PREDICTING VOTER LIKELIHOOD FOR FUTURE ELECTIONS

Vanessa Barlow
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA 22903
vmb3eb@virginia.edu

Morgan Freiberg
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA 22903
msf6wk@virginia.edu

Gracie Wright
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA 22903
gmw8hn@virginia.edu

April 26, 2020

1 Abstract

The purpose of this project is to predict the percentage of people who will vote for each major political party with the intent to identify counties in Virginia with close election races in order to increase voter participation. Using voting data from the general elections between 2000 and 2016 (5 elections), different supervised machine learning techniques were implemented in order to predict the percentage of Democratic and Republican voters in the 2016 general election. Our results have found that Random Forest Regression performs best in predicting the percentage of voters. Using this model, we formulated a prediction of the percentage of voting for the 2020 general election that we hope can be utilized to increase voter registration at the county level.

2 Introduction

Voter participation is an important aspect in a democracy. It has been acknowledged that the US has lower voter turnout than most democracies around the world. Initiatives to increase voter turnout by targeting college campuses, canvassing, and contacting local residents have been established but there is still improvement to be made [3].

A study on estimating the likeliness of voter participation was done in 2016. The study entailed analyzing past voting behaviors to generate a voter turnout score that combines voter attributes to create a scale that is used to classify individuals as a voter or non voter on election day. The score indicates the likeliness of voter turnout. For example, a score of 0.4 means that a person has a 40% chance of voting. The study implemented different methods when making their model, such as logistic regression, decision trees, and random forest. When tested against the 2014 electorate, all of the different methods made very close predictions of the likelihood of voting in the election with respect to age, gender, and race [2].

Our project's aim is to increase voter turnout in Virginia by predicting the percentage of Democratic and Republican votes in Virginia for the upcoming election using demographic data as well as data from past elections. With these predictions, we intend to identify the distribution of party lines by county in Virginia and use the data to increase voter participation, specifically in counties with close elections. Resources can then be allocated to these identified counties to increase voting registration and participation in elections throughout both political parties.

3 Method

The data was obtained and compiled using the IPUMS CPS tool [1] [4]. It details the number of votes by county for each candidate in the five general elections from 2000-2016 as well as demographic and geographic data. The percentage of voters for each party were calculated using feature engineering. The data set was split into a training and testing set, in which the features were defined as the three elections prior to the election being predicted (i.e., the training set uses the percentage of voters from the 2000, 2004, and 2008 general election to predict the voting distribution in 2012, and the testing set uses 2004, 2008, and 2012 data to predict the percentage of voters in 2016.)

Since our data is labeled, only supervised learning techniques were implemented. More specifically, since the data is continuous, we implemented various regression techniques to predict the percentage of people voting for each party. Initially, three different machine learning techniques were implemented: Linear Regression, Random Forest Linear Regression, and Support Vector Regression with a Gaussian kernel. An XGBRegressor, Elastic Net, SGDRegressor, and an Elastic Net CV model were built to explore different regression models. XGBRegressor was implemented as Random Forest performed well and we thought the additional boosting from XGB could potentially improve performance. SGDRegressor was implemented to observe if stochastic gradient descent improves our predictions. Finally, we tried two types of Elastic Net models to use a model with regularization that fits well with our data set in order to ensure our model was not over-fitting.

Each regression implemented was tuned using a grid search to minimize testing error. Additionally, a deep learning approach was constructed with 9 layers and an activation function of ReLu in the hidden layers. The Keras Regressor class was used to implement the neural network. Each implementation was fitted on the training data and predicted on the training and testing data to ensure that the model did not overfit. The root mean square error (RMSE) was then calculated for each prediction to measure how well the error in prediction is minimized.

4 Experiments

Predicting both political parties requires the implementation of two models: one to predict the Democratic percentage of voters and one to predict the Republican percentage of voters. We implemented nine baseline models on both the Democratic and Republican data: XGB Regressor, Random Forest Regression, Elastic Net, Elastic Net CV, Stochastic Gradient Descent Regression (SGD), Support Vector Regression, Deep Learning model with Keras Regression. The following table describes the results of our models on both sets of data in order to find the optimal model for each.

