An analysis of nonvoters in the United States in 2020

Final Report

Drew Bostrom May 2021

Summary

A survey from fivethirtyeight.com of American citizens was analyzed. This survey asks questions about political attitudes and voting habits, and it collects demographic information. A model was designed using Random Forest algorithm with the purpose of predicting which citizens would be 'nonvoters' based on their responses to the survey questions. This model performs reasonably well, making predictions with recall of about 74%. Clustering of survey respondents using the KMeans algorithm was also performed. Five clusters were identified – nonvoters, believers in institutions, those who perceive difficulty voting, weak Republicans, and strong Republicans. The results of this work are robust, as they were able to be replicated after transforming the data with principal component analysis.

Problem Statements

- Can an individual's likelihood to not vote be predicted based on survey data?
- What are the key factors that would cause a person to not vote?
- Can citizens be placed into clusters in which they share common characteristics?

Data

The original data consists of 5836 rows and 115 features. The rows correspond to the respondents of the survey, all United States citizens. The features are mostly multiple-choice (integer) responses to survey questions, but additionally consists of demographic information, such as age, gender, income, and race.

Some of the questions from the survey are:

2. In your view, how important are each of the following to being a good American?

(1. Very important 2. Somewhat important 3. Not so important 4. Not at all important)

- Following what happens in government and politics
- Knowing the Pledge of Allegiance

6. In general, how many of the people in elected office today are like you?

For purposes of prediction, the target data is 'voter category,' which can take the value 'always,' 'sporadic,' or 'rarely/never.'

Because all of the data is categorical (age is converted into 6 age categories), the original data is converted into dummy variables, such that each possible response (e.g., 1, 2, 3 or 4) becomes its own feature that can take on the value of 0 or 1. This conversion results in 313 features. Features related to questions 26 and 27 were dropped, since these were used to reassigned voter categories, leaving 294 features. The 'voter category' target was also encoded to

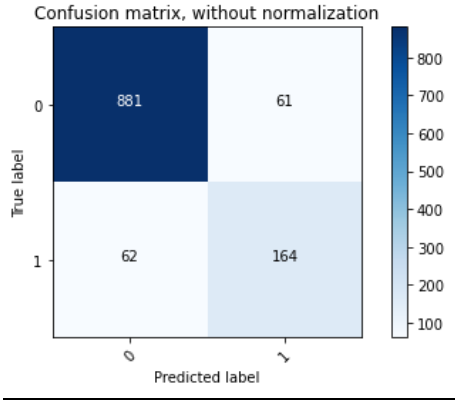
'voter_category_rarely/never,' etc. As a result, prediction consists of a binary choice in which 'rarely/never' voters are either predicted or not predicted.

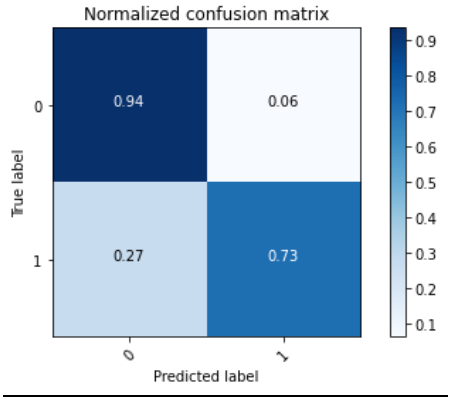
Predictive Model

To make predictions, a Random Forest model was built using the encoded features as predictors and the voter category as the target. The hyperparameters of the model were tuned using both randomized and grid search. The precision of the model (how well it avoids false positives) and its recall (how well it avoids false negatives) were both 73%.

Confusion Matrix

The confusion matrices below show the results of model prediction on just the test set of data (1178 or 20% of the rows). Because of the imbalance of the data (many more always and sporadic voters combined than rarely/never voters), most of the model's accuracy can be attributed to the correctly predicted true negatives in the top right corner. This result is again seen in the normalized confusion matrix which shows that 94% of the negatives were correctly predicted while only 73% of the positives were correctly predicted.





