Project 3 VA Voting: 2024 VA Election Model Prediction

Group 7: Anna Brown, Adaire Burnsed, Tamera Fang, Diana Nguyen, Elle Park, Carol Wu

School of Data Science, University of Virginia

DS 3001: Foundations of Machine Learning

Professor Johnson

April 29, 2024

Abstract

This project aims to build models to predict the outcome of the 2024 presidential elections in Virginia and provide qualitative information on the accuracy of the prediction. Considering that the current state of politics in the United States is very divided and polarized across the spectrum, an accurate predictive model for the 2024 Presidential Election would be very informative and provide insights into Virginia's current political leanings. In order to create our predictive model, we first used the provided start codes to import the map classify package into Python to visualize the data we found using heat and choropleth maps. We used data surveyed from 2000 to 2020 from presidential elections in Virginia and also county-level summary statistics for each county in the United States, labeled as nhgis_county_data (*nhgis*). The county_adjacencies.csv data included neighboring districts or counties.

Our group split into two subgroups: the first group cleaned the data, and the second group developed the visualizations for analysis. Cleaning included handling missing values, converting variable types, mapping categorical variables, creating new columns to organize the dataset, and merging data frames. First, we created a predicted model using linear regression after we cleaned all of the county data in the *nhgis* data frame. Using these predictive results, we then created visualizations to analyze which party is likely to win the 2024 presidential election in the state of Virginia. We also evaluated the accuracy of our predictions by considering how well our model was likely to work. Although we will not know whether we successfully predicted the election outcome after the results are certified in January 2025, our model's precision in predicting the victorious candidate will still be valuable, given the wide range of differing opinions among professionals who work to predict presidential elections each cycle.

Data

After the data cleaning group dropped duplicates from the 'county_adjacencies.csv' data frame, there were 133 rows of different counties. This quantity is consistent with Virginia's 95 counties and 38 independent cities, which are considered counties for political, administrative, and census purposes. The data contained 16 columns which included these key variables: 'County', 'Population2022', 'FIPS', 'District', 'N1', 'N2', 'N3', 'N4', 'N5', 'N6', 'N7', 'N8', 'N9', 'N10', 'N11', 'N12'.

For the training and testing data of the predictive linear regression model, the data cleaning group cleaned *nhgis* datasets of the "2020 American Community Survey: 5-Year Data [2016-2020, Tracts & Larger Areas]" and the "2022 American Community Survey: 1-Year Data" to contain only Virginia State data. The selected *nhgis* dataset variables include citizenship status, poverty status by sex, and income inequality index. Specifically, the column names were mapped *nhgis* code using the code book. The final *nhgis* dataset contains the following variables: 'county', ' U.S. citizen, born in the United States', 'U.S. citizen, born in Puerto Rico or U.S. Island Areas', ' U.S. citizen, born abroad of American parent(s)', 'U.S. citizen by naturalization', 'Not a U.S. citizen', ' Income in the past 12 months below poverty level: Male', 'Income in the past 12 months below poverty level: Female', 'Income in the past 12 months at or above poverty level: Male', 'Income in the past 12 months at or above poverty level: Female', and 'Gini Index'. Variable 'net_vote_count' is defined as the total votes for the Republican party minus the total votes for the Democratic party by county in Virginia in the 2020 Election, and is merged with the NHGIS data by Virginia county FIP into one data frame VAcounty_data. For the optimized performance of the regression model to fit the realized range of value, all explanatory NHGIS variables except 'Gini Index' and the net_vote_count are rescaled and transformed by the inverse

hyperbolic sine (arcsinh) function.

Predicted net vote counts based on cleaned 2022 NHGIS data are transformed back to original levels by hyperbolic sine into the variable 'vote_difference_2024', and are merged with the county adjacency data about neighbors along with variable 'vote_difference_2020' (same as previously defined untransformed net_vote_count) by Virginia county FIP into one data frame for visualizing heatmaps.

When preparing and cleaning 2022 NHGIS data for analysis, we debated what to do with the missing variables. Ultimately, the data cleaning group decided to substitute all predictive results due to missing values in the 2022 *nhgis* dataset with a value of 0, indicating a neutral political affiliation. For the historic votes dataset, we merged variations of candidate names, such as DONALD TRUMP and DONALD J TRUMP. We also checked for unique values several times throughout cleaning the data before combining duplicate or repeated counties in the Virginia county dataset.

