The Elusive Likely Voter - AppStat write-up*

Alice Kartvedt Paulsen, Catrine Robinson Christiansen, Phillip Henriksen and Christian Madsen

8th March 2023

1 The Problem of the 'likely voter'

When polling, i.e. surveying the opinions or choices of a population, the goal is to predict the behavior of a large population as closely as possible by only asking a small sample of the total population. One of the problems introduced once you poll a sufficiently large part of a population is that the poll will include the entire population instead of, specifically, the voting part of the population. In the case of polling for upcoming elections in the United States, the paper Gallup proposed the idea of the 'likely voter' model to try and limit this problem. A natural first step is to ask the respondents whether or not they expect to vote or not among other simple questions regarding typical voting patterns. Usually, people are asked how likely they are to vote on a scale of 1-5 among the questions, but in these types of questions, it has been found that 25%-33% of people lie, according to Rentsch, Shaffner and Gross.

The 'likely voter' model has been used for decades as the primary model (or as a foundation for more complicated models), by the major polling companies in the US. As there are problems with peoples' self-reported voting behavior, the model can therefore produce inaccurate predictions for the result of an election. We saw this in the 2016 US presidential election where a large majority of pollsters, many of them using 'likely voter', favored Hillary Clinton as the clear victor. Her loss came as a big surprise to many who were following the predictions created by the polls. To try and improve election predictions researchers and pollsters must create more exact methods of predicting election results - You cannot simply ask what people will do. Therefore, pollsters have added questions to polls to calculate predictors for future voting behavior, such as: Questions about voting history, previous political engagement, and basic demographic information. These questions are designed to let the pollsters determine how likely it is that a person saying they will vote for one thing *actually* shows up to the polls and votes for exactly that option.

1.1 Cutoff method

Simply put: A cut off-method is one that counts every entry above a certain threshold and rejects all other entries. For example, in the previously mentioned case of self-reporting the likelihood that someone will vote in a poll, a cut-off value of '4' could be implemented to represent the assumption that anyone answering 4 or 5 is likely a voter and everybody answering 1-3 are not. This would act as an initial selection, so pollsters could examine only the voting behavior of the likely voters-group.

This process was also used with the other general question, where each answer that indicated voting was awarded some points, and then the respondent's answers were summed and compared to another cutoff value. A Cutoff method is very simple and fast to implement, but is not very robust. In an example like this (i.e., an example where a significant portion of respondents are known to lie, or change their mind), many voters will not be counted while non-voters will be.

1.2 Probabilistic method

The probabilistic method involves assigning weights to the self-reported answers of voters based on their likelihood of actually voting for those. Essentially it's a more involved variant of the cutoff method: Instead of counting some people, weighted equally, and counting some not at all, a probabilistic method will assign a probability - a weight - that a particular person will vote; The cutoff method can be understood as a variant of the probabilistic method with uniform weights, 100% above the cutoff and 0% below it, and the probabilistic method uses a sliding scale

*http://justinhgross.com/wp-content/uploads/2019/08/Rentsch-Schaffner-Gross-The-Elusive-Likely-Voter.pdf

in-between the two extremes to assign weights. This method is still fairly easy to implement, but involves a big assumption: That all indicators of voting behavior have the same weight, meaning that all of the questions asked will be equal predictors of voting behavior, even when this is rarely true: Some questions asked may be irrelevant or unrelated to voting behavior. A human could manually apply weights to the question, but this requires a human to make subjective assumptions about their importance, which introduces a significant bias.

To calculate the weights without such a subjective bias, a random forest neural network is introduced. These can be trained on data of the outcomes of previous polls to create a model of the weights based on the outcome of past polls. Generally, pollsters want to train multiple similar random forest neural networks, each trained on a unique data set, and then average the results (with error-bars!). This has a significant advantage over allowing fleshy humans to make the assumptions, because machine learning is much easier to scale up, so pollsters can include demographic information to account for facts, such as younger people generally being less likely to vote than older people.

Of course, the downside of introducing machine learning is the complexity of the method is significantly greater than the simple cut-off method and probabilistic method, including the requirement for computation time, which is more-or-less trivial for the other methods. It also requires a large training data set, and because of the requirement for training data, the neural networks can be slow to adjust to changes in the general public. For example: When predicting polls for the 2016 US presidential election, the neural networks will have been trained on the latest data, from 2012. In those intervening four years, young people have become more politically active, and a neural network of this type is not designed to account for things that are not represented in its training data.

2 Results

The results of evaluating the different approaches to determining 'likely voters' are shown in table 1. The first column shows the percentage of respondents who are likely to vote, separated by method. The deterministic cutoff approaches is a straight forward calculation, and the probabilistic models take the average of the turnout propensity scores. The second and third columns are the observed democratic biases on a national and state level, where a positive score represents a bias towards the Democratic party, and a negative score represents a bias towards the Republican party. The fifth column is the average absolute error by state and the sixth is a measure of the prediction accuracy.

Approach	**Implied turnout(%)**	**National bias(%)**	**Avg. bias by state(%)**	**Avg. absolute error by state(%)**	**Predictive accuracy(A)**
Cutoff approaches:					
Already voted + will definitely vote	70.78	3.59	2.46	4.44	–0.084
Perry-Gallup 6s + 5s	60.26	2.05	1.18	3.98	–0.046
Probabilistic approaches:					
Perry-Gallup	66.55	3.29	1.75	4.17	–0.075
Perry-Gallup + Demographics	59.86	–0.19	–0.36	4.02	0.007

Table 1: The turnout rate among the target population of non-incarcerated US citizens was 59.4 percent. A total of 55.1 percent of 2016 CCES respondents were validated voters.

Among the cutoff methods used, the one relying only on intention to vote fares the worst, whereas the Perry-Gallup approach has the lowest bias and highest accuracy of the cutout methods. When looking at the probabilistic methods, the Perry-Gallup method vastly overestimates the implied turnout, thus not making any improvement on the cutoff methods. Lastly, the Perry-Gallup + Demographics method (the method proposed by the authors in this paper) produces a turnout close to the actual turnout, and has the lowest bias of all the shown methods. So the recommendation of the authors of this paper is that polling companies should use the methods that give more accurate results, which they previously have not done.