Predicting voting behavior for a generic ballot

Lauren Lawless | IST 707: Data Analytics | December 9, 2019

# Introduction

Much time and money are spent during elections in order to gauge public interest in both the candidates and in the election itself as well as to predict the outcome of the election. Many organizations conduct several polls starting more than a year before the election date and not ending until that date. The value of polling in discovering how the electorate will vote, if at all, is an irreplaceable tool used by campaigns in determining the receptivity of the public to their candidate, thus the more accurate the predictive power, the better.

Conducting such polls, however, is costly in both time and resources. Response rates to telephone surveys have fallen from 36% in 1997 to 6% in 2018, but fewer responses does not necessarily indicate lower accuracy in the results. (Kennedy & Hartig, 2019) The lower response rates mean that other measures must be taken to account for higher costs, such as reduced sample sizes, extended survey periods, or shifted resources, resulting in reduced quality elsewhere. Because of these costs and deferred resources, we would like to ensure that the results allow us to be as accurate as possible.

While we cannot eliminate polling from the political campaign process, we may increase its efficiency by asking the questions that provide the most information gain while also trying to be less privacy invasive. We may then use data mining techniques to build a model that will classify individuals by likely voting behavior based on accessible information (e.g., party registration, age, etc.) so that we may gauge voter interest and intent without the need for greater expenditure on response quantities.

Thus, we must identify which voter attributes are the most predictively valuable in a classification model without being so personal in nature that we would not be able to get the information other than by directly polling the individual.

In 2018, the Associated Press replaced their usual in-person exit poll with VoteCast, a phone and web-based survey of 138,929 registered voters across all 50 states. (NORC at the University of Chicago, 2019) The poll ran from October 29 through the time election polling locations closed on November 6, 2018. The questionnaire was designed to determine whether the individual voted; if so, for whom; and general opinions on national topics as well as demographic information. In the interest of preserving the integrity of the research, AP made the data available to academics and the public, allowing it to be peer reviewed by those who wish to do so.

Since 2018 was a midterm election (2 years after the prior presidential election and 2 years before the next) and there was no nation-wide race on which to focus, the responses to this survey will be used to predict from which party (Democrat, Republican, or Other) a voter is most likely to choose a preferred candidate on a generic ballot.

# Methodology

## Data Preparation

The data was already segmented into discrete factors for each question, but additional data preparation was done to “cleanse” the discrete responses and make them more concise and readable in short form.

A secondary data set was produced from the discrete variables in order to have a purely numeric set for use with distance-based classification models. This conversion led to a blend of binomial and ordinal numeric response and characteristic sets for each individual respondent.

In order to reduce total survey time for the majority of respondents, only one group was given all survey questions, with other survey forms randomly assigned as respondents took the surveys. I pared down the variable set to those I found most relevant to the question at hand, then filtered out respondents who had not been administered each of the questions of interest (represented by blanks in the data set).

Almost no question in the survey was required, and we see in the data an option that indicates that the person was asked the question but did not respond to it. These non-responses were left in both data sets. In the numeric data set, these responses were assigned the value of the “outgroup” for binary sets of numeric variables and the expected value (assuming equal probability of each response) of ordinal numeric variables. In the qualitative data set, the non-responses were left as-is.

Each qualitative and quantitative data set was divided into ⅔-⅓ training and testing sets on the same random index.

## Modeling

### Method: Decision Tree

Using all selected variables in the data set, the first model (Model 1) achieves an accuracy of 0.8859 with a 13-node tree. Reducing the data set from 66 variables to 24, the second model (Model 2) achieves an accuracy of 0.8590 when applied to the test data using its 3-node tree.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Model 1 | | | | |  |  | Model 2 | | | | |
|  |  | Reference | | |  |  |  |  | Reference | | |  |
|  |  | Dem. | Other | Rep. |  |  |  |  | Dem. | Other | Rep. |  |
| Prediction | Dem. | 1183 | 69 | 60 | **1312** |  | Prediction | Dem. | 1182 | 103 | 103 | **1388** |
| Other | 18 | 28 | 9 | **55** |  | Other | 0 | 0 | 0 | **0** |
| Rep. | 62 | 50 | 869 | **981** |  | Rep. | 81 | 44 | 835 | **960** |
|  |  | **1263** | **147** | **938** | **2348** |  |  |  | **1263** | **147** | **938** | **2348** |

The primary drivers of the second model are individuals’ responses to the attributes PARTY (party registration) and TRACK (direction of the country). While party registration information is typically known in the parties’ voter databases, an individual’s opinion on the direction of the country often depends on who is in power.

The diagram of Model 2[[1]](#footnote-1) shows that, as expected, registered Republicans vote Republican, registered Democrats vote Democrat, and individuals registered with neither party are likely to vote for the party currently in control—in the case of 2018, Republicans—if they think the country is going in the “right” direction and will likely vote for the party not in control—in 2018, Democrats—if they do not.[[2]](#footnote-2) This model, however, does not account for third-party options for voting, which, while they rarely win election to federal offices,[[3]](#footnote-3) third-party candidates often do obtain enough of the vote share that a strictly two-party model does not adequately represent voting behavior in some cases.[[4]](#footnote-4)

Additionally, our data set includes the ideology attribute, which, while related to party registration can be more nuanced for many people. It is also very difficult to gauge without directly inquiring about it in an interview.

Thus, in our final decision tree model (3), we remove the TRACK and IDEO attributes from the data set to get a model with 0.8390 accuracy on the test data using an 8-node tree. This tree devotes no attention to predicting third-party voting, so we are no better off than with the simpler 3-node model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Model 3 | | | | |
|  |  | Reference | | |  |
|  |  | Dem. | Other | Rep. |  |
| Prediction | Dem. | 1168 | 104 | 136 | **1408** |
| Other | 0 | 0 | 0 | **0** |
| Rep. | 95 | 43 | 802 | **940** |
|  |  | **1263** | **147** | **938** | **2348** |

### Method: Naïve Bayes

Applying the Naïve Bayes technique to the entire training set, including all attributes, produces a model (4) with 0.8756 accuracy, but does not predict any third-party voters. Using the reduced-attribute data set decreases model accuracy to 0.7649 and still does not include any third-party predictions (Model 5). This method may be more effective in a two-party prediction model, but we would prefer not to ignore the third-party option that is present in many elections.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Model 4 | | | | |  |  | Model 5 | | | | |
|  |  | Reference | | |  |  |  |  | Reference | | |  |
|  |  | Dem. | Other | Rep. |  |  |  |  | Dem. | Other | Rep. |  |
| Prediction | Dem. | 1198 | 91 | 80 | **1369** |  | Prediction | Dem. | 894 | 50 | 36 | **980** |
| Other | 0 | 0 | 0 | **0** |  | Other | 0 | 0 | 0 | **0** |
| Rep. | 65 | 56 | 858 | **979** |  | Rep. | 369 | 97 | 902 | **1368** |
|  |  | **1263** | **147** | **938** | **2348** |  |  |  | **1263** | **147** | **938** | **2348** |

