Key Demographics in Presidential Elections

DS3000 Project

By Michael Baraty, Nathan Parker, Maximus Saenz, and Kai Webber

Abstract

Elections in the United States remain a source of uncertainty for many, and its repercussions are large. There are many economic, social, and political effects that are rooted in the unclear nature of how elections are determined. In this project, we set out to understand what demographic factors have significant impacts on election outcomes. Using county level election and demographics data from the MIT election lab, we prepared and analysed the data in order to create and test machine learning models. We utilized both regressors and classifiers to understand election outcomes and swing county demographics. Following a discussion and evaluation of our results, we cover potential impacts and areas for further study that stems from our research.

Introduction

Uncertainty regarding elections and political outcomes have many effects economically. Specifically, corporate investment decision making impacted by political uncertainty, where in years with high political uncertainty, there is a significant decrease in investment-to-price sensitivity. This change means that "the company observes 6% lower sales growth over the two years following the election" (Durnev, 2010). Creating a more concrete way to think about election outcomes can help reduce this uncertainty and therefore lead to less economic instability.

Currently, elections are predicted through a number of means, historically through polls, and more recently through sentiment analysis of social media data. Polls have had their many issues, especially in recent years due to "low-response rate, mis-representation and the social desirability bias/lies" that come from traditional polling methods (Zhou et al., 2021). More recently, there has been a rise in using social media, though this is still in early stages. Some issues that exist in the current literature include noise from fake accounts, weakness of sentiment analysis algorithms, and the complexity of human communication (Chauhan et al., 2021).

Combining the importance of reducing uncertainty and utilizing new methods to analyze election outcomes can bring stability in many senses. Utilizing a data-driven approach to looking at elections, we can begin to sort through the noise of political uncertainty.

Within the democratic framework of the United States, presidential elections stand as critical junctures that shape our society's trajectory. These elections not only determine policy directions and economic conditions but also influence the international standing and relations of the nation. As we approach another presidential election year, the imperative to decode the election dynamics intensifies, underscoring the project's urgency and relevance. Our focus will be on delineating pivotal states, which will most likely be the known swing states, and identifying demographics that have historically played a decisive role in shaping election results. By employing data science techniques to analyze past elections, this project aims to forecast potential outcomes for the current year, enabling us to anticipate and prepare for the ensuing changes in our daily lives and societal structure.

Our project will rigorously maintain neutrality, providing an unbiased analysis irrespective of party involvement in the election. Aimed at revealing key electoral dynamics, our data-driven insights can guide campaigns on strategic demographic targeting. Although primarily useful for campaign strategies, the findings also offer broader insights into electoral behavior, beneficial for analysts and the public alike.

Data science is crucial for this project due to the overwhelming volume of electoral data available, which makes drawing meaningful conclusions challenging. Data science techniques, including machine learning and statistical analysis, are vital for distilling this extensive data into actionable insights. These tools allow us to identify significant patterns and demographics essential for informed electoral strategy, underscoring the necessity of data science in understanding electoral datasets.

