Predicting Voter Behavior

CMSC 35300—Mathematical Foundations of Machine Learning

Eujene Yum
eujeneyum@uchicago.edu

Carolyn Liu
crliu@uchicago.edu

Dec 2, 2022

Abstract

The act of voting is pivotal in upholding democratic values, and it is imperative that social science researchers explore factors that deter and promote citizens from going to the polls. In this paper, we aim to predict whether an individual will always vote, sometimes vote, or rarely/never vote. We fit six different machine learning models, all with varying, but promising degrees of success. We find that the random forest algorithm has the highest accuracy of classification, followed by ridge regression which also gives us some insight into possible explanations behind people's voting behaviors.[I]

[I] link to project github: https://github.com/Crliu4/cmsc353_nonvoter_project

1 Introduction

Voting plays a pivotal role in upholding democracy in the United States, but at least 33% of eligible voters do not vote in a given election.[2] Especially with recent events, many people are understandably disillusioned and discouraged. It is important to extract the causes underlying this behavior in order to combat dwindling interest in civic duties and boost voter turnout.

2 Literature Review

A National Bureau of Economic Research paper touches on the driving forces of voting behavior. This study concludes that the act of voting today directly affects future turnout, as a causal channel for explaining turnout persistence. A 1 percentage point decrease in current turnout reduces future turnout by 0.7-0.9 percentage points. This decreased turnout can be attributed to turnout today rather than other factors (persistent changes in voting costs, updating of voter beliefs over the probability of being pivotal, or changes in voters' perceived benefits from election outcomes). It is worth noting that habit formation of voting occurs through an accumulation in the expressive value citizens gain from voting. Although many have accepted the idea that voters get consumption value from the act of voting, the precise form of this consumption value and the way it develops have remained elusive.

A Pew study examines two different probabilistic methods of determining voter likelihood from the 2014 U.S. House elections. The study notes that the benefits of using probabilistic models is that more of the information contained in the survey (all response categories in each question, rather than only one or two) can be used. Logistic regression produces a predicted probability of voting for each survey respondent. One thing to note is that logistic regression assumes that the features (survey answers) are equally relevant for distinguishing voters from nonvoters in a variety of elections. The other probabilistic method highlighted is decision trees/random forest. The advantage of random forest is that it performs well with a large number of features. In the study's analysis, the predicted probabilities for a case are based only on those trees that were built using subsamples where that case was excluded.

3 Data

We found survey data from FiveThirtyEight (polling done by Ipsos) conducted in September 2020. It has 5,836 respondents and their voting behavior (our labels), which can be characterized as always, sporadic, or rarely/never. The survey asks 110 questions targeting the respondent's political

[2] https://www.census.gov/newsroom/press-releases/2021/2020-presidential-election-voting-and-registration-tables-now-available.html

leanings, sentiment toward the efficacy of government, whether they receive benefits from government programs, and the impact that Covid has had on their lives. Our project will use this data to further answer the question of what kind of belief and dispositions affect one's propensity to vote.

Before diving into modeling, we preprocessed our dataset. First, we created binary variables for each of the races present in the race column of the original dataset. We had a couple of categorical columns with string values (income, education, voter status) that were ordered in nature, so we were able to assign numerical values to them. Lastly we converted string valued columns such as gender to numeric and replaced NA's with 0's. We split the dataset into train and test sets to evaluate model performance.

4 Models

In total, we fit 6 different models in an attempt to predict voting behavior in our dataset. Our baseline approaches used truncated SVD, ridge regression, PCA, and kernel ridge regression. To further explore other machine learning methods, we implemented logistic regression and random forest.

4.1 Truncated SVD

Truncated SVD gives us the best rank K approximation of our X matrix. We use the SVD pseudo-inverse:

$$X = V\Sigma_k^+ U^T$$

where Σ_k^+ is computed by inverting the k largest singular values and setting others to zero. k is the regularization parameter and it takes values k = 1,2,...,119 to compute 119 different solutions w_k and return the solution with the smallest training error.

4.2 Ridge Regression

Ridge regression is similar to normal least squares, but adds a regularization parameter, λ. We set our $\hat{\mathbf{w}}$ to satisfy:

$$\hat{\mathbf{w}}_\lambda = \arg\min_w \|y - Xw\|_2 + \lambda\|w\|_2$$

We used the following values for the regularization parameter λ = 0, 0.001, 0.01, 0.1, 0.5, 1, 2, 4, 8, 16, 32, 64.

4.3 Principal Component Analysis (PCA)

PCA is a method that finds lower dimensional approximation of a large X matrix. This increases the interpretability of data while preserving as much variation in the data as needed.

4.4 Kernel Ridge

Kernel regression gives a linear boundary in a high-dimensional feature space. We use the Gaussian Kernel:

$$K(x_i, x_j) = e^{\frac{-\|(x_i - x_j)\|_2^2}{2\sigma^2}}$$

where σ^2 is a tuning parameter. We used the following values for σ^2 = .0001, .001, .01, .1, 0.5, 1, 5, 10, 25, 50, 100.

4.5 Logistic Regression

We implement a multinomial logistic regression using the softmax activation function since we have 3 class labels, as opposed to binary classification, where:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \quad for\ i = 1, 2, \ldots, n$$

for n samples. Our model implementation includes one-hot encoding of the labels, gradient descent to find the optimal weights and bias, and calculating the softmax for each class. After combining all the classes, we obtain a vector of size 3 whose elements sum to 1. To classify the data point, we take the maximum value (probability) of the vector.

