# Predictive Modeling of Virginia 2024 Presidential Election

Kirsten Fung, Angela Hong, Britney Hoang

DS 3001: Machine Learning

Professor Johnson

May 3, 2024

**Summary**

The following report outlines an analysis aimed at predicting the outcome of the 2024 presidential election in Virginia. The goal of the project is to develop precise prediction models that provide quantitative insights into the potential electoral landscape. Findings from these models can be used to inform political campaign strategies, allocate resources effectively, and anticipate potential electoral outcomes, thus aiding political parties, candidates, and policymakers in making informed decisions to maximize their electoral success.

The datasets utilized come from voting data from 2000 to 2020 and county level summary statistics from the NHGIS database. The cleaning process began by dropping empty columns and renaming the variables for clarity from the NHGIS data. Three null values were found, however, they were dropped as their removal was believed to have negligible impact on the significance of the results. Data-processing continued by narrowing down to focus on only Virginia counties and grouping relevant variables together. A new dataframe was created to add the necessary variables related to predicting the election such as gender, age, race, and income. To find the majority vote in each Virginia county, a function was created to calculate the most votes in each county - this party was assigned the ‘majority’ in a new column. The cleaned NHGIS and voting datasets were merged together by joining on county names, and the final dataframe was created. The final dataframe contained columns: 'County Name', 'Male', 'Female', 'Median age: Male', 'Median age: Female', 'White alone', 'Black or African American alone', 'American Indian and Alaska Native alone', 'Asian alone', 'Native Hawaiian and Other Pacific Islander alone', 'Hispanic or Latino', 'Other', 'Male Higher Education', 'Female Higher Education', 'Male Highschool Equivalent', 'Male Highschool Equivalent', 'Median Household Income', 'Year' and 'Majority'. Polynomial expansion was performed on these features to allow modeling of non-linear relationships between predictors and outcomes.

In the data analysis phase, predictive models were employed including Decision Trees and Random Forest Regressor. Initial results from the Decision Tree suggested that the outcome of the 2024 presidential election in Virginia will most likely be from the Democratic party. However, these findings exhibited a low accuracy rating, leading us to perform further analysis using the Random Forest Regressor. While this approach showed improved accuracy, there were

concerns of overfitting of the model. In the conclusion, the need for reevaluation and potential adjustments to the methodology were discussed.

**Data**

We utilized the datasets provided by Professor Johnson, voting_VA.csv and 0002_ds249_20205_county_E.csv. We specifically selected county data for its detailed demographic information that can be matched to the voting information. Data cleaning began by importing the CSV files and identifying them as comma-separated format. This required us to read in the file using sep=' ' and encoding='latin-1'. To gain a comprehensive understanding of the dataset, we called the .columns and .shape functions, revealing 3222 rows and 272 columns. We deleted empty columns using the .drop function. Next, we chose to rename the columns for clarity during the modeling phase. To address the abbreviated column naming convention, where a longer version of the name appeared in the row below, we shifted all entries upward by one row and eliminated the original header. Then, we checked for null values. Given the insignificance of the three null entries within the entire dataset, we elected to simply eliminate them, anticipating minimal impact on our analysis. Following these cleaning procedures, we exported our modifications to a new CSV file and uploaded the cleaned dataframe, labeled as "dfclean," to our GitHub repository.

While this new file was cleaned, we needed to trim down the dataframe even further in order to streamline our modeling and analysis. Beginning with the NHGIS dataset, we subsetted the data to only contain information for Virginia since the goal is to predict the outcome of the election in Virginia. Then, we combined variables that we wanted to focus on into more manageable groups. For example, ‘Male: Associate's Degree’, ‘Male: Bachelor’s Degree’, ‘Male: Master’s Degree’, ‘Male: Doctorate Degree’ and ‘Male: Professional school degree’ were all grouped into a new column called ‘Male: Higher Education’. Similar steps were taken for race and female columns. Next, we selected the variables we wanted to focus on in our models. These were: 'State Name', 'County Name', 'Male', 'Female', 'White alone', 'Black or African American alone', 'American Indian and Alaska Native alone', 'Asian alone', 'Native Hawaiian and Other Pacific Islander alone', 'Hispanic or Latino', 'Other', 'Male Higher Education', 'Female Higher Education', 'Male Highschool Equivalent', 'Female Highschool Equivalent', and 'Median household income'.

