Executive Summary

This analysis takes a deep dive into the relationship between demographic and socioeconomic factors and voting outcomes in Virginia counties during presidential election years. By integrating voting data and county-level demographic information, we looked to identify the key predictors influencing the Democratic margin of victory. Using linear regression and LASSO models, we explored how factors such as median income, poverty ratio, median rent, and the proportion of the white population shaped electoral results.

The results demonstrated that the proportion of the white population was the most statistically significant predictor, portraying a strong positive correlation with the Republican margin of victory, Economic indicators, such as higher poverty ratios and median rent, revealed positive associations with the Democratic margin of victory, highlighting the connection between economic conditions and voter preferences. The models performed moderately well, with an R^2 of 0.329, explaining about one third of the variance in the Democratic margin of victory. On the other hand, the root mean squared error (RSME) of ~9,704 votes suggests significant variability that may be attributed to unique country-level differences.

Challenges encountered included merging the datasets and cleaning duplicate vote counts occurring from voting methods, particularly in the 2020 election. These were resolved through a rigorous process, ensuring the accuracy and dependability of the analysis. LASSO regression retained all features, indicating strong predictors with minimal multicollinearity, and provided a slightly improved model interpretability through slight coefficient changes.

These findings offer useful insights towards voter behavior and electoral trends in the state of Virginia where diverse demographic and shifting political landscapes are present. While the results capture significant patterns, it still calls for further analysis into additional predictors

and external factors. This analysis allows for a deeper understanding of electoral outcomes and provides a baseline for future work on certain influences towards presidential elections.

Introduction

Understanding the factors influential to electoral outcomes is essential for analyzing certain political trends and potential implementing policies. This analysis takes a deeper dive into the relationship between demographic and socioeconomic factors and voting outcomes in Virginia counties across presidential elections. Our primary goal was to hone in on a few, key predictors of the Democratic margin of victory, utilizing a dataset that combined voting data with county level demographics.

To begin our exploration of these relationships, we focused on using linear regression and LASSO regression models. These methods provided valuable insights and through those models, avoided overfitting due to regularization. In particular, the proportion of the white population emerged as the most significant predictor amongst all other variables, demonstrating a strong positive relation with the Republican margin of victory. Certain economic indicators such as high poverty ratios and median rent had positive associations with Democratic margins, undermining the connection between demographic and economic factors in recognizing voter preferences.

The results of both models were consistent, with an R^2 value of 0.329, signaling that roughly one-third of the variance in the Democratic margin of victory is explained by the predictors. Still, the root mean squared error (RMSE) of ~9,704 votes highlighted noticeable variability across counties, suggesting an influence of additional variables or historical voting trends that were not included in the analysis. We encountered challenges during the analysis, first with merging the county data and voting_VA datasets and secondly with data inconsistencies,

such as duplicated vote counts from multiple voting methods. These issues were resolved through data cleaning and standardization, ensuring the robustness of the findings.

This analysis contributed to the understanding of how demographic and socioeconomic factors shape election outcomes. Insights for political strategies, policymakers, and researchers allow for predictions and forecasts for future election results or potentially analyze voter behavior. Virginia is a great state to begin with as it has a diverse demographic with shifting political trends, providing a sense of which patterns in electoral behavior may be acknowledged. Furthermore, incorporating log scaling variables, such as income and median rent, to better account for skewness and nonlinear relations within the dataset. This not only enhanced our model but also provided a more clear understanding of how economic factors influence voting behavior.

The remainder of this paper will first dive into our methodology including variables and modeling approaches. Next, the results, highlighting key findings, model performance metric, and interpretations. Finally, we end with the implications of our findings, addressing limitations, and propose potential insights into further studies. This detailed analysis aims to grasp a deeper understanding of electoral trends in Virginia and potentially in other states as well.

Data

This section of the paper will include the first important piece of our project, in which will further detail the variables and features of the project. Finally, it will discuss the salience of our research, challenges faced, and its relevance.