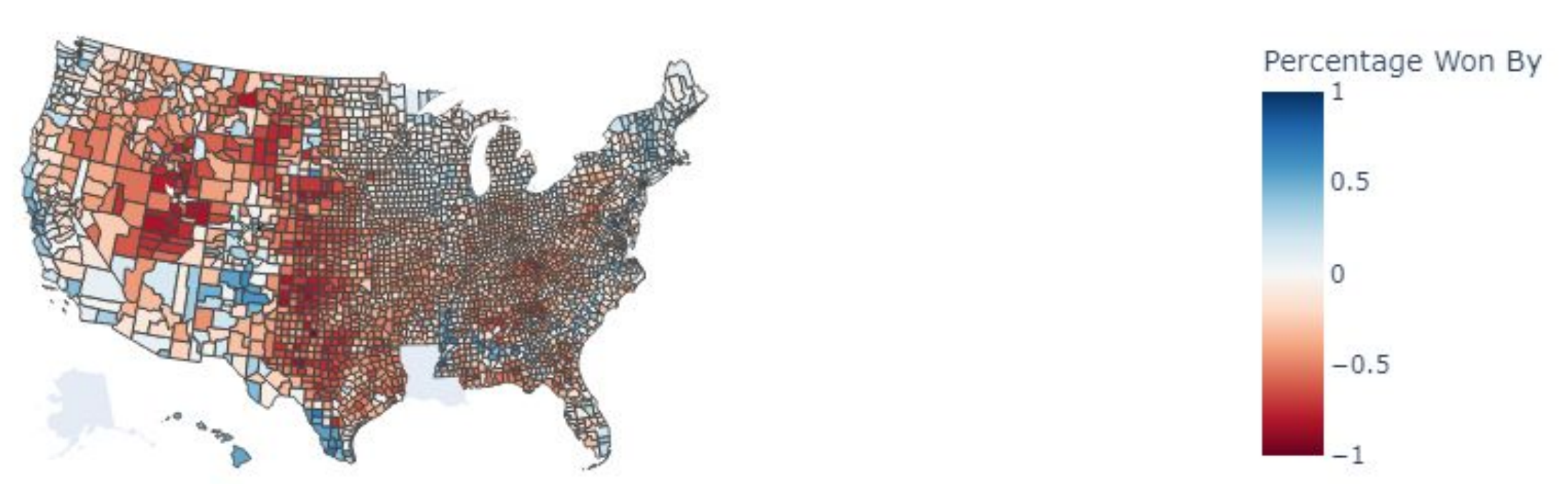
Methodology

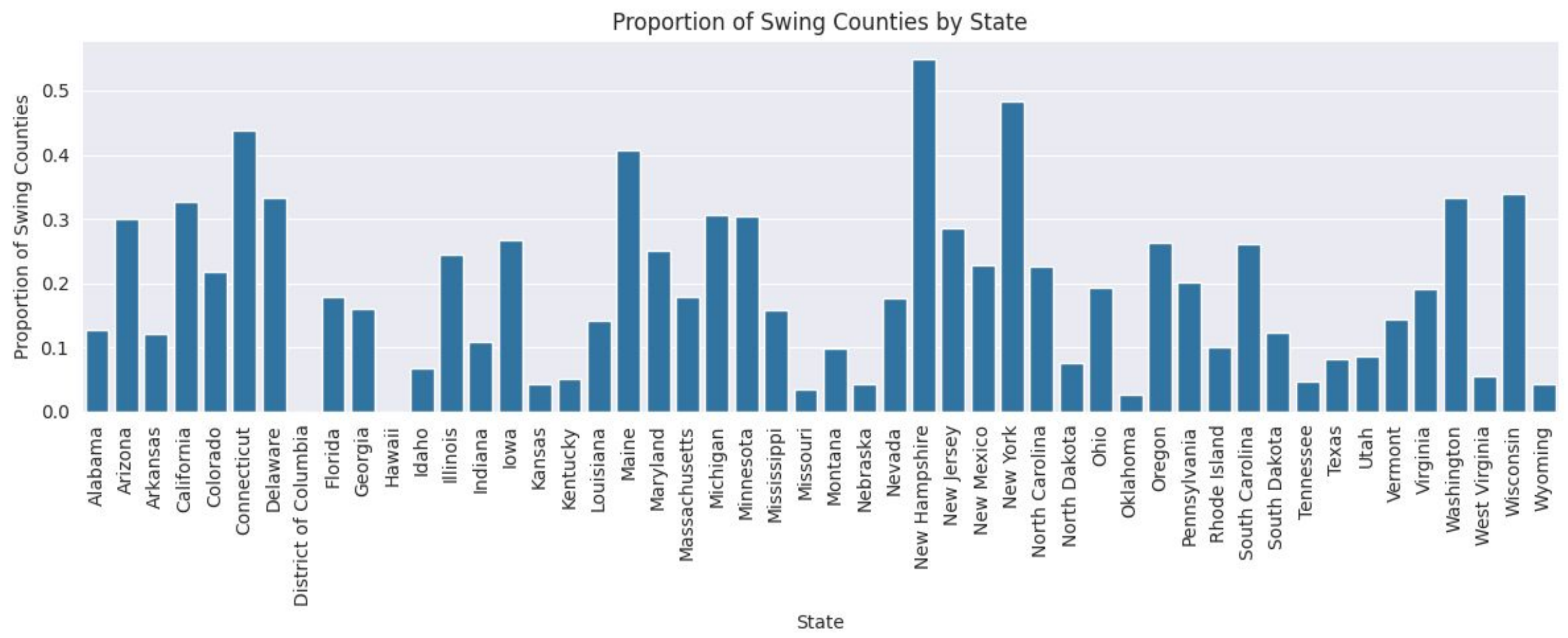
In order to determine what factors are important and significant in deciding US presidential elections, we utilized election outcomes, swing counties, and demographic data on the county level from MIT Election Lab. We focused on the 2012 and 2016 presidential elections due to the availability of the data, and some key demographics we focused on include race, gender, concentration of immigrants, age, education, employment, income, and rural/urban settings. Data cleaning included dropping columns we would not be using, converting variables to percents or total amounts, setting data types, removing/modifying missing data, and adding variables to track swing states. We created visualizations at this stage to help us understand our data, looking at the margins between parties to inform how we viewed swing counties, as well as looking at the concentration of swing counties per state to see which states are most affected by this project. We also created boxplots to see the distributions of key variables to get a better sense of the form of our data.