Moving onto the voting_VA dataset, our objective of this cleaning process was to merge it with the NHGIS data to see how demographic profiles impact party wins in each county across presidential elections. To get the voting results matched with the demographic information, we had to calculate the majority winner in each county and then join both datasets on the county name. We used multiple for loops to iterate through each presidential election year, then county name, then find the party with the maximum votes. The results were added to a new column called 'Majority'. We then dropped duplicates to obtain each county's predominant vote for each election year, categorized by the demographics mentioned above. In preparation for modeling, we also created a 'Majority Bin' column that returns numeric codes correlating to each party (0 for Democrat, 1 for Republican, and 2 for Other). Finally, polynomial expansion was used to capture nonlinear relationships, allowing us to represent interaction between our variables and enhance their predictive capabilities. The final dataset that was used for developing the models is called "merged_expanded".

**Results**

The first model constructed was a Decision Tree. Originally, the model in Figure 1 demonstrates that when the number of females in a particular group is less than or equal to 588.5, the majority party affiliation is Democratic. Otherwise, the party affiliation is Republican. However, when altering the test and train split as well as the max_depth and random_state, the variable that the model uses to predict changes as seen in Figure 2. The new figure shows that if the number of males is less than 588.5, then their party affiliation is Democratic. Also, the values in the leaf nodes return predicted values of exactly zero and one. This seems highly unusual, thus there is likely an issue with overfitting. Possibilities to address this issue are described in the conclusion. Expanding even further, the larger decision tree in Figure 3 tells a different story. This one predicts new variables including age and gender. The RSQ and RMSE values were -0.0144 and 0.4506, respectively. The RSQ was relatively low, while the RMSE was higher than desired. Thus, we found that the decision tree method is very sensitive in that the predictor variables and cutpoints change easily depending on the inputs you provide, which does not capture the data completely and accurately.