County Data contains many important demographic measures such as general population numbers broken down by sex, age, and race. There are also more specific categories such as educational attainment, household income, industry employed, home value, and rent all generally

broken down into the same categories. When joined with voting_VA trends can be observed between voting habits and household/individual traits. Voting data is provided for every national election since 2000 and demographic data since 2008 with votes for individual candidates and parties (listed as Republican, Democrat, Other).

These variables break down voting habits by county within Virginia. By joining this dataset with another that contains more specific demographic information, we hope to predict future voter tallies and election results. In doing so, we can study voter turnout (how it has changed over time) and candidate performance (compare the vote share of specific candidates across a myriad of counties to identify where they are stronger or weaker) to reveal how each county may vote.

We encountered a few initial challenges when merging the two datasets, **voting_VA** and **county_data**. The first issue was that the 2020 **voting_VA** data included excess information detailing different voting methods (e.g., absentee and ballot voting). This initially skewed the data by nearly tripling the overall vote count per county in Virginia for 2020. We have some ideas on how to address this issue to further clean our data, which will be our next step. The second issue arose after merging the two datasets. Every fourth row contained the relevant information we needed. While this was a minor issue, we resolved it by excluding all irrelevant data and keeping only every fourth row.

In the end, we don't expect any further challenges, but it was important to clean and refine the data early on to avoid any major setbacks.

Methods

With our data, we can explore the methods that were utilized to create our algorithm – supervised learning, classification, and specific models. This will serve as our pre-analysis plan which will further examine what success means in relation to our approach with the anticipation of weaknesses that may arise.

One observation in our study is one county, year pair, with variables such as median income, percent white, the winner of the county in that year, and the margin by which the winner won. To calculate the percent white population for each observation, we divided the white population by the total population for the given observation. For each observation, the input variables are demographic data for the given county, year combination.

A key component in our study is the utilization of supervised learning–using labeled data to train algorithms to comprehend patterns. Moreover, unsupervised learning is not applicable because we have labeled data such as party affiliation, demographics, margin, and more. Additionally, we will utilize regression analysis to predict the margin by which the winning party won the election. Regression is best suited to continuous tasks, and the margin of victory is the relevant variable in first-past-the-post elections such as in the United States.

For our analysis, we'll need a model that works well with supervised regression tasks; we want our output variable to be continuous, and we want our model to consider historical trends to predict classifications for 2024. For our dataset, linear classification makes the most sense given our continuous target variable. Further, linear models produce coefficients for the relevant variables of analysis, providing transparency into which factors are the strongest drivers in predicting election performance. Finally, linear classification makes the most sense for our project because it's a relatively simple algorithm that will run quickly on a large dataset.

To best assess the success of the model, we will split historical data into training and testing sets to evaluate our model's performance on county, and year combinations where we know the actual result. To measure the strength of our model, The performance of our linear regression model will be evaluated using metrics appropriate for continuous predictions, such as r^2 and root mean squared error (RMSE). The r^2 value will measure how well the model explains the variance in the Democratic margin of victory across counties, while the RMSE will provide insight into the average prediction error in votes. Together, these metrics will help assess the model's fit and its ability to provide meaningful predictions about electoral outcomes.

If our approach works, we will be able to accurately predict the outcome of an election in most counties. While election results are never completely certain, our model should be more insightful than a naive approach such as simply predicting the candidate with a higher polling average or flipping a coin.

Some key weaknesses that we anticipate being an issue may be model overfitting because if the model places too much emphasis on past election outcomes and cycles (each election has unique dynamics and characteristics), then the model may miss new emerging issues as the model may learn from previous historical anomalies. To deal with this, we must ensure the model is not biased toward a single variable or factor from previous election cycles. Additionally, cross-validation could serve as a solution to evaluate the performance of the model, which reduces overfitting. If our approach fails, it would reinforce the idea that elections are highly unpredictable, especially with a highly trained model; thus, requiring a model that is more adaptable.