Results

To predict the voting results for Election 2024, the analysis group created a linear regression model with the explanatory *nhgis* variables U.S. citizenship/nativity, property status by sex, and income inequality metric. The response variable is 'net_vote_count,' which is the difference between historic votes for the Republican and Democratic parties in 2020 by county in Virginia. All aforementioned variables are numeric, and corresponding kernel density plots are visualized in Appendix A through D, respectively.

Based on kernel density plots for *nhgis* data from 2016-2020, 'U.S. citizen, born in the United States' has a higher central tendency for its peak than other nativity status groups. 'U.S. citizen, born in Puerto Rico or U.S. Island Areas' has a bimodal kernel density: one peak occurs around 0, and the other occurs around 3.75 after arcsinh transformation. For females and males, 'Income in the past 12 months at or above property level' has the same and higher central population frequency than its corresponding group below property level. The Gini index measures the extent to which the distribution of income or consumption among individuals or households within an economy deviates from a perfectly equal distribution. A Gini index of 0 represents perfect equality, while an index of 1 implies perfect inequality. The central frequency for the Gini index from the 2016-2020 American community 5-year survey occurs at 0.45. Net_count_votes after arcsinh transformation shows a bimodal kernel density with a higher peak of around 10 and a relatively lower peak of around -10, indicating more counties in Virginia voted for the Republican party in 2020.

Kernel density plots for *nhgis* data from 2024 show the same distribution with a slightly higher central frequency for all citizenship status groups, indicating a general increase in population within Virginia state. For females and males, 'Income in the past 12 months below

property level' has a higher central frequency. Similarly, 'Income in the past 12 months at or above property level' also has a higher central frequency and a much higher density at the peak, indicating larger income inequalities between counties in Virginia. Correspondingly, the kernel density plot shows a Gini index above 0.5 in 2022, which has a higher density than the one in 2016-2020.

The analysis group split the explanatory variables (X) and response variables (y) into 80% training data and 20% testing data. The Root Mean Squared Error (RMSE) for the training linear regression model is 5.748, with an R^2 score of 1.0. The Root Mean Squared Error (RMSE) for the testing linear regression model is 7.778, with an R^2 score of -0.141. Since the R^2 score is much higher for the training model than the testing model, it indicates a sign of overfitting on the training data, resulting in the model being less generalizable for the chosen test set (Appendix E). The distribution of residuals for the model is approximately normal, indicating that the assumption of the linear model is valid, and so is the model prediction (Appendix F). We predicted the 2024 election results on the cleaned 2022 *nhgis* dataset with the regression model. According to the heatmap of the difference in vote count between the Democratic and Republican parties in 2020, the Republican party had predominantly more votes across counties in Virginia (Appendix G). However, Fairfax, Loudoun, and Prince William counties cast the most votes for the Democratic party compared to the other counties (Appendix H). Based on the predictive results, the heatmap of votes in 2024 shows that more counties will vote for the Democratic party (Appendix I). This phenomenon could be explained by the previously observed increase in the population born outside of the U.S. or not a U.S. citizen and the increased income inequality between counties in 2022. Compared to 2020 voting results, voters from Loudoun and Prince William counties will predominantly vote for the Republican party in 2024. In contrast,

Buchanan, Appomattox, and Culpeper County would primarily vote in favor of the Democratic party (Appendix J), according to our model.

Finally, using bootstrapping, we resampled our data and created hypothetical data sets to better comprehend the variation in our data and improve model performance in the future. We resampled the data, ran a regression for each bootstrap sample, created a sequence of estimates to see how noisy the coefficient is, and constructed confidence intervals (CI) to communicate the uncertainty. We are looking at the 90% CI, corresponding to .05 and .95 quantiles. If 0 was outside our interval, we could confirm that our output is statistically significant and that most VA counties are expected to vote for the Democratic party in 2024; however, based on Figure O, in 90% of the bootstrap samples of the 'vote_difference_2024' variable in the 'VAcounty_data' data frame, the estimated difference was between -0.00238 and 0.02682. Since 0 is in that 90% CI, given the bootstrapping data, we have to reject that the votes in 2024 based on the predictive results will have more counties that vote for the Democratic party.

Conclusion

In this project, we constructed models that offer insightful predictions about the 2024 presidential election in Virginia. Accompanied by visualizations, these models provide a qualitative understanding of the precision of our predictions. In essence, our research indicates a shift in the political landscape, with a projection that more VA counties will lean towards the Democratic party, potentially leading to a Democratic presidential nomination.