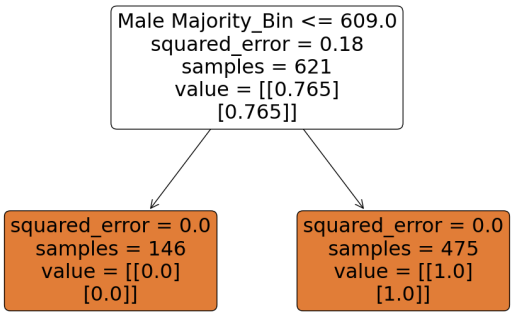



**Figure 1:** Decision Tree for Female

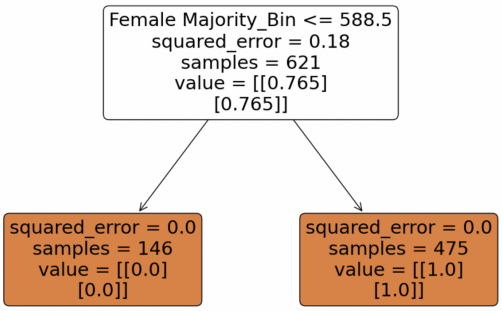



**Figure 2:** Decision Tree for Male

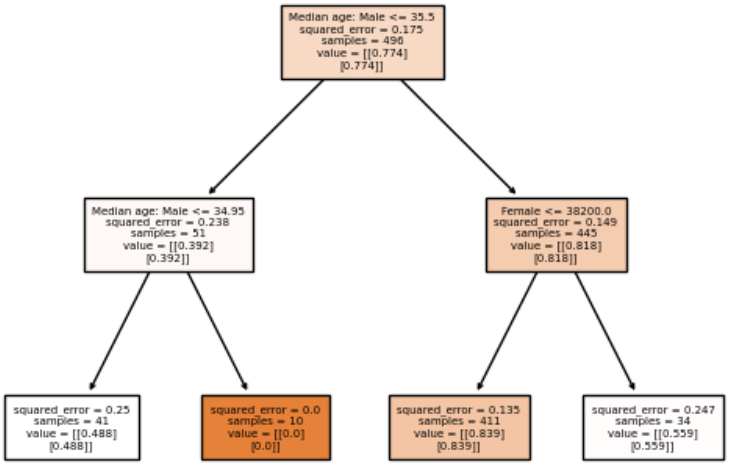



**Figure 3:** Decision Tree for Multiple Variables

The next method employed was a Random Forest Regressor, an algorithm that uses ensemble learning. Random Forests helps combat the issue regarding single decision tree fragility - the predictor becomes more robust and reduces the variance compared to decision trees due to its use of randomization or bootstrapping. Figure 4 displays the Predicted vs Actual values. The points barely deviate from 0 and 1, demonstrating that the predictions are lining up closely to what is wanted. To analyze further, Figure 5 and an RSQ value of 0.99 confirms that the model is very close to the prediction. It was found that the averaging makes a Random Forest better than a single Decision Tree hence improving its accuracy, however this is likely due to overfitting again.

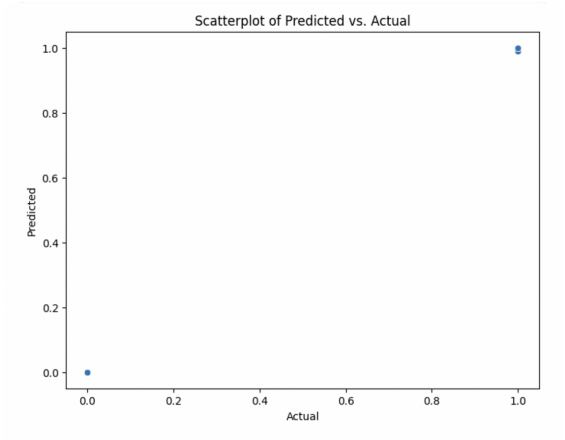



**Figure 4:** Scatter Plot of Predicted vs. Actual

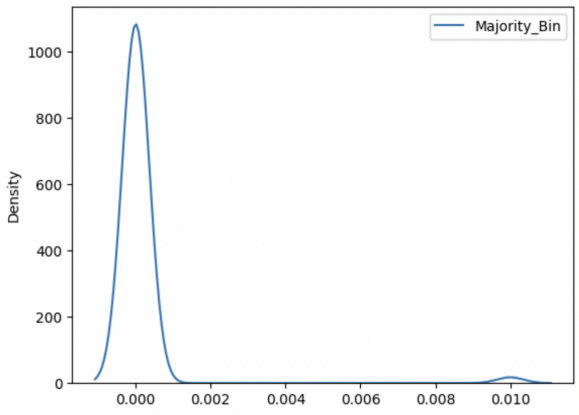



**Figure 5:** Residual Plot

Lastly, a variable importance plot was created as a way of indicating which variables ended up being the most useful for the process of constructing the forest. Figure 6 displays that "Majority Bin" was the most used variable, understandably so as it is the target variable. The other prominent variables are "White alone", "Male", "Female", and "Black or African American".