Not only that, sudden, impactful events such as economic downturns or pandemics can drastically cause a shift in voter turnout and action, as seen in the COVID-19 pandemic. This is a

difficult issue to overcome due to the unpredictability of events, and a model like this may miss those effects. Therefore, to mitigate those weaknesses, using a more adaptive model (retraining capabilities) in the event of geopolitical changes might be more effective. A lesson that we can learn from if it fails is that machine learning models may have limitations in predicting outcomes like an election which is oftentimes chaotic. It could be argued that encompassing a more "contingent model" would react to game-changing events more accurately.

Results

Finally, this section will sum our results in a comprehensive analysis which will discuss several key tables and our findings that help us better understand how selected demographic and socioeconomic factors influence the Democratic margin of victory in Virginia's presidential elections.

Prediction Question

How do demographic and socioeconomic factors, such as poverty levels, income, median rent, and the proportion of white population, influence the Democratic margin of victory in Virginia counties during a presidential election year?

Comparison of Models

The results (Figure I) show that LASSO did not exclude any features because the coefficients remained non-zero in the LASSO model.

Figure I

	variable	linear_coefficient	lasso_coefficient
0	income_log	6724.682626	6690.825736
1	poverty_ratio	1660.241808	1696.903113
2	median_rent_log	4195.132712	4206.051610
3	pctwhite	-3849.206533	-3856.592050

Key Observations

1. Consistency Across Models: The LASSO coefficients are very close to the Linear Regression coefficients. Some of those adjustments contributed to variables such as, poverty_ratio, median_rent_log, and whitepop increased slightly, indicating a slightly greater positive effect under the LASSO model. Therefore, this could indicate that the regularization penalty applied by LASSO was relatively small and didn't significantly shrink the coefficients. Moreover, the predictors are relevant to the target variable (margin), so LASSO didn't need to exclude any.
2. LASSO Retained all Features: Since all the features were retained, it indicates that each predictor contributes to the target variable and that there isn't significant multicollinearity or noise in the data that would result in irrelevant features or overfitting in the data.
3. Model Interpretation: Both models suggest that all four variables (income_log, poverty_ratio, median_rent_log, and whitepop) are significant in explaining the variation in the Democratic margin of victory. The LASSO model has slightly different coefficients which indicate that it applied a mild regularization effect – reduces some regression coefficients to zero to prevent overfitting without much bias. Thus, improving the predictive performance model.

4. Significance of Log-Scaling: The inclusion of log scaled variables, such as income_log and median_log highlights the importance of accounting for skewness or non-linear relationships within the data. This allows both models to become more interpretable and better illustrate the relationship between predictive variables and the Democratic margin of victory.

LASSO did not exclude any features due to several factors including strong predictors, low regularization strength, and minimal multicollinearity. Strong predictors reveal that all features may have significant correlations with the target variable; thus, improving predictive accuracy. As for the low regularization strength, it indicates that the penalty term's influence was small and the model retains its flexibility, leading to low bias. Finally, minimal multicollinearity reveals that LASSO will retain all variables, by doing so it improves predictive performance. This is due to low redundancy in which the coefficients directly represent each feature's predictability (better interpretability) and the model is less sensitive to small changes in the data (stronger stability).

Key Takeaways

There are three important takeaways that can be drawn. First, linear regression and LASSO produced similar R^2 values and coefficients which indicates that all predictors are relevant to the model. As a result, it contributes to the explanation of the variance in the target variable. Furthermore, it indicates that our model is robust without needing to eliminate a lot of features. Secondly, LASSO retained all features, but slightly shrunken coefficients – reinforcing the robustness of the features – without sacrificing predictive accuracy and performance. Additionally, it suggests that there is minimal multicollinearity in the predictive model because LASSO did not make coefficients to zero. Finally, the incorporation of LASSO's built-in

regularization also adds another layer of robustness by preventing overfitting, which is important because there could be a lot of noise in our model. The slight shrinkage of coefficients further indicates that it is generalized efficiently while ensuring that our model is easily interpreted and reliable for future predictions.

By understanding that LASSO achieves comparable performances, we can now transition into the interpretation of the coefficients produced by the LASSO model. This will allow us to better understand the relationship between our predictors and the margin of victory, as shown below.