### Method: kNN

The kNN technique produces similarly accurate models (0.8365 using default settings for Model 6 and 0.8343 with tuning for Model 7) but still ignores third-party voters. This indicates that third-party voters are often more dissimilar from one another in terms of the selected attributes compared the statistically identifiable similarities among Democratic and Republican voting groups.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Model 6 | | | | |  |  | Model 7 | | | | |
|  |  | Reference | | |  |  |  |  | Reference | | |  |
|  |  | Dem. | Other | Rep. |  |  |  |  | Dem. | Other | Rep. |  |
| Prediction | Dem. | 1132 | 94 | 105 | **1331** |  | Prediction | Dem. | 1132 | 95 | 112 | **1339** |
| Other | 1 | 0 | 1 | **2** |  | Other | 0 | 0 | 0 | **0** |
| Rep. | 130 | 53 | 832 | **1015** |  | Rep. | 131 | 52 | 826 | **1009** |
|  |  | **1263** | **147** | **938** | **2348** |  |  |  | **1263** | **147** | **938** | **2348** |

### Method: SVM

Model 8, produced by the svmLinear method, has 0.8445 accuracy but does not predict any third-party voters. The svmRadial method produces a model (9) with 0.6001 accuracy, so we can assume that radial projection is not optimal for this data set. Finally, using the svmPoly method for Model 10, we find the optimal degree value to be 1, which is consistent with a well-performing linear model. We also find that a 4th-degree model (11) has slightly decreased accuracy (0.7828) compared to the linear model (8), but it makes some attempt at identifying third-party voters. This is a poorer performance than most of the other models but is a notable exception that includes the third-party cases.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Model 8 | | | | | | | | | |  | |  | | Model 9 | | | | | | | | | |
|  | |  | | Reference | | | | | |  | |  | |  | |  | | Reference | | | | | |  | |
|  | |  | | Dem. | | Other | | Rep. | |  | |  | |  | |  | | Dem. | | Other | | Rep. | |  | |
| Prediction | | Dem. | | 1149 | | 83 | | 104 | | **1336** | |  | | Prediction | | Dem. | | 1248 | | 140 | | 777 | | **2165** | |
| Other | | 0 | | 0 | | 0 | | **0** | |  | | Other | | 0 | | 0 | | 0 | | **0** | |
| Rep. | | 114 | | 64 | | 834 | | **1012** | |  | | Rep. | | 15 | | 7 | | 161 | | **183** | |
|  | |  | | **1263** | | **147** | | **938** | | **2348** | |  | |  | |  | | **1263** | | **147** | | **938** | | **2348** | |
|  |  | | | | | | | | | |  | |  | |  | | | | | | | | | |
|  | Model 10 | | | | | | | | | |  | |  | | Model 11 | | | | | | | | | |
|  |  | | Reference | | | | | |  | |  | |  | |  | | Reference | | | | | |  | |
|  |  | | Dem. | | Other | | Rep. | |  | |  | |  | |  | | Dem. | | Other | | Rep. | |  | |
| Prediction | Dem. | | 1149 | | 84 | | 104 | | **1337** | |  | | Prediction | | Dem. | | 1071 | | 74 | | 169 | | **1314** | |
| Other | | 0 | | 0 | | 0 | | **0** | |  | | Other | | 39 | | 16 | | 18 | | **73** | |
| Rep. | | 114 | | 63 | | 834 | | **1011** | |  | | Rep. | | 153 | | 57 | | 751 | | **961** | |
|  |  | | **1263** | | **147** | | **938** | | **2348** | |  | |  | |  | | **1263** | | **147** | | **938** | | **2348** | |

### Method: Random Forest

Since the decision trees were the most effective in the three-party model, it is logical to try the random forest method to see if a more robust model is possible. The accuracy (0.8365) of this model (12) was not notably different from the other models and is actually slightly lower than the accuracy of the Model 3 decision tree.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Model 12 | | | | |
|  |  | Reference | | |  |
|  |  | Dem. | Other | Rep. |  |
| Prediction | Dem. | 1141 | 83 | 122 | **1346** |
| Other | 14 | 14 | 7 | **35** |
| Rep. | 108 | 50 | 809 | **967** |
|  |  | **1263** | **147** | **938** | **2348** |

### Method: Ensemble

It is understood in data science that ensemble learning is oftentimes more effective than any one modeling method, based on the premise that more than half of the models must be incorrect in their predictions in order for the overall prediction to be made incorrectly. While, conceptually, this is true, when put into practice using this data set, we see no real increase in the accuracy of the ensemble model compared to the other models (0.8454 when excluding Models 5, 9, and 11 for Model 13 and 0.8203 when excluding those models as well as the full information models, 1 and 4, for Model 14).

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Model 13 | | | | |  |  | Model 14 | | | | |
|  |  | Reference | | |  |  |  |  | Reference | | |  |
|  |  | Dem. | Other | Rep. |  |  |  |  | Dem. | Other | Rep. |  |
| Prediction | Dem. | 1134 | 68 | 61 | **1263** |  | Prediction | Dem. | 1101 | 69 | 69 | **1239** |
| Other | 73 | 46 | 72 | **191** |  | Other | 92 | 43 | 87 | **222** |
| Rep. | 56 | 33 | 805 | **894** |  | Rep. | 70 | 35 | 782 | **887** |
|  |  | **1263** | **147** | **938** | **2348** |  |  |  | **1263** | **147** | **938** | **2348** |

# Discussion

## Summary of Results

Table 1, below, shows the accuracy, precision, and recall values for each of the models. The mean of the accuracies is 0.8222 and the standard deviation is 0.0746. We see rather consistent values across most of the models with the exception of Models 5, 9, and 11, which all have accuracies lower than 0.80 and whose exclusion benefitted the accuracy of the ensemble model, despite Model 9 being the only true outlier.

As previously mentioned, the ensemble model did not have a notably different accuracy value, which means that the individual models are fairly consistent in their predictions: most of the correct predictions are correct across most models, and most of the incorrect predictions are incorrect across most models.

That said, we do see relatively higher across-the-board precision and recall values for Model 13. In fact, the only model that outperforms the ensemble model across all three categories is the Model 1 decision tree. If we examine the statistics at a two-party level (Democrats and Republicans only), the only additional model that outperforms Model 13 is Model 4 (Naïve Bayes).