Based on our bootstrapping data, the uncertainty of our predictions communicated that 0 is in our -0.00238 and 0.02682 in our 90% CI. Therefore, we reject our prediction that most VA counties will vote for the Democratic party.

Given our results, we can further train our data to develop a better predictive model for the upcoming presidential election based on Virginia residents' likelihood of voting for each party. The faults and strengths of our project also provide insight that can strengthen future research on predicting voting outcomes by identifying the predictive influence of certain variables. Because the outcomes of presidential elections are a significant public interest, this field of research will continue to be the topic of much discussion and debate for many years. Our model may not be the perfect predictor of which candidate will ultimately win the election, but it provides an additional resource for people interested in using it to understand political trends and future outcomes.

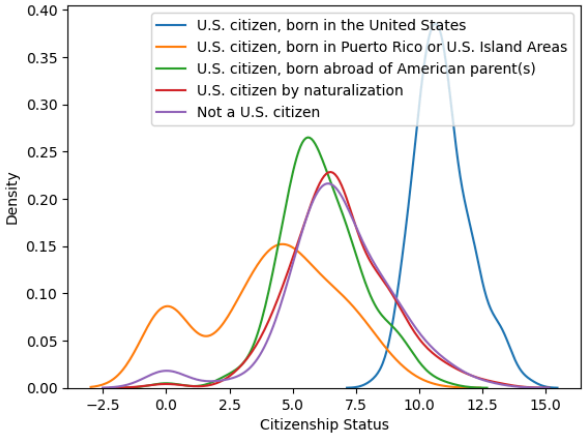
Additional work could include detailed and accurate data collection of variables on education level, racial/ethnicity demographics, and age to see what populations reside in each county/district. With these variables, there is an opportunity to have a more detailed and accurate depiction of how each population in a county/district generally votes for, if they are a population who votes heavily, or if they substantially influence the outcomes of elections. Education level,

cultural, and age differences can affect how a person or population interacts with their environment and form lasting beliefs; therefore, they significantly influence who and what party individuals choose to vote for. These variables would improve generalizability for the entire Virginia population outside of our project and, therefore, have a more accurate representation of Virginia and its county's population to make predictions. Another opportunity for additional work would be to apply the processes and strategies used for this project to other states. Since 49 other states also contribute to the democratic processes in the US, if this model proves successful, it could be scaled to predict election results nationwide, which would have significant implications for future elections.

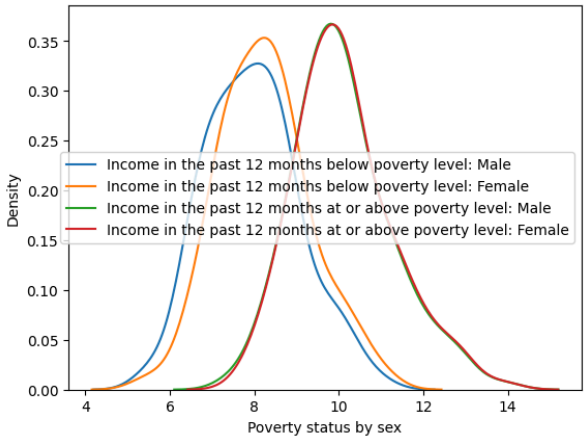
Our project can help analyze how models can accurately predict elections based on variables contributing to their votes. Our project can still contribute to our understanding of the future of our state and country, even if it is only partially accurate. With the presidential election in November, our project's results can be powerful if the actual outcome verifies them. This project is an opportunity to see if machine learning, predictive models, and visualization can predict the critical decisions of the residents of Virginia or even what factors contribute to their decision-making process.

Appendices

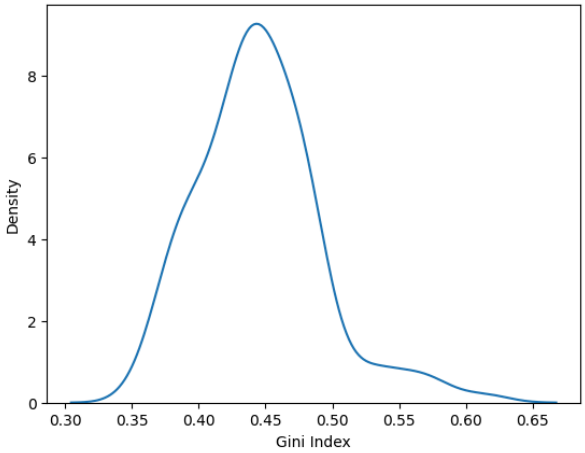
Appendix A: Kernel density plot of the numeric variables for citizenship status in 2020.