Feature Selection and Coefficients

We used an 80 / 20 train-test split for the LASSO model, and our data included four features (Appendix A: Codebook). Because the LASSO model did not exclude any features and applied only a small penalty to the linear coefficients, we will use the LASSO model coefficients to interpret the results of our project.

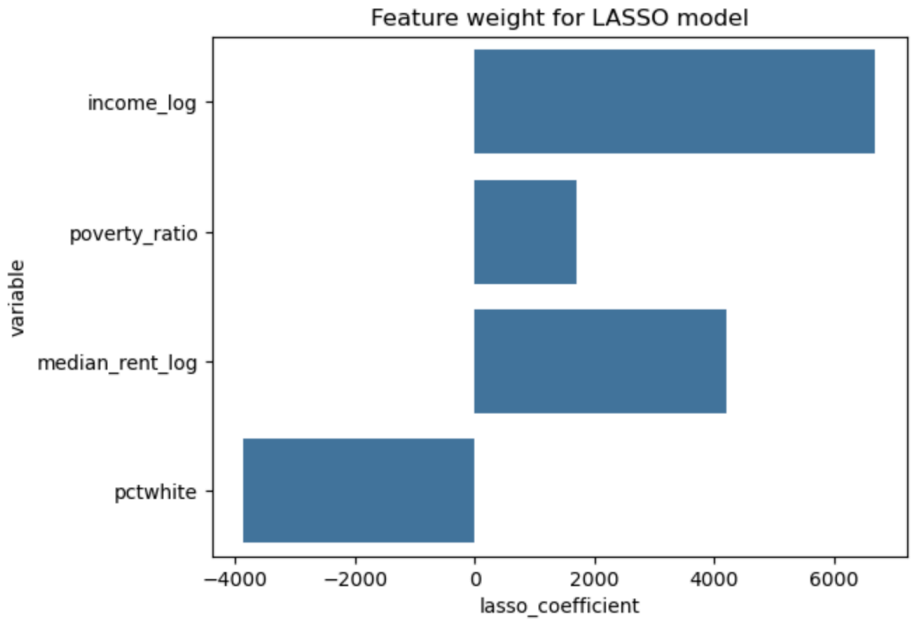
To reduce the impact of widely distributed variables, income and median rent were log-transformed to reduce skewness, and all predictors were subsequently standard-scaled (mean = 0, standard deviation = 1) to facilitate direct comparisons between coefficients.

The coefficients of each predictor in the LASSO model are shown in Figure II. These results can be interpreted as follows: “A one standard-deviation increase in variable X is associated with a β_1 change in voting margin, holding all else constant.” Variables with large positive coefficients, like log-scaled income, strongly increase the model’s predicted margin for Democrats. The only variable that increases the model’s predicted margin for Republicans is the percentage of the population identifying as white, with an effect of ~4,000 votes.

The model's performance is moderate, with an R^2 of 0.329. This indicates that approximately 32.9% of the variance in the Democratic margin of victory is explained by the features included in the model (Appendix A). The root mean squared error (RMSE) of 9,704 votes suggests that the average error for the predicted margin across counties is relatively large. This reflects the limited number of features selected and the limited number of data available, with only 532 observations in the Commonwealth of Virginia across four election cycles.

