Predicting the Election: A Voting Machine Learning Model Analysis

Catherine Xu, Jillian Haig, Erin Moulton, Riley Tomek, & Rushvi Patel

Summary:

In the following report, the team built models to predict the outcome of the 2024 election in Virginia as well as presented quantitative information about the precision of the prediction. The first dataset used, voting_VA, provided the group voting data for the presidential election for Virginia from 2000 to 2020. The second dataset, county_adjacencies, is a dataset consisting of neighbors, districts, FIPS county identifiers, and populations of all counties and cities in Virginia in 2022. Prior to choosing the significant variables, the team began the data cleaning process. Variables were cleaned by dropping NA values. The original datasets were merged to create a new dataset, merged_data, which was joined by the FIPS code to align the voting data with demographic information such as income and poverty. Unnecessary columns were then dropped from this dataset. Once the data was concise, the team ran a preliminary exploratory data analysis (EDA) that revealed significant demographics of the county predicted to influence their choice of party, AM8FE038, AM63E033, AM63E051, AM70E007. The team tested out several different models to see which would best predict the winner (winning political party) of the 2024 election in Virginia. These models include use of KNN classification and confusion matrices, ensembles, and random forests.

Data:

The variables chosen, AM8FE038, AM63E033, AM63E051, AM70E007, helped predict county influence to one party or the other as they are based on county demographics. AM8FM038 is a variable representing females 75 years and over with a disability. The next two

variables reveal poverty status in the past 12 months by sex and age. First being, AM63E033, which represents male individuals 5 years or under whose income in the past 12 months was at or above poverty level. The second variable in this section, AM63E051, represents female individuals 15 years old whose income in the past 12 months was at or above poverty level. Lastly, AM70E007, represents individuals 60 to 74 years old whose income in the past 12 months was below poverty level.

Continuing, to get a deeper understanding of the geographical location of voting districts in Virginia, a choropleth map was created:

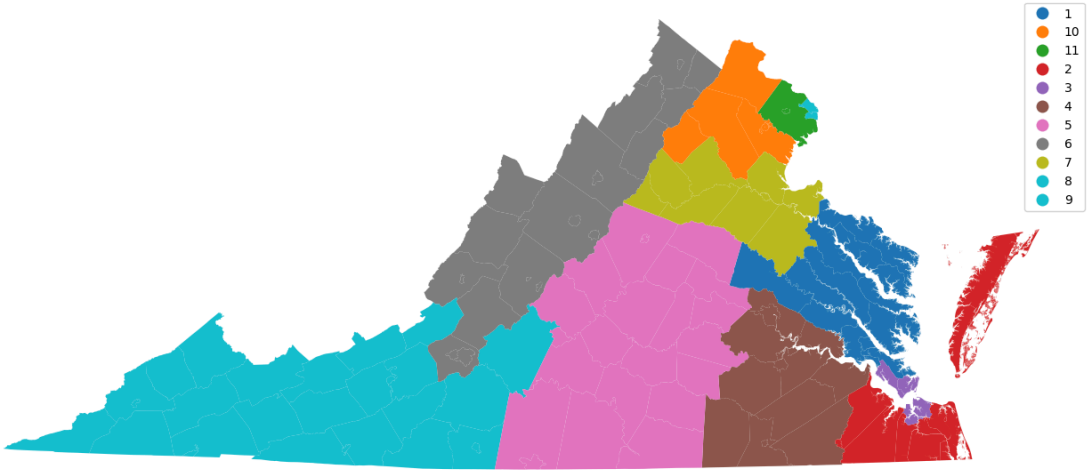

Figure 1: Virginia Voting Districts

To create this map, a shapefile containing geographic features, VirginiaCounty_ClippedToShoreline, was merged with the county_adjacencies dataset. The colors correspond with the Districts column of the merged dataset, providing a visualization of the Virginia voting districts. The next choropleth map created provides visualization of district candidate votes by political party, Democrat or Republican, in the most recent, 2020 presidential election:

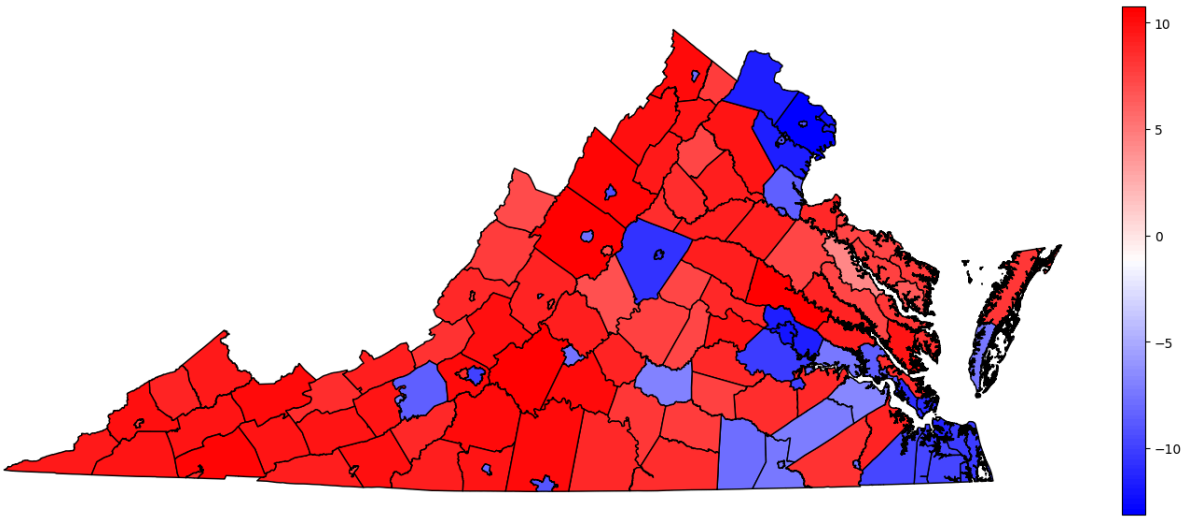

Figure 2: Democratic and Republican Votes from the 2020 Election by Virginia District

To create this map, the voting_VA dataset was filtered to include only the year 2020 and the total number of votes received by Republican and Democrat candidates. These were then further merged with the prior dataset based on the county FIPS codes. From this map, it can be seen that Democrat (blue) support is strong in the North-East and South-East parts of Virginia. Northern Virginia, comprising the suburbs of Washington, D.C., stands out as a densely populated area, which wields considerable influence over the state's total vote count. Although Republican votes (red) seem to dominate much of the state, this draws attention to the importance of considering demographic and geographic factors when assessing the 2024 election results which will be analyzed in this project.

Results:

The group tested out multiple models before deciding on the best one to predict 2024 voting outcomes in Virginia. One of these models was a KNN classification model only using data from the voting_va dataset (the unmerged data). With political party as the response variable to be predicted and year and number of candidate votes as the explanatory variable, a model was

built using the KNeighborsClassifier model. The accuracy (at optimal k) was found to be around 0.58, and a confusion matrix for the model is shown below:

col_0 party	DEMOCRAT	GREEN	LIBERTARIAN	OTHER	REPUBLICAN
DEMOCRAT	223	0	46	44	237
GREEN	7	25	0	43	1
LIBERTARIAN	11	0	89	87	2
OTHER	16	19	62	437	13
REPUBLICAN	137	0	29	49	291

Figure 3: Confusion Matrix for KNN Classification Model

The next model the group looked at was one using an ensemble of decision trees. For this model, the voting_va dataset was again used, but only the following variables were utilized: year, county_fips (FIPS code), party (political party), candidatevotes (number of votes for a specific candidate in a specific county), totalvotes (total number of votes candidate received), and version. After splitting the data into train/test and bootstrapping, the R-squared for the model was computed to be 0.547. Then, using the RandomForestRegressor package, the model was fit and a scatterplot showing actual versus predicted values for political party was created and is pictured below:

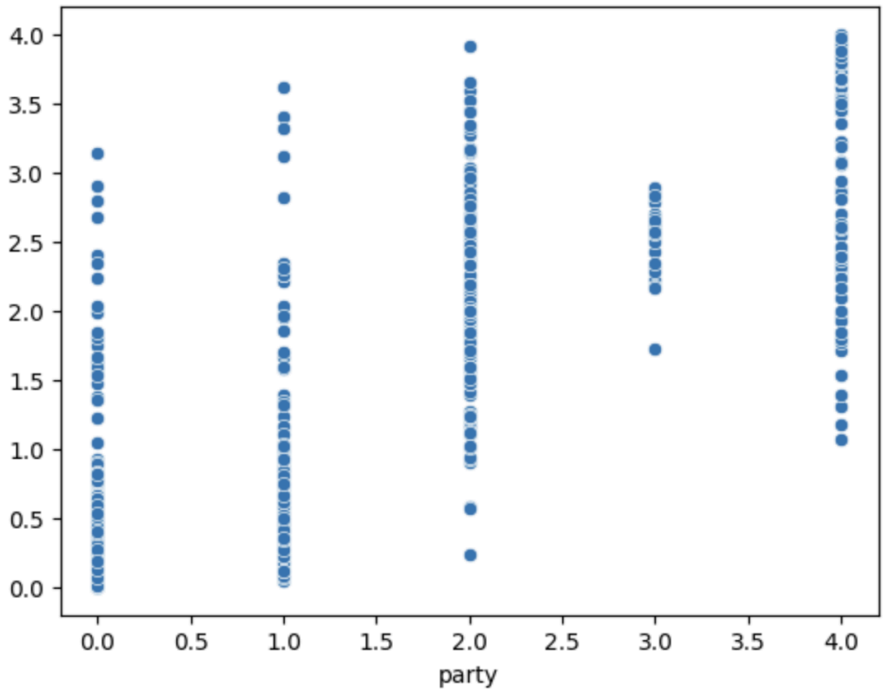

Figure 4: Scatter Plot of Predicted vs Actual Values for Political Party

In this figure, 0.0 represents a republican candidate, 1.0 represents a democrat, 2 represents other, 3.0 represents the green party, and 4.0 represents the libertarian party. As depicted, it seems as if the most data points are concentrated around 2 on both the x and y axes, indicating that the "other" party was the most predicted. However, this model attempts to predict the political party of the winning candidate which is categorical and not numerical, so using the RandomForestRegression package might not be the best approach.

Instead, a similar model–this time using the RandomForestClassifier package–was attempted. Using the same dataset and variables as the previous model, this new model also predicted a candidate belonging to the "other" political party would win, and the accuracy of the model was found to be 0.559. While the accuracy is not terribly low, one could be skeptical of this model by applying it to real life trends in previous elections. It is *not* common that someone from an "other" party wins; historically, either a candidate from either the democratic or

republican party is more likely to win. Because of this, the group decided that this should not be our final model.

The team used the merged data frame to run another Random Forest Regression. The outcome indicated whether the winning party was Republican or Democratic based on its positivity or negativity respectively. The model's ability to capture and predict the net vote counts had a reasonably high degree of accuracy. The Mean Squared Error (MSE) of 8.25 and the R-squared (R2) score of 0.86 suggest that the model explains approximately 86% of the variance in the net vote counts, indicating a good fit to the data.