Table : Model Statistics

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Accuracy** | **Democrat** | | | **Republican** | | | **Other** | | |
| **Precision** | **Recall** | **Precision** | | **Recall** | **Precision** | | **Recall** |
| **Decision Tree** | **Model 1** | All | 0.8859 | 0.9017 | 0.9367 | 0.8858 | | 0.9264 | 0.5091 | | 0.1905 |
| **Model 2** | Selected[[5]](#footnote-5) | 0.8590 | 0.8516 | 0.9359 | 0.8698 | | 0.8902 | #DIV/0! | | 0.0000 |
| **Model 3** | Selected | 0.8390 | 0.8295 | 0.9248 | 0.8532 | | 0.8550 | #DIV/0! | | 0.0000 |
| **Naïve Bayes** | **Model 4** | All | 0.8756 | 0.8751 | 0.9485 | 0.8764 | | 0.9147 | #DIV/0! | | 0.0000 |
| **Model 5** | Selected | 0.7649 | 0.9122 | 0.7078 | 0.6594 | | 0.9616 | #DIV/0! | | 0.0000 |
| **kNN** | **Model 6** | Selected | 0.8365 | 0.8505 | 0.8963 | 0.8197 | | 0.8870 | 0.0000 | | 0.0000 |
| **Model 7** | Selected | 0.8339 | 0.8454 | 0.8963 | 0.8186 | | 0.8806 | #DIV/0! | | 0.0000 |
| **SVM** | **Model 8** | Selected | 0.8445 | 0.8600 | 0.9097 | 0.8241 | | 0.8891 | #DIV/0! | | 0.0000 |
| **Model 9** | Selected | 0.6001 | 0.5764 | 0.9881 | 0.8798 | | 0.1716 | #DIV/0! | | 0.0000 |
| **Model 10** | Selected | 0.8445 | 0.8594 | 0.9097 | 0.8249 | | 0.8891 | #DIV/0! | | 0.0000 |
| **Model 11** | Selected | 0.7828 | 0.8151 | 0.8480 | 0.7815 | | 0.8006 | 0.2192 | | 0.1088 |
| **Random Forest** | **Model 12** | Selected | 0.8365 | 0.8477 | 0.9034 | 0.8366 | | 0.8625 | 0.4000 | | 0.0952 |
| **Ensemble** | **Model 13** | - [[6]](#footnote-6) | 0.8454 | 0.8979 | 0.8979 | 0.9004 | | 0.8582 | 0.2408 | | 0.3129 |
| **Model 14** | - [[7]](#footnote-7) | 0.8203 | 0.8886 | 0.8717 | 0.8816 | | 0.8337 | 0.1937 | | 0.2925 |

## Information Gain

In Appendix C, we see the information gain ratios for the top 20 most helpful variables from each of the four data sets: all qualitative, selected qualitative, all quantitative, and selected quantitative. The two full data set lists almost comprehensively contain attributes related to an individual’s personally held political beliefs and opinions about politicians themselves, which will not be very helpful in a model where we deprioritize direct polling. Again, in the selected data sets, we attempted to choose variables that would be attainable through other more affordable means, such as party voter files, public voter registration records, or even commercially available data sets. There is significant overlap between the two selected sets, but we can see that the quantitative set, by nature of how it discretizes categorical variables, allows us to see the specific cases of the attribute that are most helpful, for example, RACETH5 is the second most helpful attribute in the qualitative set, while in the quantitative set, it is actually the RACETH5\_AABlack attribute that is second, with RACETH5\_White being fourth and RACETH5\_LatHisp being nineteenth.

# Conclusion

When weighing all of the presented models against efficiency, accuracy, and applicability to non-polled information, we would likely see the best results when using Model 14. While this model is time-expensive to run, it provides the best overall prediction statistics among the partial-information models, especially for identifying third-party voters.

There are some cases, however, when we would like to prioritize time or to consider a two-party scenario (Democrats and Republicans only). These preferences would likely warrant utilization of other models. The former case does not have an equally beneficial alternative, as the only other partial-information model that predicts any third-party voters is Model 11, which heavily favors the selection of Democrats to a fault. The latter case may be adequately satisfied by Model 8, the linear SVM, which does not predict third-party voting but is among the most accurate for predicting Democratic or Republican votes.

## Limitations

This approach may be vulnerable to oversimplification of preferences, as the data provides no way to determine the likability of each candidate except by the number of votes cast for them, which will be blended into a party vote at large. We know, however, that specific candidates have such personalities that people will deviate from their normal voting behavior in order to specifically vote for or against them. Such was the case in the 2019 Kentucky gubernatorial race, where the other five state executive elections were won by Republicans candidates, but a Democrat defeated the incumbent Republican governor likely due to his high disapproval ratings across the state, especially among teachers and other public servants. (The New York Times, 2019)

Additionally, it is difficult to know which attributes are truly accessible to campaign data teams and which are not. I have done my best to eliminate attributes based on personally held beliefs that are not easily inferred from tangible attributes, such as charitable giving (ISSUES, RELIGION), but some of the more helpful attributes may be more within reach than we might think, for example, past voting behavior: while the United States has a secret ballot, it is documented that in their effort to ramp up data usage between 2008 and 2012, the Obama campaign was confident that they “knew the name of every one of the 69,456,897 Americans whose votes had put him in the White House” by way of combining precinct-level Democratic vote totals with prior internal analysis on who was most likely to have supported him. (Issenberg, 2012)