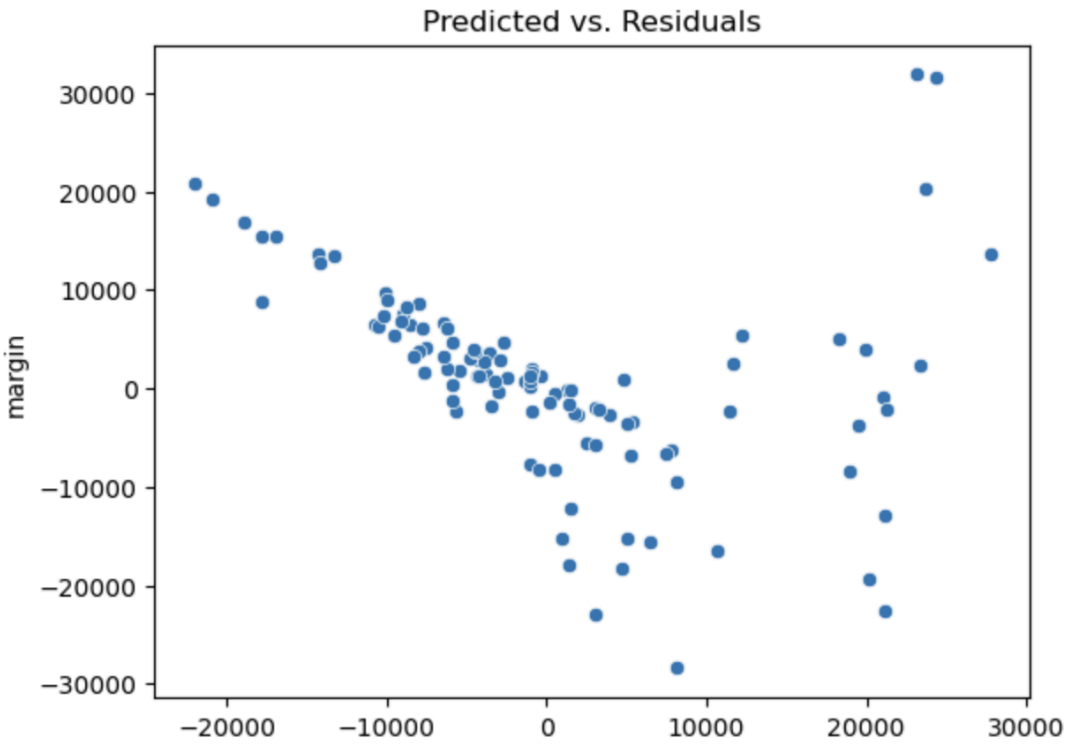
Among the selected features, percent white was the only statistically significant predictor at the 0.05 level, with a coefficient of -3849.21. This can be interpreted as "a one-standard-deviation increase in the proportion of the white population is associated with a 3,849 increase for the Republican margin, holding all else constant." This variable remains largely unchanged between the two models, suggesting its strength as a key critical predictor. Other predictors, such as percent living under the poverty line (1,660.24) and scaled median rent (4,195.13), demonstrated positive associations with the Democratic margin, albeit with smaller magnitudes.

Figure II



The residual plot (Figure III) reveals that the model's residuals are not evenly distributed around zero, suggesting heteroscedasticity. The variance of the residuals increases greatly at higher predicted margins, indicating that the model performs quite inconsistently for counties with extreme Democratic margins. This reflects the limited set of predictors and the unique variance of specific counties (e.g., Fairfax County) that are not fully captured by the model.

Figure III



Summary of Findings

These results emphasize the relationships between demographic and socioeconomic factors and voting outcomes in Virginia. The influence of the percent White variable on the Republican margin of victory suggests a pivotal role in the composition of certain demographics and how it shapes electoral outcomes. On the other hand, the income_log variable has a negative relationship with the Democratic margin of victory, possibly indicating that higher-income areas are more likely to align with Democratic preferences and the overall wealth gap. When looking at the last two variables of poverty_ratio and median_rent_log, indications of economic well-being may be presumed, and while influential, are secondary in demographic factors within the models conducted. Lastly, a moderate R^2 value suggests that these features don't capture the entire story, but remain important predictors, providing valuable information into trends and tendencies among voter characteristics.

Among the selected features, only percent identifying as white was statistically significant at the 0.05 level, with a strong positive relationship to the Republican margin. The other predictors showed no statistically significant relationships but still may offer some explanatory value in specific contexts.

Conclusion

This report aimed to understand how selected demographic and socioeconomic factors—measures of income, poverty levels, median rent, and racial composition—influence the Democratic margin of victory in Virginia's presidential elections. By merging county-level demographic data with historical election results, this report aimed to reveal patterns that could help predict and interpret future voting behavior. Our model yielded insights into the factors that shaped electoral outcomes but also revealed limitations and opportunities to refine our approach and expand our understanding.