Important features

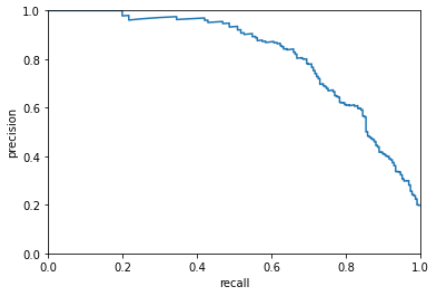
According to the results of Random Forest modeling, the most important features were (from most to least):

1. 28_1_1.0 – Question 28 asks what factors caused respondent to vote. Choice 1 is that voting is an important civic duty. 1.0 means the choice was selected.
2. 21_1 – Question 21 asks if respondent plans to vote in 2020 national election. Answer 1 is 'yes'.
3. 20_1 – Question 20 asks if the respondent is currently registered to vote. Answer 1 is 'yes'.
4. 2_1_1 – Question 2 asks how important a factor is in being a good American. Choice 1 is voting in elections. Answer 1 is 'very important'
5. 20_2 – See question 20 above. Answer 2 is 'no'.
6. 21_2 – See Question 21 above. Answer 2 is 'no'.
7. 29_3_1.0 – Question 29 (asked only of irregular or nonvoters) asks "which of the following were the most important reasons in your decision not to vote? Please choose all that apply." Choice 3 states "No matter who wins, nothing will change for people like me." 1.0 means the choice was selected.

These features generally relate to being registered to vote, planning to vote, voting being an important civic duty and part of being a good American. The last three relate to not wanting to vote, with #7 expressing a lack of belief that anything can change by voting.

An important finding from these results is that there does not appear to be a specific demographic profile (race, income, gender) that strongly predicts that a person will vote or not vote.

Precision-Recall curve and Area Under Curve (AUC)



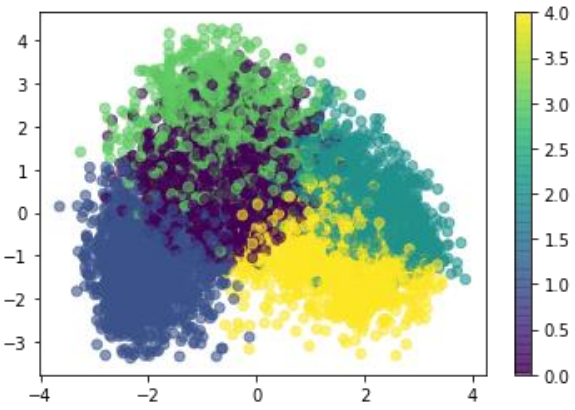
The area-under-curve is 0.62. The curve and the AUC confirm that the model has some success in separating 'rarely/never' voters from others but the prediction is not perfect.

KMeans Clustering

A graph of sum of squares versus number of clusters revealed the survey weakly forms five clusters. These clusters were visualized by graphing all points on axes composed of the first two principal components. As seen below, there are five clear clusters, but with substantial overlap.

Silhouette Score

The silhouette score was 0.033. This score confirms some clustering both with overlap. The low score also indicates the high dimensionality of the model, which used all 294 features.



Characterization of Clusters

Looking at these clusters in more detail reveals some interesting commonalities among the respondents. These commonalities arise from the features that most distinguished one cluster from the other clusters. Listed below are the most important features for each cluster along with some information about their tendency to vote. I've given each cluster a name based on these details.