# Appendix A: Data Dictionary

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable (Alternative Label)** | **Description** | **Response options (numeric value)** | |
| **RACE5\_VOTE** (RACE5HouseGen) | Race 5 candidate vote - Generic House Ballot for all cases | Other (0) Republican (1) Democrat (-1) | |
| **MODE** (MODE\_Web) | Survey mode. | Web (1) Phone (0) | |
| **State** | State - Character variable - State abbreviation | Alabama (2) Alaska (1) Arizona (4) Arkansas (3) California (5) Colorado (6) Connecticut (7) Delaware (8) Florida (9) Georgia (10) Hawaii (11) Idaho (13) Illinois (14) Indiana (15) Iowa (12) Kansas (16) Kentucky (17) Louisiana (18) Maine (21) Maryland (20) Massachusetts (19) Michigan (22) Minnesota (23) Mississippi (25) Missouri (24) | Montana (26) Nebraska (29) Nevada (33) New Hampshire (30) New Jersey (31) New Mexico (32) New York (34) North Carolina (27) North Dakota (28) Ohio (35) Oklahoma (36) Oregon (37) Pennsylvania (38) Rhode Island (39) South Carolina (40) South Dakota (41) Tennessee (42) Texas (43) Utah (44) Vermont (46) Virginia (45) Washington (47) West Virginia (49) Wisconsin (48) Wyoming (50) |
| **LV** | Please indicate how likely it is that you will vote in this election. | 0-Certainly will not vote 1 2 3 4 5 6 7 8 9 10-Certain will vote Already voted (12) | |
| **LIKELYVOTER** | Likely Voter Model | Nonvoter (0) Likely voter (1) | |
| **TRACK** | Direction of the country | Right direction (1) Wrong direction (-1) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **ISSUES** (ISSUES\_Jobs ISSUES\_Healthcare ISSUES\_Immigration ISSUES\_Taxes ISSUES\_Abortion ISSUES\_Guns ISSUES\_Environment ISSUES\_ForeignPolicy ISSUES\_Terrorism) | Most important issue facing the country | Abortion Foreign policy Gun policy Health care Immigration Taxes Terrorism The economy and jobs The environment DON’T KNOW/SKIPPED/REFUSED (VOL) | |
| **FAVTRUMP** | Opinion of Donald Trump | Very favorable (2) Somewhat favorable (1) Somewhat unfavorable (-1) Very unfavorable (-2) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **FAVREP** | Opinion of the Republican Party | Very favorable (2) Somewhat favorable (1) Somewhat unfavorable (-1) Very unfavorable (-2) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **FAVDEM** | Opinion of the Democratic Party | Very favorable (2) Somewhat favorable (1) Somewhat unfavorable (-1) Very unfavorable (-2) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **SUPREMECOURT** | Importance of debate over Kavanaugh's confirmation to vote | Somewhat important (3) Very important (2) Not very important (1) Not at all important (0) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **GOVTDO** | Views of government's role | Government should do more to solve problems (1) Government doing too many things better left to business and individuals (-1) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **NEC** | Condition of national economy | Excellent (3) Good (2) Not so good (1) Poor (0) DON’T KNOW/SKIPPED/REFUSED (1.5) | |
| **GETAHEAD** | Family's financial situation | Getting ahead (1) Holding steady (0) Falling behind (-1) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **TRADENATIONALECON** | Impact of Trump admin's trade policies on national economy | Help (1) No difference (0) Hurt (-1) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **TRADELOCALECON** | Impact of Trump admin's trade policies on local economy | Help (1) No difference (0) Hurt (-1) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **ECONFAIRWEALTHY** | How the economic system treats the wealthy | Favors too much (1) Treats about right (0) Does not favor enough (-1) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **ECONFAIRMIDDLE** | How the economic system treats the middle class | Favors too much (1) Treats about right (0) Does not favor enough (-1) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **ECONFAIRPOOR** | How the economic system treats the poor | Favors too much (1) Treats about right (0) Does not favor enough (-1) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **TAXCUTS** | Opinion of 2017 tax law | Approve strongly (2) Approve somewhat (1) Disapprove somewhat (-1) Disapprove strongly (-2) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **HEALTHLAW** | Views of the Affordable Care Act | Expand the law (1) Leave the law as is (0) Repeal parts of the law (-1) Repeal the law entirely (-2) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **HEALTHGOV** | Government responsibility to provide healthcare | Government should be responsible (1) Government should not be responsible (-1) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **IMMDEPORT** | Immigration policy - immigrants living in the U.S. illegally | Offered a chance to apply for legal status (1) Deported to the country they came from (-1) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **IMMBETTER** | Views of immigrants in the U.S. | Do more to help the country (1) Do more to hurt the country (-1) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **IMMWALL** | U.S.-Mexico border wall | Strongly favor (2) Somewhat favor (1) Somewhat oppose (-1) Strongly oppose (-2) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **RACEREL** | Advantages of blacks and whites in US society | Whites have more advantages than blacks (1) Neither has much advantage over the other (0) Blacks have more advantages than whites (-1) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **CLIMATE** | Concern over effects of climate change | Very concerned (3) Somewhat concerned (2) Not too concerned (1) Not at all concerned (0) DON’T KNOW/SKIPPED/REFUSED (1.5) | |
| **GUNPOLICY** | Views of gun laws | Should be more strict (1) Should be kept as they are (0) Should be less strict (-1) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **ABORTION** | Views on abortion | Legal in all cases (2) Legal in most cases (1) Illegal in most cases (-1) Illegal in all cases (-2) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **METOOBELIEVE** | Concern about women not being believed when the make allegations of sexual misconduct | Very concerned (3) Somewhat concerned (2) Not too concerned (1) Not at all concerned (0) DON’T KNOW/SKIPPED/REFUSED (1.5) | |