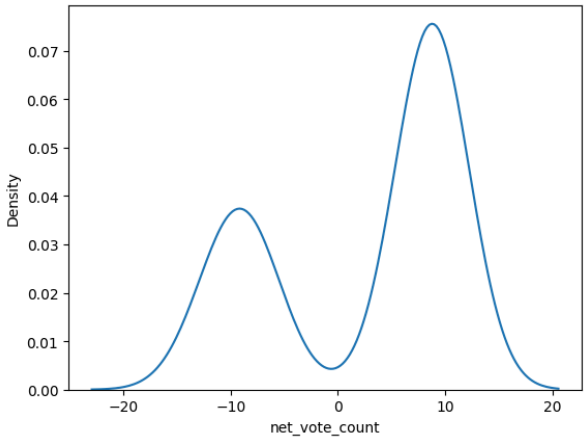
Appendix B: Kernel density plot of the numeric variables for poverty status by sex in 2020.



Appendix C: Kernel density plot of the numeric variable Gini index in 2020.

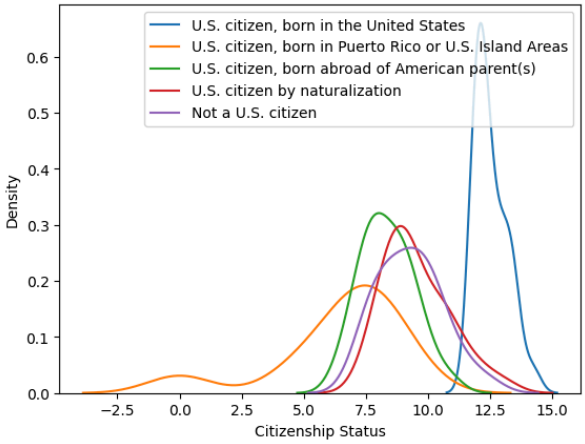


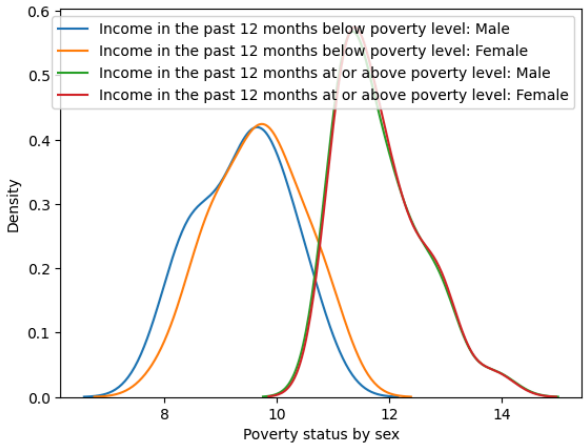
Appendix D: Kernel density plot of the numeric variable Net vote counts (Total votes for Republican - Total votes for Democratic) by county in 2020.



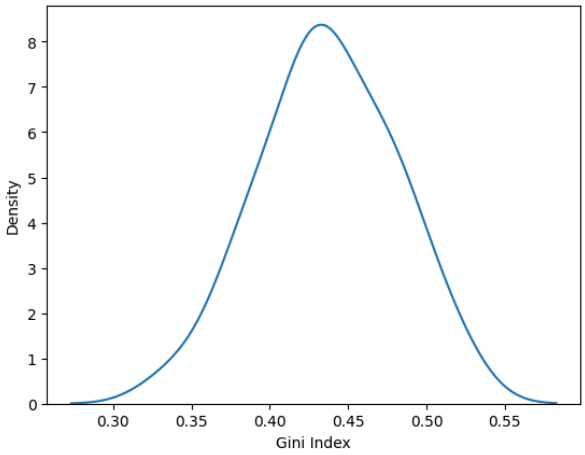
Appendix E: Kernel density plot of the numeric variables for citizenship status in 2024.

Appendix F: Kernel density plot of the numeric variables for poverty status by sex in 2024.

