CSCI 5525 Final Project Report

Meredith Bain & Anna Ton Nu

Introduction

Being able to vote is a critical part of democracy in the US. It allows for citizens to have a voice on governance and public policy but not all eligible voters are taking advantage of this opportunity. In the 2024 presidential election only 64% of eligible voters voted. Encouraging voter turnout is an important step towards securing a win for a candidate so figuring out who to target can be advantageous. As political campaigns develop more technical savvy, an increasingly interesting question has been that of using machine learning techniques to model outcomes of an election on both the aggregate and individual levels. This has an outsized effect on efficiency in distributing campaign resources, as it is critical that a campaign reaches high support/low turnout voters and high turnout/ambivalent support voters before election day.

Overview of the Data

Our dataset is retrieved from the Minnesota Secretary of State, the governing body that has authority over elections in Minnesota. The voter file comes in 8 files (one for each congressional district) of voter registration lists that include one row per voter and encodes geographic and demographic information about each voter. There are also 8 files of vote history lists, which have one row per voter-election combination and indicate whether a voter participated in a given election.

A few patterns tend to dominate the vote history data. Many voters either vote in every even-year election for which they are eligible, and others vote in every even-year election only if it is a year with a presidential contest. A small proportion of voters have a more sporadic and unpredictable voter record. This is an unusually salient pattern in Minnesota relative to other states, given our status as a voter turnout leader in the US. The pattern is magnified by the MN SOS’s regular purges of voters who are inactive, deceased, or moved out of state, so that the denominator of participants from the voter file is expected to be smaller than that in other states.

Preprocessing Steps

All 16 files were uploaded to BigQuery. Selecting the unique election dates and names, we exported to Google Sheets and selected the statewide primary and general elections (township, parks, specials, etc. were excluded, as were the 2020 and 2024 presidential primaries which were held as separate elections from the state primaries in those years). A matchkey was designed that concatenated the election date and election description.

The selected elections were used as a filter on the election history files. We used a query to pivot the data into a table that contains one row per voter, and included their election history as boolean columns based on their participation in primary and general elections over the years.

Literature Review

On the support modeling front, Hare & Kutsuris (2023) sought to use an ensemble of machine learning models to identify which segments of the electorate might be classified as swing voters, by using techniques such as SVM, ANN, and random forests to tease out what they term as "cross-pressures" that cause voters to oscillate between multiple parties. Similarly, Stoetzer et al (2019) leveraged a dynamic Bayesian measurement model based on ideological survey data to attempt to parse out on a macro level the margins by which political parties in the UK might win shares of Parliament. Kim and Zilinsky (2021) performed a similar analysis on the American electorate using regression trees and random forests.

Overall turnout models have also been explored in the literature, particularly as those estimates are needed prior to election day for campaigns to determine a target number of votes they must earn in order to win; similarly, media and research organizations must have an estimated turnout figure in order to call elections prior to all ballots being counted by the election administration. Piotr (2019) shows that ANN is the best technique over random forest and RNN for predicting overall turnout by province in Poland. Fitrani et al (2022) confirmed this result using ANN's to predict participation rates in Indonesian elections, and found that decision trees provided a decent alternative approach.

Ansolabehere et al (2024) improved upon these results by finding that in American elections the simplest model, a ratio of the number of ballots cast in the previous election to the number of voters registered on election day, is the best predictor of overall turnout, beating out much more intricate models such as those based on demographic information or early vote data from the same year. This is largely due to the fact that registration numbers tend to be fairly stable cycle-to-cycle. Kim et al (2020) found similar results, determining via fuzzy forests (an extension of the random forests algorithm that works best in high-dimensional, highly correlated datasets) that demographic data other than age were not as salient of a predictor as voter registration status or political issue importance in determining whether a voter will participate in an election.