The key finding of our analysis was the influence of the proportion of the population identifying as white on election margins. In our model, the percent white was the only statistically significant predictor at the 0.05 level, associated with a stronger Republican margin, all else held constant. This relationship persisted across the different modeling approaches. Collectively, our selected features explained approximately one-third of the variation in the Democratic margin, suggesting that future research should include additional variables and modeling strategies to better understand electoral behavior.

One challenge that emerged in our analysis was the disproportionate influence of Fairfax County on the election predictions. As the largest county in the dataset by a margin of over 700,000 votes, the characteristics of the four observations in Fairfax differ substantially from other counties in Virginia. Its demographic and economic conditions, as well as the magnitude of

its vote totals, influencing the margin of victory, could skew our residuals and introduce heteroscedasticity into the model. However, rather than view the outlier as a problem, future work should treat Fairfax as an informative case study. Future studies should examine Fairfax separately to understand how the largest voting bloc in the state behaves given changing economic and demographic conditions. Future work could also consider using hierarchical models like trees to make distinctions between urban and rural counties, which could prove more effective in understanding how the urban/rural divide defines electoral outcomes in Virginia. Other models could incorporate more contextual variables, like workforce composition, commuting and migration patterns, or level of education to better explain Fairfax's persistent Democratic lean.

Most critically, our current model is limited to only a few socioeconomic and demographic variables, limiting its power to explain variance in electoral outcomes. Including more detailed racial and ethnic breakdowns, indicators of political engagement, or religious affiliation, among many other features, could more precisely identify predictive factors in voting behavior. At the statewide level, future models could incorporate policy stances on issues specific to the state, or measure the political salience of those issues. Finally, incorporating time-series elements to explicitly model changes across different election cycles is an essential step to understanding how shifting demographics influence voting patterns over time.

Also, future research could explore using different modeling techniques to capture richer insights. Linear and LASSO regression provided useful starting points for the purpose of our project, but hierarchical models like trees could also prove valuable in understanding voting behavior.

Another overarching limitation is the inherent difficulty in predicting elections in the first place. Unforeseen events significantly alter voter behavior and preferences regularly. For example, the COVID-19 pandemic changed how people vote and raised the salience of health and economic issues (Baccini et al., 2021). In the future, research should focus on developing prediction frameworks that can be updated rapidly in the face of disruptions.

Understanding the demographic and economic dimensions of voting behavior offers practical implications for political stakeholders. The most obvious application is for political campaigns to understand how to better tailor their outreach and messaging to have the most significant effect on electoral outcomes. Understanding what factors are the strongest predictors of electoral behavior would be an invaluable tool to campaigners to make the best use of limited resources. Policymakers might also benefit from recognizing that certain communities are more politically responsive to particular issues. Finally, this work contributes to the growing academic interrogation of electoral behavior and its determinants. Our experience highlights the need for careful data preparation, a broader range of variables, and consideration of further modeling techniques for future researchers to build upon.

In conclusion, while our study sheds light on how selected demographic and socioeconomic variables relate to electoral outcomes in Virginia, it also highlights the need for richer data and more comprehensive approaches. We hope that subsequent efforts may provide a deeper understanding of how evolving demographics, economic conditions, and social forces shape the political landscape in Virginia and across the nation.

References/ Bibliography

MIT Election Data and Science Lab, 2018, "County Presidential Election Returns 2000-2020", https://doi.org/10.7910/DVN/VOQCHQ, Harvard Dataverse, V13, UNF:6:GILlTHRWH0LbH2TItBsb2w== [fileUNF]

Steven Manson, Jonathan Schroeder, David Van Riper, Katherine Knowles, Tracy Kugler, Finn Roberts, and Steven Ruggles. IPUMS National Historical Geographic Information System: Version 19.0 [dataset]. Minneapolis, MN: IPUMS. 2024. http://doi.org/10.18128/D050.V19.0

U.S. Census Bureau. County Adjacency File [dataset]. U.S. Census Bureau. 2023. https://www.census.gov/geographies/reference-files/2023/geo/county-adjacency.html