| **METOODEFEND** | Concern about men not being given the opportunity to defend themselves against allegations of sexual misconduct | Very concerned (3) Somewhat concerned (2) Not too concerned (1) Not at all concerned (0) DON’T KNOW/SKIPPED/REFUSED (1.5) | |
| **PCSPEECH** | Too much pressure to be politically correct these days | Pressure to be politically correct has gone too far these days (1) There's not too much pressure these days to be politically correct (0) DON'T KNOW/SKIPPED/REFUSED (0) | |
| **MARIJUANA** | Should use of marijuana be legal nationwide | Yes, legal (1) No (0) DON'T KNOW/SKIPPED/REFUSED (0) | |
| **OPIOID** | Concern about the use of opioids in your community | Very concerned (3) Somewhat concerned (2) Not too concerned (1) Not at all concerned (0) DON’T KNOW/SKIPPED/REFUSED (1.5) | |
| **SAFETERROR** | Trump administration impact on U.S. safety - terrorism | Made the U.S. safer (1) Hasn't made much difference (0) Made the U.S. less safe (-1) DON'T KNOW/SKIPPED/REFUSED (0) | |
| **SAFECRIME** | Trump administration impact on U.S. safety - crime | Made the U.S. safer (1) Hasn't made much difference (0) Made the U.S. less safe (-1) DON'T KNOW/SKIPPED/REFUSED (0) | |
| **SAFECYBER** | Trump administration impact on U.S. safety - cyberattacks | Made the U.S. safer (1) Hasn't made much difference (0) Made the U.S. less safe (-1) DON'T KNOW/SKIPPED/REFUSED (0) | |
| **RUSSIA** | Do you think the Trump campaign coordinated with the Russian government during the 2016 presidential election? | Yes (1) No (0) DON'T KNOW/SKIPPED/REFUSED (0) | |
| **INTERFERENCE** | Concerns about foreign government interference impacting the outcome of the 2018 midterms | Very concerned (3) Somewhat concerned (2) Not too concerned (1) Not at all concerned (0) DON’T KNOW/SKIPPED/REFUSED (1.5) | |
| **Q2020VOTE** | 2020 vote choice | Donald Trump, the Republican (1) The Democratic candidate (-1) It depends (0) Would not vote (0) DON'T KNOW/SKIPPED/REFUSED (0) | |
| **REPINTENTIONS** | Intentions of Republicans | They mostly try to do what's best for the country, even if they don't always get it right (1) They mostly try to do what's best for their party, even when it's bad for the country (-1) DON'T KNOW/SKIPPED/REFUSED (0) | |
| **DEMINTENTIONS** | Intentions of Democrats | They mostly try to do what's best for the country, even if they don't always get it right (1) They mostly try to do what's best for their party, even when it's bad for the country (-1) DON'T KNOW/SKIPPED/REFUSED (0) | |
| **ATTENDANCE** | Church attendance | Once a week or more (4) A few times a month (3) About once a month (2) A few times a year or less (1) Never (0) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **MARRIED** | Marital status | Married Divorced Single/Never married Separated Widowed DON’T KNOW/SKIPPED/REFUSED REMOVED FOR DISCLOSURE RISK | |
| **MARRIED2** | Marital status, 2 categories | Married (1) Not married (0) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **GUNOWNER** | Gun owner | Gun owner, self (2) Gun in household (1) No (0) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **UNION** | Union member | Union member, self (2) Union member in household (1) No (0) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **VET** | Veteran | Veteran, self (2) Veteran in household (1) No (0) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **LGB** | Lesbian, gay, bisexual identification | Yes (1) No (0) DON'T KNOW/SKIPPED/REFUSED (0) | |
| **TRANSGENDER** | Transgender identification | Yes (1) No (0) DON'T KNOW/SKIPPED/REFUSED (0) REMOVED FOR DISCLOSURE RISK (0) | |
| **BORNCITIZEN** | Born a citizen, or not | Yes (1) No (0) DON'T KNOW/SKIPPED/REFUSED (0) REMOVED FOR DISCLOSURE RISK (0) | |
| **Q2016VOTE2** | 2016 vote, 3 categories | Trump (1) Clinton (-1) Neither (0) | |
| **FIRSTTIME** | First time voter | First time voting (1) Not (0) | |
| **QPVVOTE3** (Vote2014) | Voted in 2014 | I'm sure I voted (1) I thought about voting in the 2014 election for Congress, but didn't (0) I did not vote in the 2014 election for Congress (0) I usually vote, but I didn't in the 2014 election for Congress (0) DON'T KNOW/SKIPPED/REFUSED (0) | |
| **BREAKA** (DemViolence) | Do you think the way Democrats talk about politics these days is leading to an increase in acts of violence, or don't you think so? | Yes, it is (1) No, it is not (0) DON'T KNOW/SKIPPED/REFUSED (0) | |
| **BREAKB** (RepViolence) | Do you think the way Republicants talk about politics these days is leading to an increase in acts of violence, or don't you think so? | Yes, it is (1) No, it is not (0) DON'T KNOW/SKIPPED/REFUSED (0) | |
| **SEX** (SEX\_Men) | Sex | Men (1) Women (0) | |
| **AGE65** (Age18\_24 Age25\_29 Age30\_39 Age40\_49 Age50\_64) | Age, granular | 18-24 25-29 30-39 40-49 50-64 65+ | |
| **RACETH5** (RACETH5\_White RACETH5\_AABlack RACETH5\_LatHisp RACETH5\_Asian) | Race/Ethnicity | White Latino or Hispanic African American or Black Asian Other DON’T KNOW/SKIPPED/REFUSED (VOL) | |
| **EDUC** (EDUC\_SomeCollege EDUC\_CollegeGrad EDUC\_PostGrad) | Education, 4 categories | High school or less Some college/assoc. degree College graduate Postgraduate study DON’T KNOW/SKIPPED/REFUSED (VOL) | |
| **INCOME** (Income25\_49 Income50\_74 Income75\_99 Income100) | 2017 household income | Under $25,000 $25,000-$49,999 $50,000-$74,999 $75,000-$99,999 $100,000 or more DON’T KNOW/SKIPPED/REFUSED | |
| **PARTY** | Party ID (no leaners) | Republican (1) Neither (0) Democrat (-1) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **IDEO** | Ideology | Very conservative (1) Somewhat conservative (0.5) Moderate (0) Somewhat liberal (-0.5) Very liberal (-1) DON’T KNOW/SKIPPED/REFUSED (0) | |
| **RELIG** (RELIG\_Chris RELIG\_Cath RELIG\_Jew RELIG\_Mus RELIG\_None) | Religion | Protestant/Other Christian Catholic Muslim Jewish None Other | |
| **SIZEPLACE** | Community type | Urban (3) Suburban (2) Small town (1) Rural (0) DON’T KNOW/SKIPPED/REFUSED (1.5) | |