The attempt to model individual turnout probabilities is the focus of this paper. Previous studies include Moses & Box-Steffensmeier's (2021) attempt to parse out the features that may correlate with higher turnout propensity, and found via decision trees that a combination of registration status, age, and self-described intensity of particular political issues are the biggest factors influencing individual turnout. This is a slight improvement upon the overall turnout features discussed above, showing that voter registration is still the best factor for determining participation, and adding that age and ideology play a large contributing role on the individual level.

Two student projects were identified as confirming the efficacy of SVM's in determining individual-level turnout propensity. Challenor (2017) and Pollard (2020) both point to SVM's as having the highest Matthews correlation coefficient for predicting turnout on the individual level.

As SVM's and Random Forests tended to show up the most in our survey of the literature and only confirmed by other student projects to be among the most optimal methods, we proceed by running both models on Minnesota voter data and compare the results between the two methods. Since neither piece of literature that included SVM's and Random Forests trained their models on the person-level vote history we have access to from the Minnesota Secretary of State – registration status was the closest proxy – we expected our models to have an edge over the others in the literature.

Objective

We seek to determine which model between SVMs and Random Forests is best suited to predict individual turnout propensity based on individual voter file data including demographics and vote history.

Random Forest

Theoretical Background

Random Forest is a supervised machine learning algorithm that uses a combination of predictions from multiple decision trees to create a final prediction. A decision tree is a flowchart like structure that replicates the decision making process. It is useful for determining which features have high impact on a prediction and data can be split at each node in a way that will lead to the most accurate result. Each decision tree is trained using different subsets of data and features. The averaging of results or voting, where the majority prediction wins, allows for increased accuracy of the final prediction compared to just using one single decision tree.

Implementation

For our Random Forest implementation we used the Random Forest Classifier from the python package scikit-learn. This package was also useful for creating the test/ train splits and creating accuracy reports. For our experiment, we set the classifier to use 10 trees in a forest. In general, more trees can lead to more accurate predictions but it takes more time to train. We trained 4 different models to predict a voter's involvement in the primary election or the general election for 2024.

We knew from the literature review that the most important features are registration status and age, so we were able to focus on vote history and age as features in the model.

We trained all models below with a 70/30 training/testing split.

Results on Minnesota Voter Data

Predicting General Voters based on General Election History

To begin with, we used vote history in general even-year elections from 1994 to 2022 as features to predict participation in the 2024 general election.

Accuracy of the Random Forest model: 1.0				
Classification Report:				
	precision	recall	f1-score	support
0	1	1	1	79,679
1	1	1	1	972,021
accuracy			1	1,051,700
macro avg	1	1	1	1,051,700
weighted avg	1	1	1	1,051,700

When using random forest to predict if a voter is going to vote based on their general election history and age, the model is 100% accurate. Even when age was not included, general election history alone was enough to give an accurate prediction.

Predicting Primary Voters based on Primary Election History

To see if random forest and data works for another type of election, we created a model for the primary election.

Accuracy of the Random Forest model: 1.0				
Classification Report:				
	precision	recall	f1-score	support
0	1	1	1	890,719
1	1	1	1	160,981
accuracy			1	1,051,700
macro avg	1	1	1	1,051,700
weighted avg	1	1	1	1,051,700

Similar to the results of predicting general voters based on general election history, there is a 100% accuracy when predicting if a voter is going to vote in the primary for 2024 based on primary election voting history.

Predicting General Voters based on Primary Election History

Since general and primary elections are related to each other, we wanted to see if general voters could be predicted with primary election history.

Accuracy of the Random Forest model: 0.924233146334506				
Classification Report:				
	precision	recall	f1-score	support
0	.14	0	0	79,679
1	0.92	1	0.96	972,021
accuracy			0.92	1,051,700
macro avg	0.53	0.5	0.48	1,051,700
weighted avg	0.75	0.92	0.89	1,051,700

It gets a little bit more complicated when predicting general election voters based on primary election history. The prediction is only 92% accurate. This model seems to have trouble determining which voters will not participate in the general election based on their primary election history.

Predicting Primary Voters based on General Election History

We did the same thing as the previous but we flipped it to predict primary voters based on general election history.