4.6 Random Forest

Random forest is a supervised learning algorithm that uses ensemble learning method, allowing for more accurate prediction than a single decision trees. We incorporate bootstrapping, where we randomly sample subsets of a data with replacement. We implemented this model for its high accuracy, but acknowledge that it is difficult to interpret and over-fitting can easily occur. Because our y label classes are quite balanced, we will consider the accuracy score over precision, recall, or F1.

5 Results

While evaluating the results of our six models, we decided that the metric of success would be an accuracy greater than 33%, which is equivalent to randomly guessing someone's voting behavior.

As shown in Appendix Table 1, Random Forest returned the most accurate predictions, which aligns with the general consensus that this model has high predictive performance. Ridge regression was the next best model with an accuracy rate of 59.8% with a λ = 0.001. Truncated SVD and Kernel Ridge had similar outcomes of 58%. The ideal number of columns for Truncated SVD was 117 and the best sigma squared value was σ^2 = 25 . PCA and logistic regression gave the lowest accuracy rate among the models that we implemented.

Table 1: Test Set Accuracy Metrics

Model Type	Accuracy	Best Parameter
$Truncated\ SVD$	58.5%	117 columns
$Ridge\ Regression$	59.8%	$\lambda = 0.001$
PCA	54.2%	4 dimensions
$Kernel\ Ridge$	58.1%	82.3%
$Logistic\ Regression$	53.7%	$\eta = 0.1$
$Random\ Forest$	64.4%	N/A

When running the PCA model, we wanted to come to a reasonable number of dimensions before computing principal components. We plotted the sigma values (see Appendix Figure 1) to see where drop-off point was and decided that 4 dimensions were appropriate to use.

We wanted to confirm that our intuition was correct and plotted the accuracy of both the training and testing sets (see Appendix Figure 2). There seems to be minimal increases in accuracy rates after 20 dimensions. From this, we can gather that there is sufficient variation captured by 20 principal components.

The regularized least square method gave us some features that have higher importance (greatest absolute value of the weights) when predicting voter behavior. Some features to note were whether the survey respondent had any members of their household miss the voter registration deadline or could not obtain necessary assistance to fill out a ballot, their race, and whether they receive long-term disability benefits.

6 Conclusion

In this project, we attempted to predict how engaged a person is with the democratic system using survey data. We implemented six different models, of which the random forest model gave the highest accuracy and garnered some potential factors (e.g. receiving long-term disability benefits, couldn't obtain necessary assistance to fill out ballot) that may impact one's voting behavior through regularized least squares. Potential next steps for future research would include expanding the dataset to include more observations in order to create more robust models.

References

Fujiwara, T., Meng, K., and Vogl, T. (2013). Estimating Habit Formation in Voting. NBER Working Paper 19721. Available at https://www.nber.org/system/files/working$_papers/w19721/w19721.pdf(Accessed\ November 30, 2022)$.

Keeter, Scott, and Ruth Igielnik. "2: Measuring the Likelihood to Vote." Pew Research Center Methods, Pew Research Center, 30 May 2020, https://www.pewresearch.org/methods/2016/01/07/measuring-the-likelihood-to-vote/.

Kreiger, J. R. (2021, December 13). Evaluating a Random Forest model - Analytics Vidhya. Medium. https://medium.com/analytics-vidhya/evaluating-a-random-forest-model-9d165595ad56

Radecic, D. (2022, March 23). Master machine learning: Random forest from scratch with python. Medium. Retrieved November 30, 2022, from https://towardsdatascience.com/master-machine-learning-random-forest-from-scratch-with-python-3efdd51b6d7a

Verma, Suraj. "Softmax Regression in Python: Multi-Class Classification." Medium, Towards Data Science, 2 June 2021, https://towardsdatascience.com/softmax-regression-in-python-multi-class-classification-3cb560d90cb2.

Appendix

Figure 1: PCA Sigmas

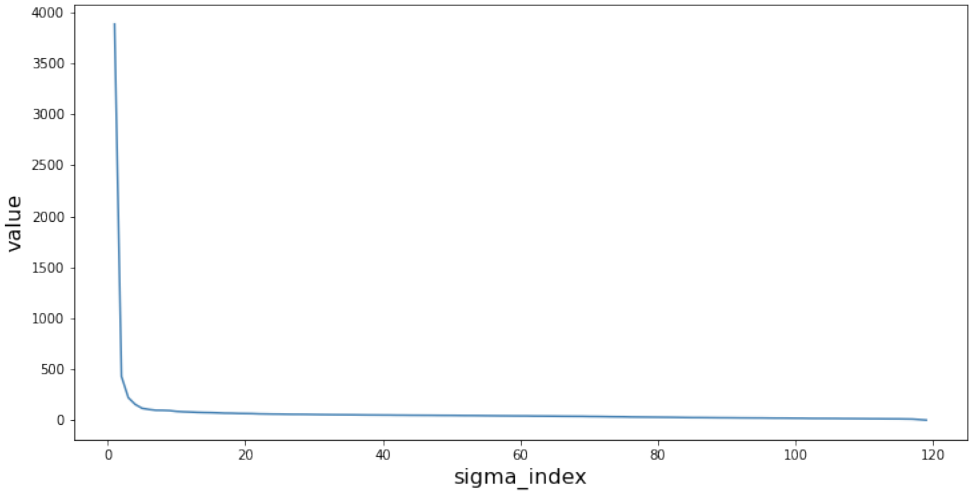

Figure 2: PCA Train Test Accuracy