# Appendix B: Model Diagrams

# A close up of a map Description automatically generated

# A close up of a map Description automatically generated

# A close up of a map Description automatically generated

# Appendix C: Information Gain Ratios Ranked by Data Set

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Qualitative** | | **Quantitative** | |
|  | Attributes | Importance | Attributes | Importance |
| 1 | Q2020VOTE | 0.3721 | Q2020VOTE | 0.3968 |
| 2 | PARTY | 0.3662 | RUSSIA | 0.3777 |
| 3 | Q2016VOTE2 | 0.3546 | PARTY | 0.3674 |
| 4 | RUSSIA | 0.3469 | Q2016VOTE2 | 0.3546 |
| 5 | TRACK | 0.3050 | TRACK | 0.3296 |
| 6 | REPINTENTIONS | 0.3045 | REPINTENTIONS | 0.3045 |
| 7 | FAVTRUMP | 0.2896 | FAVTRUMP | 0.3028 |
| 8 | DEMINTENTIONS | 0.2650 | TRADENATIONALECON | 0.2738 |
| 9 | FAVREP | 0.2627 | DEMINTENTIONS | 0.2650 |
| 10 | TRADENATIONALECON | 0.2627 | FAVREP | 0.2627 |
| 11 | RepViolence | 0.2390 | RepViolence | 0.2599 |
| 12 | FAVDEM | 0.2362 | HEALTHGOV | 0.2382 |
| 13 | SAFETERROR | 0.2239 | FAVDEM | 0.2362 |
| 14 | HEALTHGOV | 0.2209 | SAFETERROR | 0.2330 |
| 15 | TRADELOCALECON | 0.2172 | TRADELOCALECON | 0.2268 |
| 16 | IMMWALL | 0.2081 | DemViolence | 0.2238 |
| 17 | DemViolence | 0.2043 | IMMWALL | 0.2133 |
| 18 | SAFECRIME | 0.1872 | GUNPOLICY | 0.2042 |
| 19 | HEALTHLAW | 0.1788 | ECONFAIRPOOR | 0.2003 |
| 20 | SAFECYBER | 0.1725 | SAFECRIME | 0.1959 |
|  | **Selected Qualitative** | | **Selected Quantitative** | |
|  | Attributes | Importance | Attributes | Importance |
| 1 | PARTY | 0.3662 | PARTY | 0.3674 |
| 2 | RACETH5 | 0.0697 | RACETH5\_AABlack | 0.1327 |
| 3 | ISSUES | 0.0542 | ISSUES\_Immigration | 0.1043 |
| 4 | GUNOWNER | 0.0378 | RACETH5\_White | 0.0824 |
| 5 | LIKELYVOTER | 0.0317 | ISSUES\_Environment | 0.0675 |
| 6 | LGB | 0.0271 | ISSUES\_Abortion | 0.0556 |
| 7 | RELIG | 0.0249 | ISSUES\_Guns | 0.0527 |
| 8 | MARRIED2 | 0.0229 | GUNOWNER | 0.0472 |
| 9 | BORNAGAIN | 0.0220 | ISSUES\_Healthcare | 0.0459 |
| 10 | FIRSTTIME | 0.0156 | RELIG\_None | 0.0425 |
| 11 | SIZEPLACE | 0.0147 | LIKELYVOTER | 0.0317 |
| 12 | ATTENDANCE | 0.0124 | LGB | 0.0308 |
| 13 | VET | 0.0076 | RELIG\_Chris | 0.0247 |
| 14 | UNION | 0.0074 | MARRIED2 | 0.0232 |
| 15 | BORNCITIZEN | 0.0070 | SIZEPLACE | 0.0193 |
| 16 | EDUC | 0.0060 | ATTENDANCE | 0.0171 |
| 17 | INCOME | 0.0057 | Age18\_24 | 0.0152 |
| 18 | AGE65 | 0.0055 | UNION | 0.0119 |
| 19 | SEX | 0.0046 | RACETH5\_LatHisp | 0.0113 |
| 20 | TRANSGENDER | 0.0015 | EDUC\_PostGrad | 0.0113 |