Accuracy of the Random Forest model: 0.8465226563784971				
Classification Report:				
	precision	recall	f1-score	support
0	0.85	1	0.92	890,719
1	.71	0	0	160,981
accuracy			0.85	1,051,700
macro avg	0.78	0.5	0.46	1,051,700
weighted avg	0.83	0.85	0.78	1,051,700

Predicting primary voters based on general election history was the most challenging for the random forest model. This model had an 84% accuracy. There seems to be trouble in determining if a voter will participate in the primary based on the general election results.

Support Vector Machines

Theoretical Background

The Support Vector Machine (SVM) is a method of supervised learning that attempts to classify data by identifying a hyperplane that can separate the classes. The optimal hyperplane is found by controlling the margin by which classes are separated and maximizing the perpendicular distance between data points and a valid separating hyperplane; larger margins result in more certain classifications.

Implementation

We used the Python package scikit-learn to process data, train, and test our SVM.

Since we expected our vote history data to be roughly linearly separable, given that we are only seeking class 1 “voted” and class 0 “did not vote,” we chose to use a linear kernel to define the boundary of separation. A 12-hour stint of training on a radial basis kernel yielded no results, so we are comfortable leaving further explorations of kernels to a future time. We did not choose to use a soft margin hyperplane and use or perturb the C or ξ values in the default package in our initial exploration of the model, again because we anticipated that the data would be largely separable based on the clear patterns of voting behavior.

Similarly to Random Forest, we focused on vote history and age as features in the model.

We trained all models below with a 70/30 training/testing split.

Results on Minnesota Voter Data

Predicting General Voters based on General Election History

To begin with, we used vote history in general even-year elections from 1994 to 2022 as features to predict participation in the 2024 general election.

Accuracy of the SVM model: 1.0				
Classification Report:				
	precision	recall	f1-score	support
0	1	1	1	79,526
1	1	1	1	972,297
accuracy			1	1,051,823

macro avg	1	1	1	1,051,823
weighted avg	1	1	1	1,051,823

While we expected the data to be relatively linearly separable, and we also expected individual-level vote history to be a stronger predictor of voter behavior than registration status, the linear SVM was shockingly reliable at predicting which voters would participate in an upcoming general based on their previous participation in historical generals.

These results improve upon the individual-level voter predictions seen in the literature on SVM's as individual-level prediction machines using registration status and age as primary features, though these results are only focused on Minnesotan voters. This could indicate the superiority of using vote history rather than simple registration status as a feature of predicting future behavior.

Predicting Primary Voters based on Primary Election History

To see if our results from the general predictions hold up in other contests, we turned our attention to even-year statewide primaries, using those from 1994 to 2022 to predict participation in the 2024 primary. IT is important to note that this is the statewide primary, held in August, which is separate from the Super Tuesday presidential primary contest, which is held in March. Presidential primary vote history was not considered in this exploration.

Accuracy of the SVM model: 1.0				
Classification Report:				
	precision	recall	f1-score	support
0	1	1	1	890,269
1	1	1	1	161,554
accuracy			1	1,051,823
macro avg	1	1	1	1,051,823
weighted avg	1	1	1	1,051,823

Surprisingly, we got the same results for primary contests. While we were concerned that there may be something wrong with the model (perfect results feel too good to be true), we believe that the results below prove that our model was not ruined by including training labels in the test results, but rather vote history is a very reliable feature to use for prediction in Minnesota.

Predicting General Voters based on Primary Election History

To test our model further, we trained an SVM to predict general election turnout using only primary contests.

Accuracy of the SVM model: 0.9243922218852412				
Classification Report:				
	precision	recall	f1-score	support
0	0	0	0	79,526
1	0.92	1	0.96	972,297
accuracy			0.92	1,051,823
macro avg	0.46	0.5	0.48	1,051,823
weighted avg	0.85	0.92	0.89	1,051,823

Here, we can see that the results are still highly reliable at an overall accuracy of 92%, though not perfect. Clearly, it is easy for the model to identify a general voter from a consistent record of primary participation. However, the model finds it impossible to identify a voter who would abstain from the general based on a record of abstention from the primary.