Due to the nature of the two party system in the United States, we were able to use both classifier and regression machine learning models to evaluate our data. For classifiers, we looked at which party won the county, and for regressors, we looked at the proportion and total amount of votes that each candidate received, as well as figuring out what demographics were important in making swing counties. In order to find the most robust algorithm, we tested four types of machine learning models: K-Nearest Neighbors, Random Forest Regressor, Random Forest Classifier, and SVM. For each model, we tuned parameters appropriate to each model, including different k values, n-estimators, max depth, max features, c, gamma, and kernel variables respectively.

In order to determine the fitness of our regression models, we will look at accuracy scores, mean squared error, and R-squared. For our classifier models, we will mainly focus on our f1-scores, since it is a combination of recall and precision, and gives us a broader view as to the accuracy of our model.

Initial results from our models performed well, and we will further discuss the model that serves as the most useful in the results and evaluation sectio.





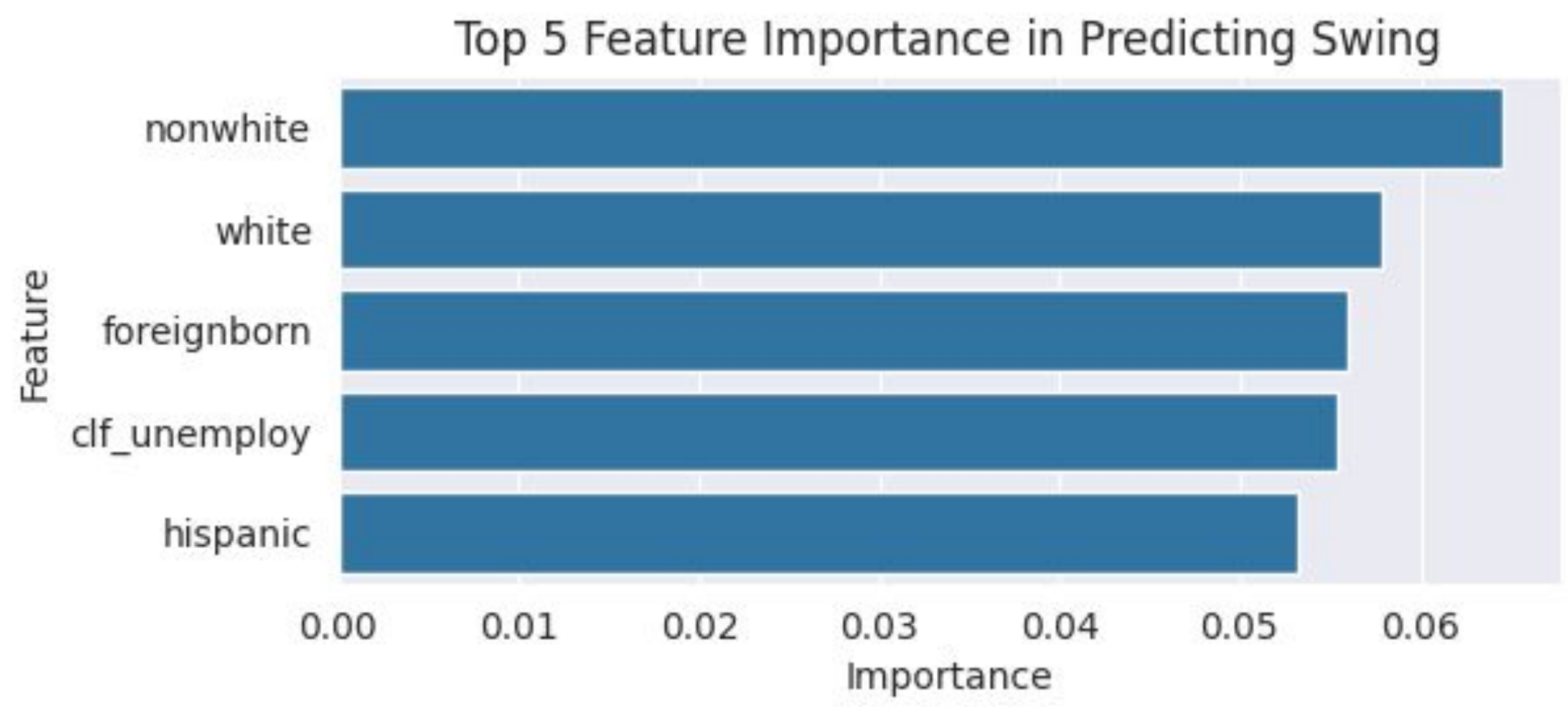
Results and Evaluation

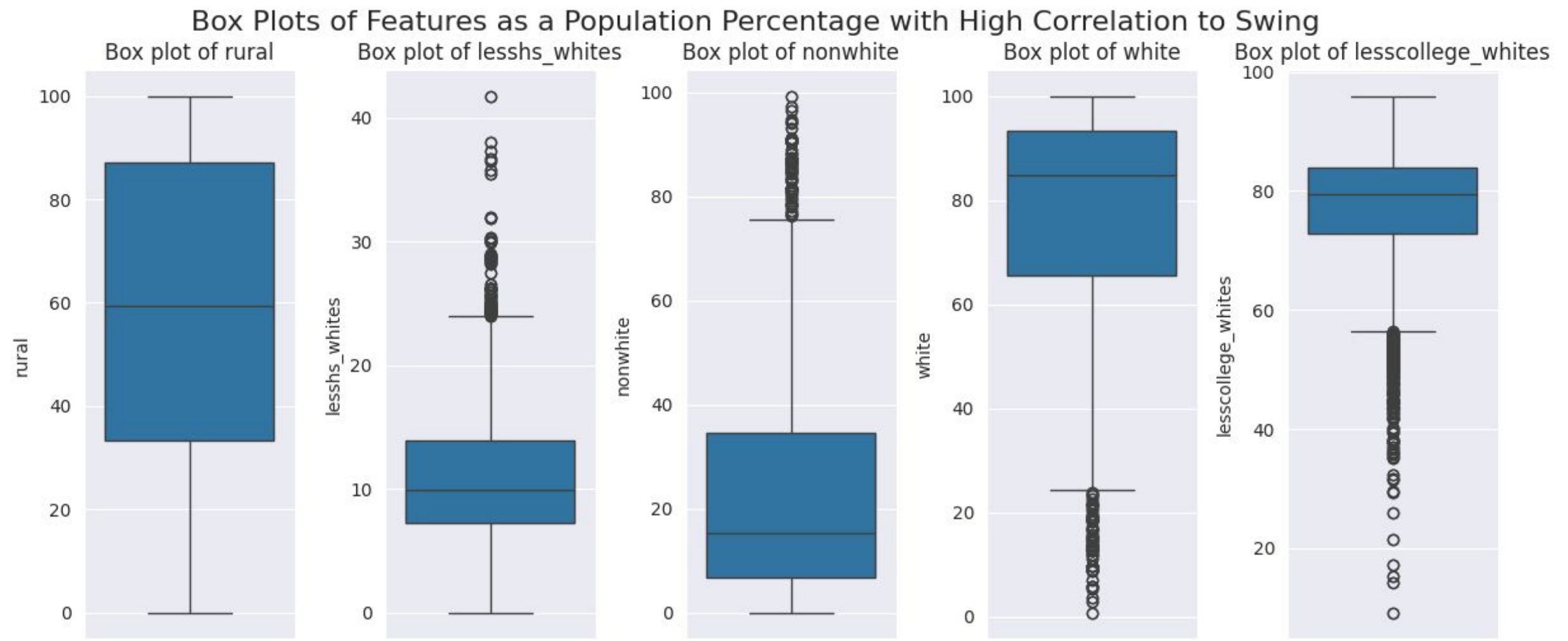
The model that we ended up selecting was the random forest classifier that determined whether or not a county would be classified as a swing county. We decided to focus on the demographics that determine a swing county as opposed to election outcomes, because it is a more unique question that can shift the focus on who and where to appeal to as a candidate. We also felt that focusing on predicting the winner is a task that can beyond beyond the scope of a purely data approach, and we felt we had a better chance at creating a stronger, more robust model using this classification problem.