Table 1: Model Implementation Results for Democratic Data

Model	Train RMSE	Test RMSE
XGB	0.000879	0.136032
Random Forest	0.012699	0.134908
Linear	0.015694	0.135207
Elastic Net	0.098414	0.150286
Elastic Net CV	0.127708	0.163889
SGD	0.162867	0.179615
SVR	0.121413	0.166933
Deep Learning	0.087899	0.153241

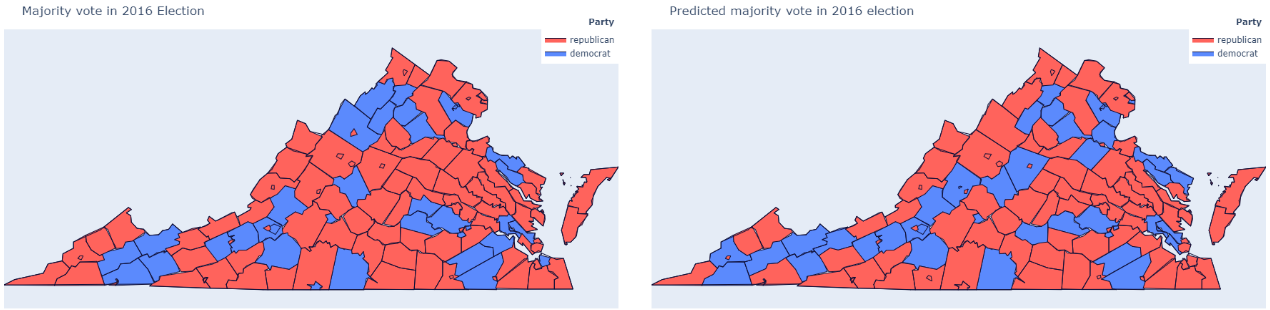
Table 2: Model Implementation Results for Republican Data

Model	Train RMSE	Test RMSE
XGB	0.000868	0.134854
Random Forest	0.011638	0.133398
Linear	0.015956	0.139634
Elastic Net	0.097009	0.146458
Elastic Net CV	0.126036	0.162784
SGD	0.181207	0.207846
SVR	0.119102	0.160696
Deep Learning	0.072186	0.144998

5 Results

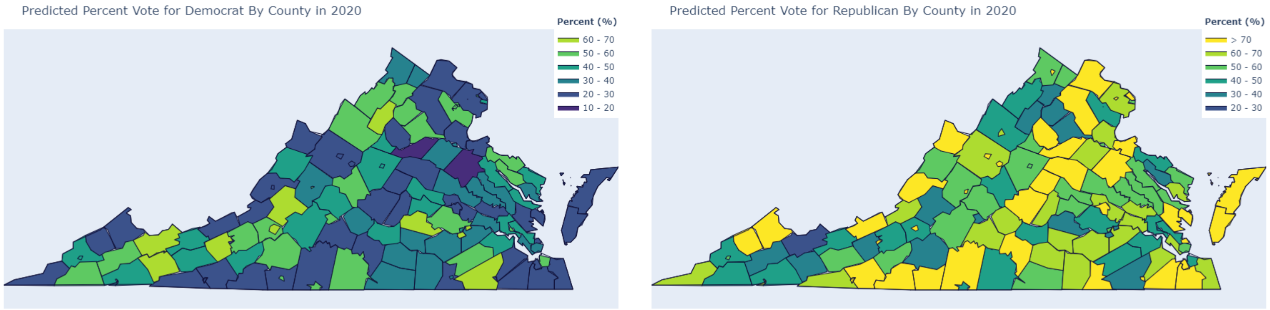
Experimentation yielded Random Forests as the most suitable model for predicting voting results with our data set, as chosen through RMSE comparison. Another important metric is the accuracy in this problem's tangential two-class classification problem. By viewing the party that was predicted to have a higher percentage of the votes as the majority party, the results of predictions can be transformed into two class classification: majority Republican or majority Democrat. By using the outputs of our Random Forest Regressor we achieved a 100% accuracy on the training set and an 84.96% accuracy on the testing set as a two-class classification problem.

Figure 1: Predicted and Actual Majority Vote in 2016


Majority vote in 2016 Election
Party
republican
democrat
Predicted majority vote in 2016 election
Party
republican
democrat

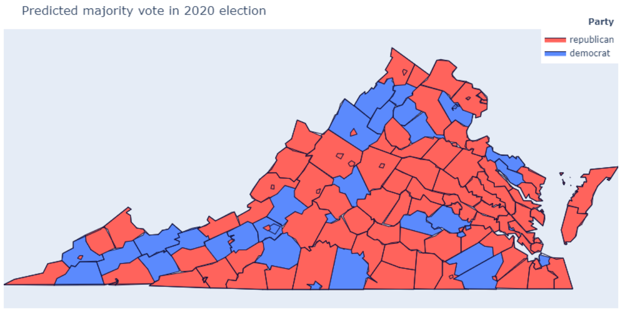
This same model was then fit on the test set, in order to predict the upcoming 2020 election. The following results are the products of this process:

Figure 2: Predicted Democrat and Republican Percent Vote in 2020


Predicted Percent Vote for Democrat By County in 2020
Percent (%)
60 - 70
50 - 60
40 - 50
30 - 40
20 - 30
10 - 20
Predicted Percent Vote for Republican By County in 2020
Percent (%)
> 70
60 - 70
50 - 60
40 - 50
30 - 40
20 - 30

These results were ultimately aggregated to predict the winning party within each county:

Figure 3: Predicted Majority Party in 2020



This model achieved training RMSEs of 0.048268 for the Democratic Party and 0.046872 for the Republican Party and a 95.489% accuracy on the training data. This loss in training accuracy (in comparison to predicting the 2016 election) is likely the loss of overfitting that may have happened in the tuning of the parameters. Obviously these results cannot be validated until the election has happened but we look forward to seeing how our predictions stand up!

The link to all the work can be found here:
`https://colab.research.google.com/drive/1u3i6btbkaG0vsrJC39mZ-mjQKp_h6Ojk`.

6 Conclusion

We hope that our model for predicting the 2020 general election can be utilized by political parties to determine which counties in Virginia they should focus their campaigning efforts. Looking at Figure 3, it is evident that the majority of the counties are expected to have a Republican majority in the 2020 election. Both parties can use this information to find potential swing counties, in which increased voter participation would be crucial to determine the election outcome, and can allot their resources and funds to these respective areas. Also, in some cases, the party that generally has won in past elections is not predicted to obtain the majority of votes. For example, there are a significant number of counties predicted to have a Democratic majority in Southwest Virginia, but this area generally has trended toward Republican in past elections. Likewise, in the Virginia Beach area, there are a significant number of counties expected to have Republican majorities, but these areas have generally voted Democratic in the past. These areas would be crucial to analyze further to determine ways to increase voting.

Some potential shortcomings of our model is the lack of inclusion of third party voting. Our model strictly uses Democratic and Republican data. While the number of people voting for third party candidates is minimal compared to the number of people voting for the major political parties, these voters could cause close elections at county levels, which could contribute to some inaccuracy in our findings. The models could also be improved by incorporating more demographic data such as wealth and family statuses. Furthermore, our final model may slightly over-fit the training data but it is difficult to tell given the sparse data available. More measures or regularization could be put in place to counteract this. The most effective measure against over-fitting may be the reorganization of the data to incorporate more than one data point for each county.

Some possibilities for future work would be to apply our model to new data to predict voter participation for other states. More specifically, our model would be extremely useful for determining voter participation in swing states, which are notoriously known for having close elections. Another useful utilization of our model would to use Virginia data that included third party data. The inclusion of this data would allow us to identify trends in third party voting and its influence on the general election, which could be used by political parties and voter registration organizations.

7 Contribution

The team as a whole was involved in the analysis and evaluation of the models generated. All team members participated in Zoom meetings and discussed approaches that were taken and what the project fully entailed. Specific details regarding what each teammate can be found below.

Morgan: I researched and found the data used. In conjunction with Vanessa I reorganized the data to be effective in a machine learning model and split the data into train and test sets. After this I worked on researching and creating more of our models. Finally, I created the choropleth maps as seen in the report.

Vanessa: I initially made my own copy of the project and did my own visualizations of the data, data cleaning, and transformation of training and testing data. After pre-processing the data, I trained a Random Forest Regressor model that was then evaluated on the training/testing data. After this, I worked on tuning some of the existing models we had, implemented new models, and tuned the new models I added.

Gracie: I compiled the results from our preliminary work, and completed the write-up for the checkpoint. After the checkpoint, I trained and tuned some models, and helped with the final report by working on the experiments and conclusion section.

References

[1] "Bridged-Race Population Estimates, United States July 1st resident population by state, county, age, sex, bridged-race, and Hispanic origin." CDC WONDER Online Database Compiled, 25 June 2019, https://wonder.cdc.gov/bridged-race-v2018.html

[2] “Can Likely U.S. Voter Models Be Improved?” Pew Research Center Methods, Pew Research Center, 30 Dec. 2019, www.pewresearch.org/methods/2016/01/07/measuring-the-likelihood-to-vote/.

[3] FairVote.org. “Voter Turnout.” FairVote, www.fairvote.org/voter_turnout#how_can_we_increase_voter_turnout.

[4] Project Data:
https://docs.google.com/spreadsheets/d/1g1cVAuCU-bV-nuhmsbxSfxDCPkBepNS_Kr4zEAJ1HkE/edit?usp=sharing,
https://www.dropbox.com/s/0qfcxu8uigfv9oo/more_ml_data.csv?dl=0