Predicting Primary Voters based on General Election History

This was an experiment to see how well the SVM could predict voter turnout in the 2024 primary based only on the vote history from generals up to 2022.

Accuracy of the SVM model: 0.8464057165511688				
Classification Report:				
	precision	recall	f1-score	support
0	0.85	1	0.92	890,269
1	0	0	0	161,554
accuracy			0.85	1,051,823
macro avg	0.42	0.5	0.46	1,051,823
weighted avg	0.72	0.85	0.78	1,051,823

Here, the SVM shows that it is capable of identifying a non-voter with an accuracy of about 85%, but is completely incapable of correctly identifying any voters who actually voted. In addition to the explanation above of the profile of voters who may always participate in one type of contest and never the other, this can be explained by the fact that, using only 30% of the data to test, a low-turnout primary is unlikely to generate many samples of active voters. Nonvoters in general would be particularly unlikely to show up to a primary, and a low-turnout primary would not capture many of the sporadic or even high-turnout general election regulars.

Model Comparison and Analysis

When comparing the accuracy results of the 4 models of Random Forest and the 4 models of SVM, we can see that the accuracy is practically the same.

For the “Predicting General Voters based on General Election History” models, both have an accuracy of 100%. Several confounding factors may have contributed to this result. Firstly, the Minnesota Secretary of State is generally fairly ruthless at purging inactive voters from the rolls. This is largely due to the fact that Minnesota offers same-day voter registration, so it is fairly low stakes to remove an inactive voter from the rolls. Since we needed access to the 2024 vote history in order to label our model, we only received the latest post-election purged file. If we used these results to predict a future election, say the 2026 midterms, we would certainly see a different result, as there would be new voters on that file that had not yet registered in Minnesota in 2024, and some of the voters in the 2024 result would no longer have participation recorded in 2026, as they would have moved states or deceased. Secondly, Minnesota also has an atypically high level of voter turnout relative to other states, which contributes to the overwhelming trend of consistent even-year voters as well as consistent presidential surge voters. Additionally, the voter rolls are predominantly composed of older voters, who tend to have long vote histories behind them.

For the “Predicting Primary Voters based on Primary Election History” models, both have an accuracy of 100% also. We think that this makes sense for the same reasons stated for predicting general voters based on the general election history case. The clean voting records and Minnesota’s high and consistent voting turnout makes it easy for voter turnout prediction.

For the “Predicting General Voters based on Primary Election History” models, both random forest and SVM have an accuracy of 92% which is lower than the 100% accuracy when using the general election history to predict a general election voter. It appears like both SVM and random forest struggled with identifying voters who decide not to vote in the general election based on their choice to vote in the primary election. This makes sense as one might expect a high conversion rate between the subset of voters who participate in the primary to those who participate in the general. However, many voters in our highly partisan state will always skip a primary and never skip a general, making it nearly impossible to identify which of those absentions are confined to the primary and which extend to the general.

For the “Predicting Primary Voters based on General Election History” models, both random forest and SVM have an accuracy of 84% which is lower than the 100% accuracy when using the primary election history to predict a primary election voter. Both SVM and random forest struggled with predicting the voters that would vote in the primary based on general election history. This is likely due to the fact that most people vote in the general election but do not care as much to vote in the primary election due to the voter caring more about voting for the party they identify most with rather than a specific candidate.

Since the accuracy for SVM and random forest models are identical, we can not say that one is better than the other. Our results show that voting history and age are sufficient in predicting voter participation but there are other features of voters that will need to be further looked into such as voter behaviors and possibly other voter demographics other than just age. Once these other features are identified, further comparisons on SVM and random forest can be done.