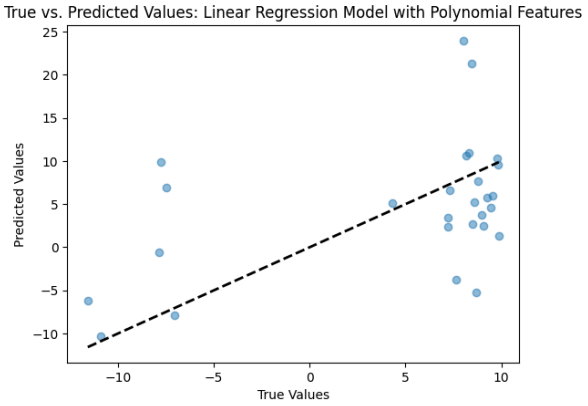




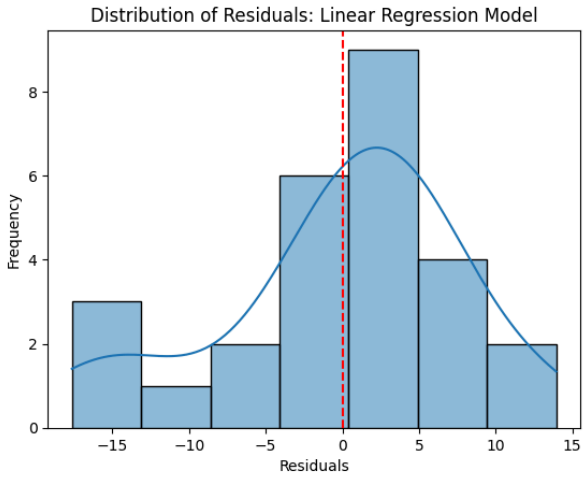
Appendix G: Kernel density plot of the numeric variable Gini index in 2020.



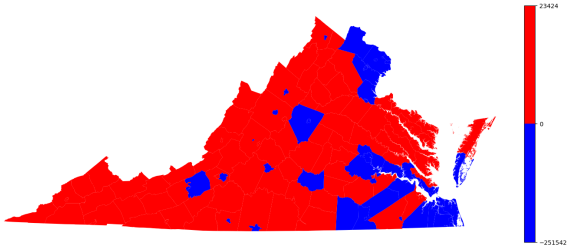
Appendix H: Scatter plot of true voting results versus predicted voting results.



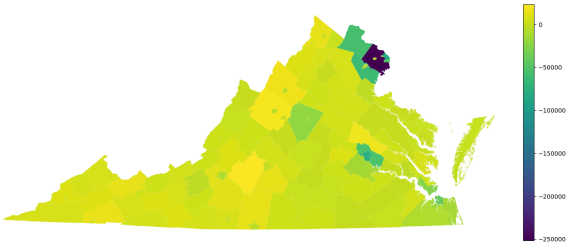
Appendix I: Histogram of distribution of residuals of model for each political party



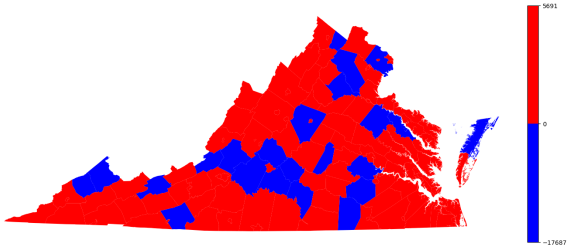
Appendix J: Heatmap of votes for Net vote counts in 2020 by county in Virginia. Blue indicates win for the Democratic party, red indicates win for the Republican party.



Appendix K: Heatmap of votes for Net vote counts in 2020 by county in Virginia



Appendix L: Heatmap of votes for Net vote counts in 2024 by county in Virginia. Blue indicates win for the Democratic party, red indicates win for the Republican party.



Appendix M: Heatmap of votes for Net vote counts in 2024 by county in Virginia

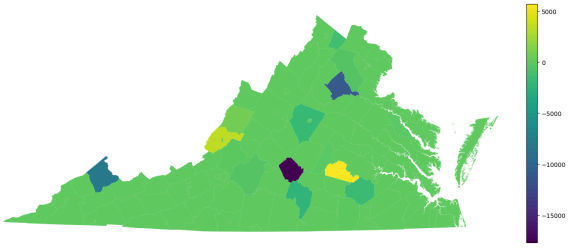


Figure N: Bootstrapping data for ‘vote_difference_2024’ variable that includes confidence interval -0.00238 and 0.026812.

Figure O: Bootstrapping data for ‘vote_difference_2024’ variable that includes confidence interval -0.00238 and 0.026812.

vote_difference_2024
Point Estimate: 0.009751772686017032
CI: [-0.00238053 0.02681546]