# Appendix D: R Code

# # Lauren Lawless

# # IST 707

# # 11/25/19

# # Final Project

# #install.packages("caret")

# library(caret)

# #install.packages("rpart")

# library(rpart)

# #install.packages("rattle")

# library(rattle)

# #install.packages("rpart.plot")

# library(rpart.plot)

# #install.packages("e1071")

# library(e1071)

# #install.packages("klaR")

# library(klaR)

# #install.packages("kernlab")

# library(kernlab)

# #install.packages("randomForest")

# library(randomForest)

# #install.packages("FSelectorRcpp")

# library(FSelectorRcpp)

# split.fun <- function(x, labs, digits, varlen, faclen)

# {

# # replace commas with spaces (needed for strwrap)

# labs <- gsub(",", " ", labs)

# for(i in 1:length(labs)) {

# # split labs[i] into multiple lines

# labs[i] <- paste(strwrap(labs[i], width = 22), collapse = "\n")

# }

# labs

# }

# df\_og <- read.csv("~/Documents/Grad School Things/IST 707 - Data Analysis/rstudio-export/AP\_VOTECAST\_2018\_DATA\_V3.csv")

# set.seed(5)

# randIndex <- sample(1:dim(df\_og)[1])

# cutpoint <- floor(2\*dim(df\_og)[1]/3)

# # Qualitative Set

# df\_qual <- read.csv("~/Documents/Grad School Things/IST 707 - Data Analysis/rstudio-export/Votecast Qual.csv")

# train\_qual <- df\_qual[randIndex[1:cutpoint],]

# test\_qual <- df\_qual[randIndex[(cutpoint+1):dim(df\_qual)[1]],]

# # Choose variables

# #####

# vars\_qual <- c("RACE5HouseGen",

# # "MODE",

# # "State",

# # "LV",

# "LIKELYVOTER",

# # "TRACK",

# "ISSUES",

# # "FAVTRUMP",

# # "FAVREP",

# # "FAVDEM",

# # "SUPREMECOURT",

# # "GOVTDO",

# # "NatEcon",

# # "GETAHEAD",

# # "TRADENATIONALECON",

# # "TRADELOCALECON",

# # "ECONFAIRWEALTHY",

# # "ECONFAIRMIDDLE",

# # "ECONFAIRPOOR",

# # "TAXCUTS",

# # "HEALTHLAW",

# # "HEALTHGOV",

# # "IMMDEPORT",

# # "IMMBETTER",

# # "IMMWALL",

# # "RACEREL",

# # "CLIMATE",

# # "GUNPOLICY",

# # "ABORTION",

# # "METOOBELIEVE",

# # "METOODEFEND",

# # "PCSPEECH",

# # "MARIJUANA",

# # "OPIOID",

# # "SAFETERROR",

# # "SAFECRIME",

# # "SAFECYBER",

# # "RUSSIA",

# # "INTERFERENCE",

# # "Q2020VOTE",

# # "REPINTENTIONS",

# # "DEMINTENTIONS",

# "ATTENDANCE",

# # "MARRIED",

# "MARRIED2",

# "GUNOWNER",

# "UNION",

# "VET",

# "LGB",

# "TRANSGENDER",

# "BORNCITIZEN",

# # "Q2016VOTE2",

# "FIRSTTIME",

# # "Vote2014",

# # "DemViolence",

# # "RepViolence",

# "SEX",

# "AGE65",

# "RACETH5",

# "EDUC",

# "INCOME",

# "PARTY",

# # "IDEO",

# "RELIG",

# "BORNAGAIN",

# "SIZEPLACE")

# #####

# sml\_qual <- df\_qual[,vars\_qual]

# train\_smlqual <- sml\_qual[randIndex[1:cutpoint],]

# test\_smlqual <- sml\_qual[randIndex[(cutpoint+1):dim(sml\_qual)[1]],]

# # Quantitative Set

# df\_quant <- read.csv("~/Documents/Grad School Things/IST 707 - Data Analysis/rstudio-export/Votecast Quant.csv")

# train\_quant <- df\_quant[randIndex[1:cutpoint],]

# test\_quant <- df\_quant[randIndex[(cutpoint+1):dim(df\_quant)[1]],]

# # Choose variables

# #####

# vars\_quant <- c("RACE5HouseGen",

# # "MODE\_Web",

# # "State",

# # "LV",

# "LIKELYVOTER",

# # "TRACK",

# "ISSUES\_Jobs",

# "ISSUES\_Healthcare",

# "ISSUES\_Immigration",

# "ISSUES\_Taxes",

# "ISSUES\_Abortion",

# "ISSUES\_Guns",

# "ISSUES\_Environment",

# "ISSUES\_ForeignPolicy",

# "ISSUES\_Terrorism",

# #"FAVTRUMP",

# #"FAVREP",

# #"FAVDEM",

# #"SUPREMECOURT",

# #"GOVTDO",

# #"NatEcon",

# #"GETAHEAD",

# #"TRADENATIONALECON",

# #"TRADELOCALECON",

# #"ECONFAIRWEALTHY",

# #"ECONFAIRMIDDLE",

# #"ECONFAIRPOOR",

# #"TAXCUTS",

# #"HEALTHLAW",

# #"HEALTHGOV",

# #"IMMDEPORT",

# #"IMMBETTER",

# #"IMMWALL",

# #"RACEREL",

# #"CLIMATE",

# #"GUNPOLICY",

# #"ABORTION",

# #"METOOBELIEVE",

# #"METOODEFEND",

# #"PCSPEECH",

# #"MARIJUANA",

# #"OPIOID",

# #"SAFETERROR",

# #"SAFECRIME",

# #"SAFECYBER",

# #"RUSSIA",

# #"INTERFERENCE",

# #"Q2020VOTE",

# #"REPINTENTIONS",

# #"DEMINTENTIONS",

# "ATTENDANCE",

# "MARRIED2",

# "GUNOWNER",

# "UNION",

# "VET",

# "LGB",

# "TRANSGENDER",

# "BORNCITIZEN",

# #"Q2016VOTE2",

# "FIRSTTIME",

# #"Vote2014",

# #"DemViolence",

# #"RepViolence",

# "SEX\_Men",

# "Age18\_24",

# "Age25\_29",

# "Age30\_39",

# "Age40\_49",

# "Age50\_64",

# "RACETH5\_White",

# "RACETH5\_AABlack",

# "RACETH5\_LatHisp",

# "RACETH5\_Asian",

# "EDUC\_SomeCollege",

# "EDUC\_CollegeGrad",

# "EDUC\_PostGrad",

# "Income25\_49",

# "Income50\_74",

# "Income75\_99",

# "Income100",

# "PARTY",

# # "IDEO",

# "RELIG\_Chris",

# "RELIG\_Cath",

# "RELIG\_Jew",

# "RELIG\_Mus",

# "RELIG\_None",

# "SIZEPLACE")

# #####

# sml\_quant <- df\_quant[,vars\_quant]

# train\_smlquant <- sml\_quant[randIndex[1:cutpoint],]

# test\_smlquant <- sml\_quant[randIndex[(cutpoint+1):dim(sml\_quant)[1]],]

# # Mixed Set (factor dependent; numeric independent)

# df\_mixed <- data.frame(df\_qual[,1],df\_quant[,2:dim(df\_quant)[2]])

# colnames(df\_mixed)[1] <- "RACE5HouseGen"

# train\_mixed <- df\_mixed[randIndex[1:cutpoint],]

# test\_mixed <- df\_mixed[randIndex[(cutpoint+1):dim(df\_mixed)[1]],]

# sml\_mixed <- df\_mixed[,vars\_quant]

# train\_smlmixed <- sml\_mixed[randIndex[1:cutpoint],]

# test\_smlmixed <- sml\_mixed[randIndex[(cutpoint+1):dim(sml\_mixed)[1]],]

# pred <- data.frame(test\_qual$RACE5HouseGen, test\_quant$RACE5HouseGen)

# # Decision Tree

# # Model 1 - All Qual

# set.seed(10)

# start <- Sys.time()

# model\_dt1 <- train(RACE5HouseGen~., data=train\_qual, metric="Accuracy", method="rpart", tuneLength = 10, cp = 0, trControl=trainControl(method="cv", number=3))

# stop <- Sys.time()

# time\_dt1 <- stop-start

# model\_dt1

# model\_dt1$finalModel

# fancyRpartPlot(model\_dt1$finalModel, main="Decision Tree Model 1", palettes=c("Blues","Greens","Reds"), type=4, cex=0.6, split.fun=split.fun)

# pred$pred\_dt1 <- predict(model\_dt1, newdata=test\_qual, type="raw")

# confusionMatrix(pred$pred\_dt1, pred$test\_qual.RACE5HouseGen)

# # Model 2 - Small Qual

# set.seed(10)

# start <- Sys.time()

# model\_dt2 <- train(RACE5HouseGen~., data=train\_smlqual, metric="Accuracy", method="rpart", tuneLength = 8, cp = 0, trControl=trainControl(method="cv", number=3))

# stop <- Sys.time()

# time\_dt2 <- stop-start

# model\_dt2

# #model\_dt2$finalModel

# fancyRpartPlot(model\_dt2$finalModel, main="Decision Tree Model 2", palettes=c("Blues","Greens","Reds"), type=4, cex=0.6, split.fun=split.fun)

# pred$pred\_dt2 <- predict(model\_dt2, newdata=test\_smlqual, type="raw")

# confusionMatrix(pred$pred\_dt2, pred$test\_qual.RACE5HouseGen)

# # Model 3 - Small Qual minus TRACK

# set.seed(10)

# start <- Sys.time()

# model\_dt3 <- train(RACE5HouseGen~., data=train\_smlqual, metric="Accuracy", method="rpart", tuneLength = 8, cp = 0, trControl=trainControl(method="cv", number=3))

# stop <- Sys.time()

# time\_dt3 <- stop-start

# model\_dt3

# #model\_dt3$finalModel

# fancyRpartPlot(model\_dt3$finalModel, main="Decision Tree Model 3", palettes=c("Blues","Greens","Reds"), type=4, cex=0.6, split.fun=split.fun)

# pred$pred\_dt3 <- predict(model\_dt3, newdata=test\_smlqual, type="raw")

# confusionMatrix(pred$pred\_dt3, pred$test\_qual.RACE5HouseGen)

# # Naive Bayes

# # Model 4 - All Mixed

# set.seed(10)

# start <- Sys.time()

# model\_nb4 <- train(RACE5HouseGen~., data=train\_mixed, method="nb", trControl=trainControl(method="cv", number=3), tuneGrid=expand.grid(usekernel=c(TRUE,FALSE), fL=0:2, adjust=0:2))

# stop <- Sys.time()

# time\_nb4 <- stop-start

# time\_nb4

# model\_nb4

# pred$pred\_nb4 <- predict(model\_nb4, newdata=test\_mixed, type="raw")

# confusionMatrix(pred$pred\_nb4, pred$test\_qual.RACE5HouseGen)

# # Model 5 - Small Mixed

# set.seed(10)

# start <- Sys.time()

# model\_nb5 <- train(RACE5HouseGen~., data=train\_smlmixed, method="nb", trControl=trainControl(method="cv", number=3), tuneGrid=expand.grid(usekernel=c(TRUE,FALSE), fL=0:2, adjust=0:2))

