Data Science for Political Science Final Project

Reese Feldman

I. Introduction

Who will vote in a given midterm election? This question is very important to candidates and "get out the vote" non-profits. Using the Cooperative Election Study pre and post election survey data from the 2018 midterm election, I built a model that estimates how likely it is that a given person will cast a ballot, based on a number of factors, such as age, internet access, gender, education, and more. A more detailed explanation of variables considered will be provided later.

Overall, it is very difficult to predict the likelihood of voting. There are many factors at play for a given person that affect their ability to cast a ballot. However, it is possible to make a prediction based on factors we do know. One surprising result from my analysis is that race is not an effective predictor when more informative variables are considered. Factors such as internet access, the cost of voting index score in a person's state, a person's approval rating of congress, and whether or not they have a history of voting proved to be much more informative predictors of voter turnout. Knowing what factors are more important when predicting voting likelihood allows the model to be as efficient as possible while still being as accurate as possible.

II. Background

As mentioned above, the question of who will cast a ballot is important for multiple reasons. For candidates, knowing who is a likely voter, who has no chance of voting, and who may be on the fence, allows their campaign to prioritize spending and better target various advertisements. Someone they are confident will vote, and will vote for them, could be targeted with advertisements encouraging them to donate to the campaign. People on the fence of voting can be strongly encouraged to vote, and vote for the candidate. And campaigns won't have to

waste money or time on people who will never vote. For non-profits focused on increasing voter turnout, knowledge about who is likely to vote, and the factors that affect voter turnout allows them to better target their campaigns and money.

Because this question is so important, people have been trying to answer it for decades. One prior study looks at voter turnout in Canada, and aims to fit the most accurate model possible using over three dozen explanatory variables. The authors explain that even with over three dozen explanatory variables, there is still a lot of information they are missing, such as the cost and benefits of voting for each individual person, rather than the general cost of voting score in the state overall. The authors do say though, that whether or not someone has voted in the past could be a good indicator and summary of their personal views and cost-benefit analysis of voting (Matsuska 1999). A different study, focusing primarily on the impact of education on voter turnout, found that in American presidential and midterm elections, one additional year of education had a very slight positive impact on voting likelihood. Interestingly, the study found that student status, that is being in school versus not, had a more significant positive effect on both voter registration and voter turnout (Tenn 2007). Unfortunately, it was difficult to include student status in my predictive model based on the questions asked in the CES 2018 survey, as the survey only asks what level of education has been completed. Nonetheless, this study encouraged me to include education as a predictor in my model.

III. Data, Approach, and Results

For this analysis, I built a model to predict if a given registered voter will vote in the upcoming midterm election, using the *2018 Cooperative Congressional Election Study* survey data. This survey asked respondents a number of questions, so to build an efficient model, I narrowed the predictor variables down to ones that I was inclined to believe would build a good

model based on prior research. I first wanted to see what a model built purely on physical characteristics about a person would result in, as this data is the easiest for future researchers to collect and study. In the supplementing r code, this model is referred to as base2018, and uses age, gender, and race as the only predictors of voting behavior. This model is somewhat weak, and when using a cutoff point of .9, has a total accuracy rate of .5892. The cutoff point refers to what likelihood will result in a will vote or will not vote prediction, and this model uses a likelihood of .9 or higher to classify someone as a predicted voter. The significance levels of each predictor variable was very interesting, as age and gender were both statistically significant at the .001 significance level, while race had a p-value of .148, meaning we cannot conclude that race as a predictor had a significant impact on the model.

Taking the prior research into account, the next model, titled new2018 in the corresponding r code, expands on the previous one by taking more subjective factors into account, including the cost of voting index score of the state of residence, interest in the news and current events, family income, internet access, congressional approval, and previous voting history, specifically if the person voted in the previous presidential election. In order to include the cost of voting index score, I needed to merge an additional row of data into the dataset, as the CES survey asked state of residence, but did not include the cost of voting index for those states. This data was extracted from *Cost of Voting in the American States*.