Conclusions

Overall, the machine learning models we have selected based on our review of the literature suggest that individual level vote history is the most compelling feature available for predicting individual turnout probabilities. These predictions convert particularly well among the same type of election (primary history to predict future primary behavior, and general history to predict future general behavior). Predicting who will participate in a primary based on general history or vice versa can yield generally reliable results for identifying a subset of voters who will vote, but it is impossible to use those features to decide who will not vote. At this stage in our project, further research must be done on features that can assist in determining when voters do not vote. With the current dataset we have now, both models succeed and fail in the same areas. There is no clear answer on which model is best but we can see that either model can be used to get a fairly accurate prediction.

Further Questions

It would be an interesting extension of this project to apply the question to a multilayer perceptron model and see if that model comes out with a higher accuracy of prediction in a future election.

Furthermore, it would be interesting to hold on to this model and compare it against election results that are currently unseen in our current voter file. As mentioned above, the voter file purges in Minnesota contribute to a higher proportion of valid predictions than might be expected if we had access to multiple versions of the voter file over time, to account for the flux of moving, inactive, and deceased voters over time.

Furthermore, it would be interesting to see this work done to predict more niche elections. For example, the decision we made in the preprocessing step to eliminate niche contests such as off-year school board, municipal, or special elections would make this a poor model for predicting turnout in anything other than an even-year primary or general contest. Especially given that a non-statewide election would reveal patterns of voters who have moved around the state, it would be interesting to see if a change in geography perturbs the modeled turnout propensity of a more local contest.

Citations

Ansolabehere, S., Brown, J., Khanna, K., Phillips, C., & Stewart III, C. (2024). Forecasting Turnout. Harvard Data Science Review, 6(4). https://doi.org/10.1162/99608f92.62881547

A. S. Fitrani, N. E. Pratama, A. B. Raharjo, Y. Purwananto and D. Purwitasari, "A Comparative Study on Machine Learning based Prediction Models for Public Participation Rate in an Election Voting," 2022 3rd International Conference on Electrical Engineering and Informatics (ICon EEI), Pekanbaru, Indonesia, 2022, pp. 86-91, doi: 10.1109/IConEEI55709.2022.9972283.

Challenor, Tynan (2017). Predicting Votes from Census Data. Stanford University. Retrieved from https://cs229.stanford.edu/proj2017/final-reports/5232542.pdf

Hare, C., & Kutsuris, M. (2023). Measuring Swing Voters with a Supervised Machine Learning Ensemble. Political Analysis, 31(4), 537–553. doi:10.1017/pan.2022.24

Kim, S.-y.S., Alvarez, R.M. and Ramirez, C.M. (2020), Who Voted in 2016? Using Fuzzy Forests to Understand Voter Turnout. Social Science Quarterly, 101: 978-988. https://doi.org/10.1111/ssqu.12777

Kim, Seo-young Silvia, and Jan Zilinsky. 2021. “The Divided (But Not More Predictable) Electorate: A Machine Learning Analysis of Voting in American Presidential Elections.” APSA Preprints. doi: 10.33774/apsa-2021-45w3m-v2. This content is a preprint and has not been peer-reviewed.

Michalak, Piotr. "Application of artificial neural networks in predicting voter turnout based on the analysis of demographic data" Polish Cartographical Review, vol. 51, no. 3, Sciendo, 2019, pp. 109-116. https://doi.org/10.2478/pcr-2019-0010

Moses, L., & Box-Steffensmeier, J. M. (2021, November 11). Considerations for machine learning use in political research with application to voter turnout. Society for Political Methodology. https://polmeth.org/publications/considerations-machine-learning-use-political-research-application-voter

Pollard, Rebecca D, Pollard, Sara M, Streitc, Scott E. "Predicting Propensity to Vote with Machine Learning". 2020. Retrieved from https://arxiv.org/pdf/2102.01535.

Stoetzer, L. F., Neunhoeffer, M., Gschwend, T., Munzert, S., & Sternberg, S. (2019). Forecasting Elections in Multiparty Systems: A Bayesian Approach Combining Polls and Fundamentals. Political Analysis, 27(2), 255–262. doi:10.1017/pan.2018.49

Appendix

Our code and data can be found at https://github.com/Meredith-Bain/csci-5525/tree/main/CSCI5525_project.