400
Detection Prevalence	.4957	.0000	.5043
Balanced Accuracy	.9252	.5000	.9196

5. Summary and Conclusions

This report analyzes U.S. voting behavior, including models to predict 2016 presidential election voting behavior given certain voter information. Data was collected from the VOTER (Views of the Electorate Research) Survey, conducted by YouGov. From my analysis on voting behavior in the 2012 and 2016 presidential elections, political party appears to have the greatest effect on voting behavior. A large majority of Democrats and Republicans voted for their party's presidential nominee in the both the 2012 and 2016 presidential election. 87% and 84% of Democratic respondents voted for Clinton and Obama, respectively, in the presidential election. Additionally, 77% and 80% of Republican respondents voted for Trump and Romney, respectively, in the presidential election. To further explore these findings, I analyzed whether voters consistently vote for the same party. Across four different positions, U.S. House of Representative for 2010 and 2012 and U.S. president for 2012 and 2016, 79% of respondents voted for their party's nominee in all four positions. 84% of Democratic respondents and 82% of Republican respondents voted for their party's nominee in all four positions.

To model 2016 presidential voting, I ran four different algorithms—K-nearest neighbors, Naive Bayes classifier, penalized multinomial regression, and random forest—using 10-fold cross-validation. Of the four algorithms, random forest performed the best at predicting 2016 presidential voting behavior, receiving an accuracy of .9327 and area under the curve of .9921 on the test dataset. Additionally, random forest correctly predicted 97% and 98% of respondents who voted for Clinton and Trump, respectively.

Acknowledgments

I am grateful to Dr. Hasan Kurban for all his assistance.

References

[1] Democracy Fund Voter Study Group. 2016 VOTER Survey Data Set, August 2017.

[2] Harshitha Mekala. Dealing With Missing Data using R, June 2018.

[3] CNN. Presidential Election Results 2016.

[4] New York Times. Presidential Map - Election 2012.

[5] Straight Ticket Voting, June 2021.

[6] Christopher Klein. Here's How Third-Party Candidates Have Changed Elections, November 2019.

[7] Alboukadel Kassambara. Penalized Logistic Regression Essentials in R: Ridge, Lasso and Elastic Net, November 2018.

[8] IBM. Random Forest.

[9] Thomas Wood. What is a Random Forest?