This model resulted in a much wider range of voter likelihood values, and the overall accuracy went up to .7765, an increase of .1873 from the previous model, at the same cutoff point of .9. In addition to the new predictor variables, all predictors from the previous model were included. All predictors were significant at the .01 significance level, and most at the .001 significance level, except for race. In this model, the p-value for race increased to .709, meaning

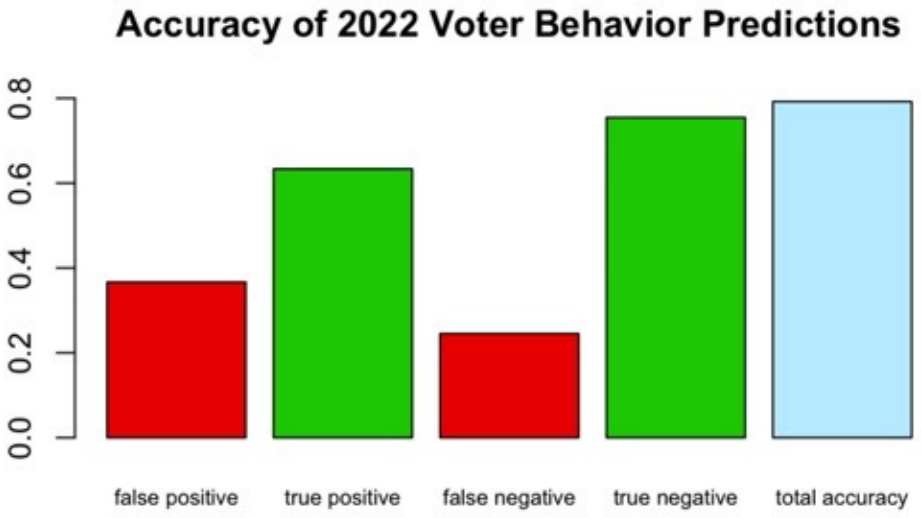
it has even less of a significant impact on this model than the previous one. This could imply that the reasons one's race might affect their choice or ability to cast a ballot is better represented by qualitative measures like internet access, education, family income, and others, rather than making generalizations about those qualitative variables because of one's race.

Knowing the lack of significance of race on the model, to further fine tune the predictions, I eliminated race from the list of predictors and left everything else the same. This final model is titled best2018 in the corresponding r code. In this model, all predictors have a statistically significant impact on the model. By eliminating race from the list of predictors, the model is more efficient, as it is making similar predictions with less information. The overall accuracy rate of this model is .7764, meaning that by eliminating a predictor the accuracy only decreased by .0001, a marginal and insignificant difference.

Now that an effective and efficient model has been developed, it can be used to predict voter turnout in future elections. To demonstrate this, I applied the model to the information collected in the *2022 Cooperative Congressional Election Study* and the updated for 2022 *Cost of Voting in the American States*. Using the same cutoff point of .9, the overall accuracy of the predictions made is .7923, which is actually slightly higher than the overall accuracy of the model when checked against the 2018 data. However, while the false negative rate went down slightly, the false positive rate increased. A summary of the accuracy rates of the prediction can be seen below.

IV. Conclusion

The overall goal of this analysis was to be able to predict who will vote in an upcoming midterm election. In line with my prior expectations, this proved to be a difficult thing to predict, but the more variables accounted for, the more accurate the predictions become, especially when



including person specific qualitative variables, like someone's interest in current events, their opinion of the current congress, the ease of voting where they live, and others. While it is not easy to predict voter turnout, when campaigns and non profits are able to have some idea of how likely someone is to vote in an upcoming election, they can better allocate their resources, time, energy, and money.

Evaluating what factors influence the likelihood that someone will vote also allows for future research in how to increase voter turnout. For example, knowing that access to the internet increases the likelihood of someone voting, could inspire future research into how and why internet access increases voter turnout and what areas could be most improved by increasing internet availability. Social programs and campaigns to increase voter turnout improve our democracy.

Works Cited

Matsusaka, J.G., Palda, F. *Voter turnout: How much can we explain?*. Public Choice 98, 431–446 (1999). https://doi.org/10.1023/A:1018328621580

Quan Li, Michael J. PomanteII, and Scot Schraufnagel. *Cost of Voting in the American States*. Election Law Journal: Rules, Politics, and Policy. Sep 2018. 234-247. http://doi.org/10.1089/elj.2017.0478

Schaffner, Brian; Stephen Ansolabehere; Sam Luks, 2019, "CCES Common Content, 2018", https://doi.org/10.7910/DVN/ZSBZ7K, Harvard Dataverse, V6

Schaffner, Brian; Ansolabehere, Stephen; Shih, Marissa, 2023, "Cooperative Election Study Common Content, 2022", https://doi.org/10.7910/DVN/PR4L8P, Harvard Dataverse, V3

Tenn, Steven. "The Effect of Education on Voter Turnout." *Political Analysis*, vol. 15, no. 4, 2007, pp. 446–64. *JSTOR*, http://www.jstor.org/stable/25791906. Accessed 19 Dec. 2023.