Cluster 1: Nonvoters

- Q2_1_4 Voting in elections -- not at all important
- Q29_9_1.0 I don't believe in voting
- Q2_3_4 Following what happens in government and politics -- not at all important
- Q22_2.0 I don't trust the political system to serve my needs
- Q21_2 Do not plan to vote in November 2020 election
- Never vote - 72%

Cluster 2: Difficulty voting/low enthusiasm

- Q2_9_4 Believing in God -- not at all important
- Q2_4_4 Displaying the American flag -- not at all important
- Q18_6_1 Had to cast a provisional ballot
- Q2_7_3 Supporting the military -- not so important
- Q18_7_1 Couldn't get off work to vote when polls were open
- Always vote - 59%, Sporadic - 29%

Cluster 3: Trust in Institutions

- Q8_7_1 Trust the news media -- a lot
- Q3_4_4 The mainstream media is more interested in making money than telling the truth -- strongly disagree

- Q17_4_1 Electronic votes submitted online or by email safe/secure -- very confident
- Q8_6_1 Trust the intelligence community -- a lot
- Q8_2_1 Trust Congress -- a lot
- Always vote - 82%

Cluster 4: Strong Republican/Skeptical of racism

- Q8_1_1 Trust the presidency - a lot
- Q15_5 The Democratic Party does not want people like me to vote, and works hard to keep us from being able to vote
- Q31_1.0 Strong Republican
- Q3_2_4 Systemic racism in policing is a bigger problem than violence and vandalism in protests - Strongly disagree
- Q3_1_4 Systemic racism is a problem in the United States - Strongly disagree
- Always vote - 77%

Cluster 5: Weak Republican

- Q3_1_2 Systemic racism is a problem in the United States - Somewhat agree
- Q3_2_3 Systemic racism in policing is a bigger problem than violence and vandalism in protests - Somewhat disagree
- Q25_2 Following the 2020 presidential race -- somewhat closely
- Q14_2 The Republican Party wants people like me to vote but does not work hard to earn our votes
- Q31_2.0 Not very strong Republican
- Always vote - 51%, Sporadic - 29%

PCA

Principal Component Analysis was used in conjunction with Random Forest modeling and KMeans clustering in an attempt to improve results. Random Forest on principal components alone did not yield good results. Random Forest on principal components, with original survey data, and clustering labels produced results similar to the original Random Forest model. Notably, KMeans clustering performed on only the first 10 principal components resulted in five clusters almost identical to the clusters identified by all

Logistic Regression

A second predictive model was built with Logistic Regression. The model suffered from either poor recall or poor precision depending on the inputs.

Conclusion

A model can be built that predicts who will not vote with recall of 73%. It is doubtful, using just the survey data available, that a much more predictive model is possible. Anyone deploying this model, therefore, would have to be satisfied with not identifying every person who is likely not to vote.

On the other hand, the information about important features and clustering revealed some interesting characteristics about voters and nonvoters. Believing that voting is a civic duty is a strong indicator of tendency to vote. Conversely, believing that the results of an election will not change anything indicates someone will not vote. Demographic information such as income and race were not important predictors of voting or not. Similarly, barriers to vote (such as long lines) did not strongly indicate that a person would never or rarely vote.

It should be emphasized that the model identifies correlation and not causation. For instance, being registered to vote is an important feature in determining who will vote. However, that does not mean that registering people to vote will cause them to change from never voters to always voters. To really cause somebody to want to vote, it's possible that a large education effort would be needed that emphasized that voting is a civic duty and that the outcome of the election makes a difference.

Recommendations

1. In a targeted voter drive effort with limited resources, the predictive model can identify some citizens who are not likely to vote regularly. Some potential voters will not be identified also there will be a fair number of false positives. The logistic model may be better for identifying more nonvoters but with more false positives.
2. The model can also be used to identify who is likely to vote with a great deal of accuracy, in order eliminate individuals from outreach efforts.
3. Clustering can be used to identify groups of people and tailor messages that best resonate with them. There was a strong cluster of nonvoters who will share some characteristics in common.

Further Work

- Explore relationships between demographic information and model prediction/clustering.
- Determine the effects of barriers to vote on tendency to vote
- Characterize differences between Republican and Democratic voters
- Find additional data, including a state of residence, to identify voters without a survey or to improve modeling,