# stop <- Sys.time()

# time\_nb5 <- stop-start

# time\_nb5

# model\_nb5

# pred$pred\_nb5 <- predict(model\_nb5, newdata=test\_smlmixed, type="raw")

# confusionMatrix(pred$pred\_nb5, pred$test\_qual.RACE5HouseGen)

# # kNN

# # Model 6 - Default

# set.seed(10)

# start <- Sys.time()

# model\_knn6 <- train(RACE5HouseGen~., data=train\_smlmixed, method="knn")

# stop <- Sys.time()

# time\_knn6 <- stop-start

# time\_knn6

# model\_knn6

# pred$pred\_knn6 <- predict(model\_knn6, newdata=test\_smlmixed)

# confusionMatrix(pred$pred\_knn6, pred$test\_qual.RACE5HouseGen)

# # Model 7 - Tuning

# set.seed(10)

# start <- Sys.time()

# model\_knn7 <- train(RACE5HouseGen~., data=train\_smlmixed, method="knn",

# tuneGrid=data.frame(k=seq(0,25,1)),

# trControl=trainControl(method="repeatedcv",

# number=10, repeats=3

# )

# )

# stop <- Sys.time()

# time\_knn7 <- stop-start

# time\_knn7

# model\_knn7

# pred$pred\_knn7 <- predict(model\_knn7, newdata=test\_smlmixed)

# confusionMatrix(pred$pred\_knn7, pred$test\_qual.RACE5HouseGen)

# # SVM

# # Model 8 - Linear

# set.seed(10)

# start <- Sys.time()

# model\_svm8 <- train(RACE5HouseGen~., data=train\_smlmixed,

# method="svmLinear",

# preProcess=c("center", "scale"),

# trControl=trainControl(method="boot", number=25),

# tuneGrid=expand.grid(C=seq(2.5, 5, 0.5)

# )

# )

# stop <- Sys.time()

# time\_svm8 <- stop-start

# time\_svm8

# model\_svm8

# plot(model\_svm8)

# pred$pred\_svm8 <- predict(model\_svm8, newdata=test\_smlmixed)

# confusionMatrix(pred$pred\_svm8, pred$test\_qual.RACE5HouseGen)

# # Model 9 - Nonlinear separability (radial)

# set.seed(10)

# start <- Sys.time()

# model\_svm9 <- train(RACE5HouseGen~., data=train\_smlmixed,

# method="svmRadial",

# preProcess=c("center", "scale"),

# trControl=trainControl(method="boot", number=25),

# tuneGrid=expand.grid(sigma=seq(0.5, 2, 0.5),

# C=seq(0.5, 2, 0.5)

# )

# )

# stop <- Sys.time()

# time\_svm9 <- stop-start

# time\_svm9

# model\_svm9

# plot(model\_svm9)

# pred$pred\_svm9 <- predict(model\_svm9, newdata=test\_smlmixed)

# confusionMatrix(pred$pred\_svm9, pred$test\_qual.RACE5HouseGen)

# # Model 10 - Nonlinear separability (polynomial)

# set.seed(10)

# start <- Sys.time()

# model\_svm10 <- train(RACE5HouseGen~., data=train\_smlmixed,

# method = "svmPoly",

# preProcess=c("center", "scale"),

# trControl=trainControl(method="boot", number=25),

# tuneGrid=expand.grid(degree=seq(0, 2, 1),

# scale=seq(0.5, 2, 0.5),

# C=seq(0.5, 2, 0.5)

# )

# )

# stop <- Sys.time()

# time\_svm10 <- stop-start

# time\_svm10

# model\_svm10

# plot(model\_svm10)

# pred$pred\_svm10 <- predict(model\_svm10, newdata=test\_smlmixed)

# confusionMatrix(pred$pred\_svm10, pred$test\_qual.RACE5HouseGen)

# # Model 11 - Nonlinear separability (polynomial)

# set.seed(10)

# start <- Sys.time()

# model\_svm11 <- train(RACE5HouseGen~., data=train\_smlmixed,

# method = "svmPoly",

# preProcess=c("center", "scale"),

# trControl=trainControl(method="boot", number=25),

# tuneGrid=expand.grid(degree=seq(3, 4, 1),

# scale=seq(0.5, 2, 0.5),

# C=seq(0.5, 2, 0.5)

# )

# )

# stop <- Sys.time()

# time\_svm11 <- stop-start

# time\_svm11

# model\_svm11

# plot(model\_svm11)

# pred$pred\_svm11 <- predict(model\_svm11, newdata=test\_smlmixed)

# confusionMatrix(pred$pred\_svm11, pred$test\_qual.RACE5HouseGen)

# # Random Forest

# # Model 12

# set.seed(10)

# start <- Sys.time()

# model\_rf12 <- train(RACE5HouseGen~., data=train\_smlqual, method="rf")

# stop <- Sys.time()

# time\_rf12 <- stop-start

# time\_rf12

# model\_rf12$finalModel

# pred$pred\_rf12 <- predict(model\_rf12, newdata=test\_smlqual)

# confusionMatrix(pred$pred\_rf12, pred$test\_qual.RACE5HouseGen)

# varimp\_rf <- varImp(model\_rf12)

# varimp\_rf

# plot(varimp\_rf, main = "Variable Importance with Random Forest")

# # Pred Results

# pred\_num <- pred[,-1]

# for (i in 2:dim(pred\_num)[2]){

# pred\_num[,i] <- as.integer(ifelse(pred\_num[,i]=="Republican",1,ifelse(pred\_num[,i]=="Democrat",-1,0)))

# }

# str(pred\_num)

# vars\_score <- c(3,4,7,8,9,11,13)

# pred\_num$score <- rowSums(pred\_num[,vars\_score])/(length(vars\_score))

# pred\_num$ensemble <- round(pred\_num$score)

# confusionMatrix(as.factor(pred\_num$ensemble), as.factor(pred\_num$test\_quant.RACE5HouseGen))

# View(pred\_num[which(pred\_num$test\_quant.RACE5HouseGen==0),])

# # Gain Ratio

# info\_qual <- information\_gain(RACE5HouseGen~., data=train\_qual, type="gainratio")

# info\_smlqual <- information\_gain(RACE5HouseGen~., data=train\_smlqual, type="gainratio")

# info\_quant <- information\_gain(RACE5HouseGen~., data=train\_quant, type="gainratio")

# info\_smlquant <- information\_gain(RACE5HouseGen~., data=train\_smlquant, type="gainratio")

# head(info\_qual[order(-info\_qual$importance),],20)

# head(info\_smlqual[order(-info\_smlqual$importance),],20)

# head(info\_quant[order(-info\_quant$importance),],20)

# head(info\_smlquant[order(-info\_smlquant$importance),],20)

# References

Ballotpedia. (2018, November). *Margin of victory analysis for the 2018 congressional elections*. Retrieved from Ballotpedia: https://ballotpedia.org/Margin\_of\_victory\_analysis\_for\_the\_2018\_congressional\_elections

Issenberg, S. (2012, December 19). *How Obama’s Team Used Big Data to Rally Voters*. Retrieved from Technology Review: https://www.technologyreview.com/s/509026/how-obamas-team-used-big-data-to-rally-voters/

Kennedy, C., & Hartig, H. (2019, February 27). *Response rates in telephone surveys have resumed their decline*. Retrieved from Pew Research Center: https://www.pewresearch.org/fact-tank/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/

NORC at the University of Chicago. (2019, May). *AP VoteCast 2018 Public Use File Codebook.* Retrieved from data.world: https://data.world/apnorc/ap-votecast-public-use-file/workspace/file?filename=VC\_PUF\_codebook\_final.pdf

The New York Times. (2019, November 14). *2019 Kentucky General Election Results*. Retrieved from The New York Times: https://www.nytimes.com/interactive/2019/11/05/us/elections/results-kentucky-general-elections.html

­

1. See Appendix B: Model Diagrams. [↑](#footnote-ref-1)
2. Leading up to the 2018 midterm elections, Republicans controlled both chambers of Congress and the Presidency. [↑](#footnote-ref-2)
3. At the time of writing (December 3, 2019), only 3 of the 535 voting members of Congress are registered Independents: Senator Angus King (ME) and Senator Bernie Sanders (VT), who formally caucus with the Democratic Party, and Representative Justin Amash (MI), who left the Republican Party in July 2019. [↑](#footnote-ref-3)
4. Third-party candidates were the runners-up in 15 of the 435 2018 House elections. (Ballotpedia, 2018) [↑](#footnote-ref-4)
5. Selected attributes not yet excluding TRACK and IDEO [↑](#footnote-ref-5)
6. All models except accuracies below 0.8 (5, 9, 11) [↑](#footnote-ref-6)
7. All models except accuracies below 0.8 (5, 9, 11) or models using all attributes (1, 4) [↑](#footnote-ref-7)