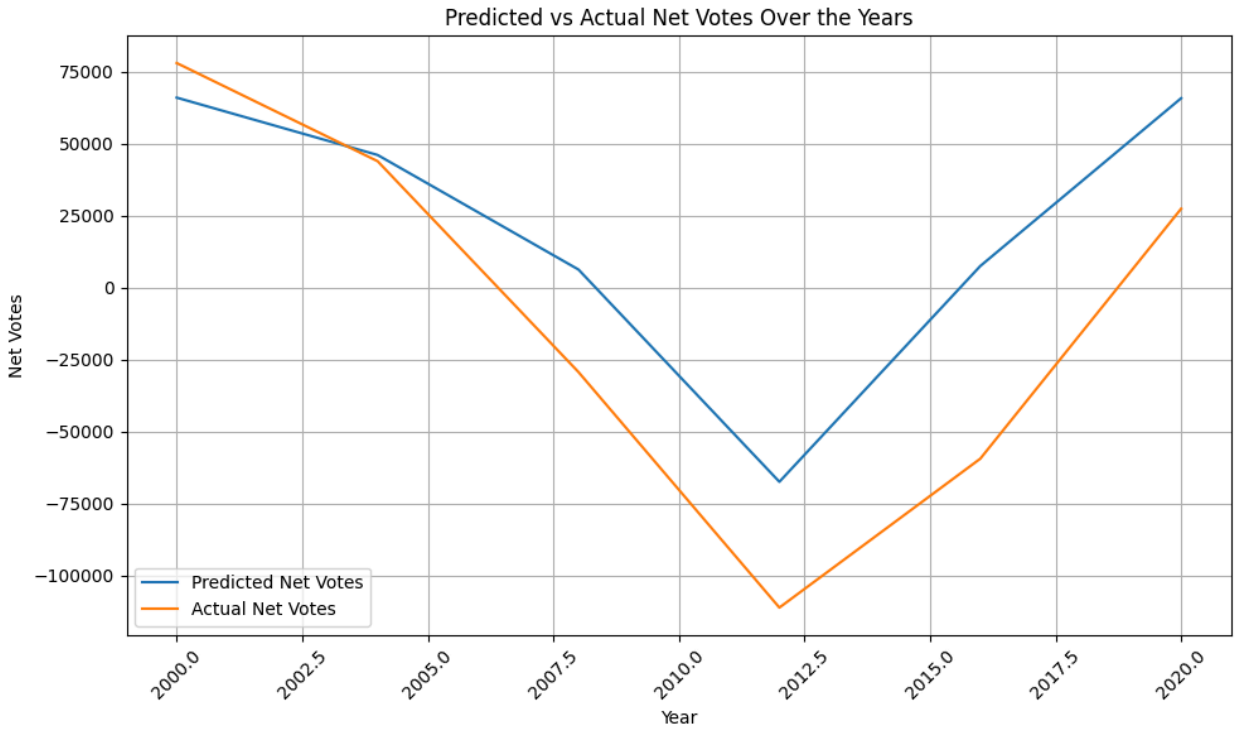

Figure 6: Plot of Predicted vs Actual Net Votes

The plot above shows the predicted and actual net votes over the years. Examining the predictions on a yearly basis reveals interesting trends. In the year 2000, the model predicted a net vote count in favor of the Republican Party, which aligns with the actual outcome where the Republican Party led in net votes. Similarly, in 2004 and 2016, the model predicts Republican

leads that were in line with the actual results. However, in 2008 and 2012, the model predicted Republican leads when the actual outcomes favored the Democratic Party.

Lastly, the study aimed to identify the features of most importance to predicting a boosting model in order to determine what demographics and population factors could point to a role in shifting potential voters towards a party. The previous models utilized the following analysis to isolate Virginia counties which significantly contribute to the populations of the features listed. In order to predict which features of county_adjacenies correspond with which counties vote Republican vs. Democrat, data merged by FIPS code as an inner join matched each county with their respective population demographics. From this merged data, 80% was split as training and 20% was split for testing the model outputs. A varying array of trees were created each with varying levels of depth and learning estimators as an XGBoost classification method. In an attempt to isolate variables within the model which predict more strongly Democrat vs. Republican, 15 defined features were isolated from the rest of the dataset and designated as strong predictors for swaying a county's party choice.

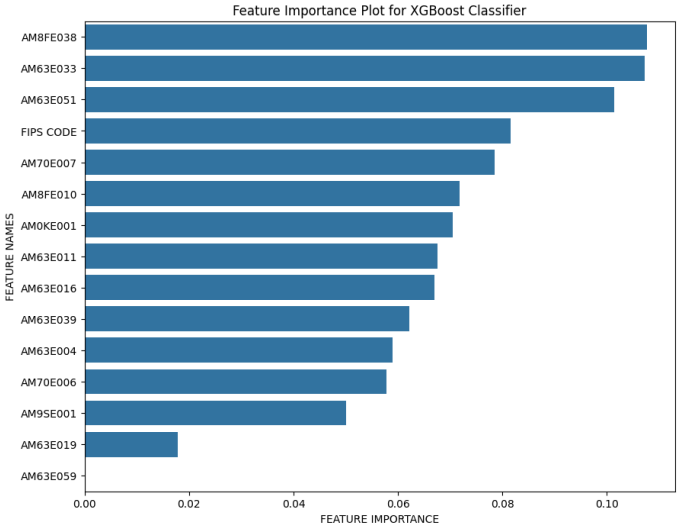

Figure 6: Bar Plot of Distinct Demographic Variables

In the above plot, distinct demographic variables were isolated, with the most notable being 1) females above the age of 75 with a disability and 2) men aged 60-74 years old with income in the past 12 months below the poverty level. This feature importance plot strongly suggests that the model bases its predictions on variables that are impacted by income level, and older individuals above the age of 60 at or below the poverty line determine the state of elections by county. The resulting predictions may not take into account the quantity of this demographic as well as other socio-economic factors at play, but further analysis may reveal insights into which party these features tend to sway towards, as well as the implications and validity of the model on the current elections occurring.

Conclusion:

In this study, we explored various models including KNN classification, random forest classification, random forest regression, and boosting. Based on the results, the group came to the conclusion that the most successful model we tried was the Random Forest Regression model using the merged dataset. After a feature importance plot was created to highlight significant variables, these variables were then used in the merged dataset for this final model; because a wider scope of variables are used, the team believes that this model is most representative of actual election outcomes. Since this model predicted a Republican candidate to win in the years 2000, 2004, 2008, 2016, and 2020, we believe a Republican candidate will likely win the next presidential election in Virginia (based on the historical pattern of this model). Also, the model had an R-squared of nearly 0.86, indicating that this model is a good fit for the data utilized. However, it should be noted that a wide range of additional factors weigh into which candidate will win the election that are not captured in the data we decided to use, such as specific policies

each candidate favors and how voters in Virginia feel about certain topics related to these policies, for example. In the future, newer census data could also be used to further improve these predictions and provide more accurate insight.