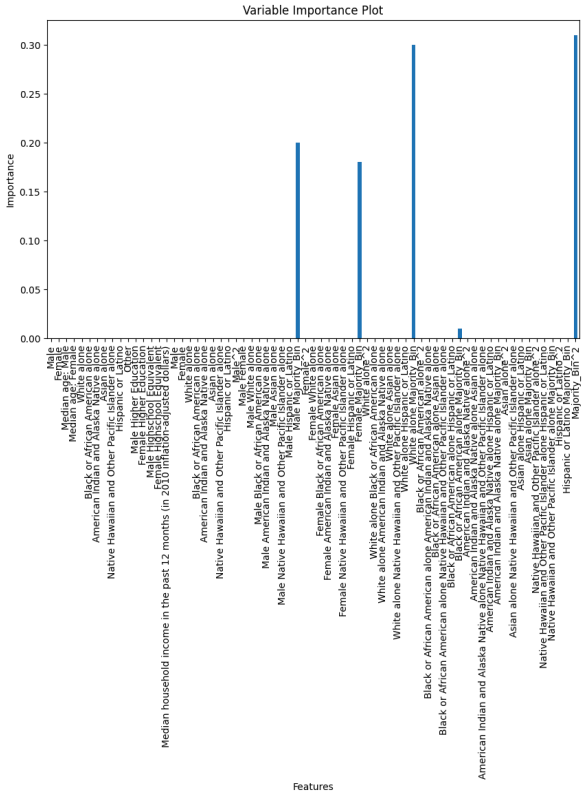



**Figure 6:** Variable Importance Plot

In order to calculate the vote of each county to get the result of the Virginia presidential election, the probabilities from the decision tree were used. Based on the variable importance plot, “Female” was one of the top variables used in predicting the model. This variable was also displayed in the decision tree, and thus chosen to determine the voting outcome. The decision tree shows that 83.9% percent of female individuals would vote democrat and 16.1% of females would vote republican in the state of Virginia. The female vote count was multiplied by these probabilities to obtain the total vote share of democratic and republican females in Virginia. This resulted in 3,346,966 Female Democratic votes and 642,266 Female Republican votes, leading to the conclusion that the outcome in Virginia will be a democratic win. Looking at the results, it is acknowledged that there is a very apparent discrepancy in vote counts between the Female Democratic and Republican votes. This is likely not indicative of a true difference in support levels due to the model's accuracy only being 0.42.

**Conclusion**

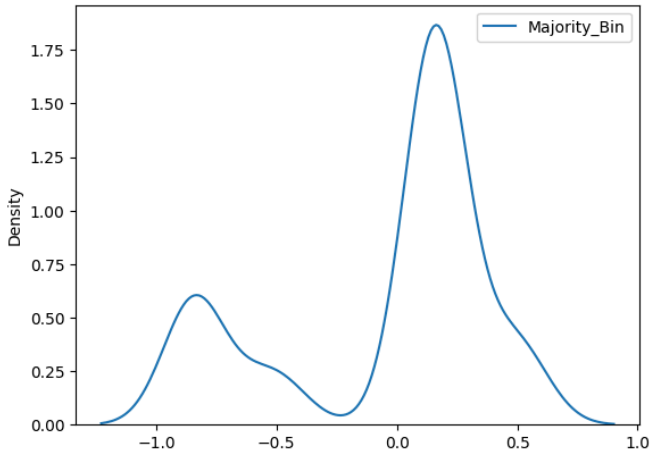
One potential line of criticism towards our project could be the absence of some historical data within our dataset. Due to the misalignment in time frames, some data was not captured, resulting in a loss of information. However, despite this limitation, a significant portion of the data remained available, particularly from recent years, which serves as a more reliable indicator of future voting trends. Another possible line of criticism was the decision to not incorporate age as a demographic variable alongside education, gender, and income. It was collectively decided to prioritize what we deemed to be the most influential factors in our analysis. Looking back, we could have used K Means Clustering to identify county outliers or find the best variables to work with instead of choosing by hand, which would allow us to focus on more accurate predictors of the voting outcome.

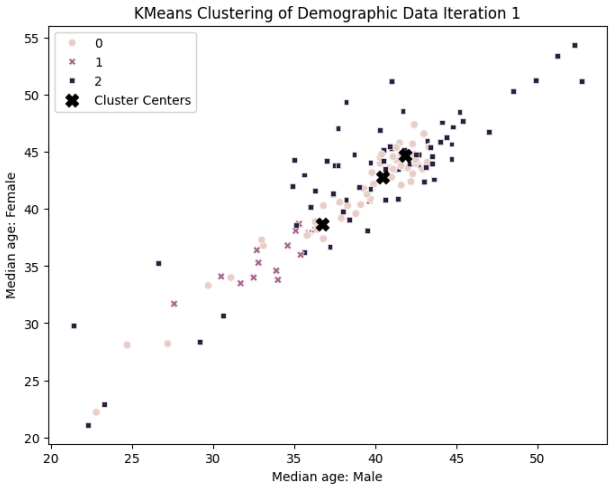
If we could have approached things differently, we would have revisited our data preprocessing methods to mitigate overfitting and guarantee the accurate capture of patterns. This would entail ensuring the appropriate selection, scaling, and encoding of features to prevent the modeling of irrelevant patterns. Additionally, further modeling techniques could have been done to yield more accurate predictions.

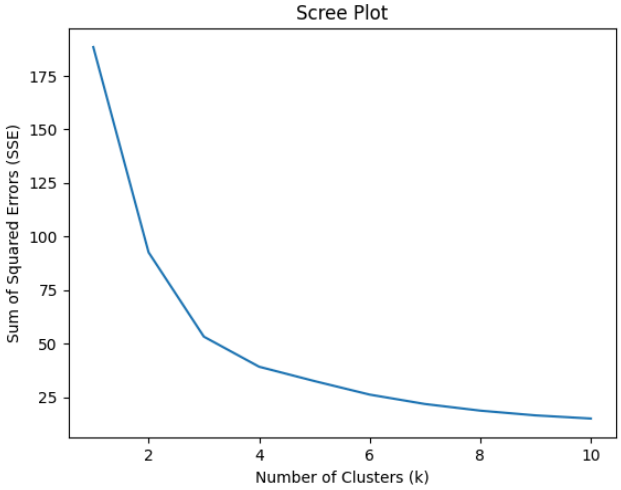
There are a multitude of factors that influence voters that are not information that can be found in the census and are outside the scope of this project. Each county has different lifestyles and priorities which heavily influences the way they vote in elections. For instance, northern Virginia could focus more on education and vote for a candidate that promises to funnel more money in school systems while the South could focus on other policies. So, while race, education, gender, and income, are helpful factors in predicting how individuals will vote, it is certainly not all encompassing.

Throughout the project, we encountered various challenges and made assumptions. For instance, we decided to code each county in Virginia as “0” (Democrat) and “1” (Republican) corresponding to their majority vote. However, when we needed to calculate the probabilities of groups voting democratic and republican of each county through the decision trees, we struggled with how to return to a total vote count when our results were binary. In the future, we would change how we calculated the votes. By coding each country as strictly democrat and republican, it disregards the number of votes from the opposing party. For instance, if we predicted that Fairfax County will most likely vote democrat, there are still republicans that we did not consider at the state level. Thus, in retrospect, we recognize the need to refine our method, particularly in accounting for votes from opposing parties, which were disregarded in our initial approach. Furthermore, it is important to note that the way in which we calculated the binary variable was not the best because it fails to factor in the entire population of voters. The presidential election is not dependent on majorities within counties but rather cumulative electoral votes. Thus, it is not a completely accurate representation to extrapolate our results.

# Appendix


Majority_Bin
Density
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−1.0
−0.5
0.0
0.5
1.0


KMeans Clustering of Demographic Data Iteration 1
0
1
2
Cluster Centers
Median age: Female
Median age: Male
55
50
45
40
35
30
25
20
20
25
30
35
40
45
50


Scree Plot
Sum of Squared Errors (SSE)
Number of Clusters (k)
175
150
125
100
75
50
25
2
4
6
8
10