The random forest classifier had an accuracy score of 85%, and an f-1 score of .91 for non swing counties, and .25 for swing counties. While there is high variation, we found it was our best model with the least overfitting or underfitting of the data. By hypertuning our parameters, we found that the best max depth was 10, the best minimum number of leaf samples was 1, best minimum number of split samples was 5, and best n estimators was 10.

Based on this model, we found that the most important demographics were non-white, white, foreign born, unemployment, and hispanic. These demographics are particularly interesting, because it shows how a lot of whether a not a county is a swing county relies on race and ethnicity. For further study, would want to examine which race groups are the most important, along with why. Predicting the unpredictable can bring along certainty to elections and help demystify an unclear process.

As a whole, our models all performed pretty well given the nature of the problem that we set out to solve. Elections are a game that go beyond demographics, and we would need to include other important metrics, including sentiment and current events to paint a full picture to predict elections, informing why we switched towards county level swing status.





The Impacts

The results of this project will likely be most impactful for those who speculate on US presidential elections in any form. From politicians and their campaigns to economists planning future policies, the results here can help inform how they approach elections going forward. While demographics are not perfect predictors, they can be added to the robust tools that are used to inform how elections are thought about as older tools like polls fall out of favor. Similarly, the techniques we used can be replicated on further historical elections, as well as tested on recent demographic data to provide some insight as to how this election may go. Incorporating data, statistical analysis, and machine learning can provide a more complete picture and reduce uncertainties that surround elections, and we hope that our results can inform how future experts of all disciplines see presidential elections.

Conclusion

Based on the results of our model, while we were not able to create a perfect algorithm that can completely predict election outcomes or predict swing counties, but we learned a lot about what factors go into shaping US presidential elections. Through shifting our focus to swing counties, we were able to determine some key features related to race and ethnicity when it came to shaping whether or not a county would be a swing county.

At the outset of this project, we had hoped to have a better understanding than when we started and when compared to the previous literature. We achieved this goal by providing a look at demographics in a way that had not been previously done. What we learned most importantly is that, like most problems that involve human behavior and emotion, concrete outcomes cannot be perfectly predicted with data. The goal of utilizing data in these scenarios should be help inform, but not to serve as a concrete truth.

Beyond its use in election prediction, the methodology that we used, combining demographic data with some other outcome, could be an important step in understanding how other processes function. In a lot of social science research, increasing emphasis has been placed on understanding how marginalized groups are most impacted by certain programs or issues, and incorporating demographic information can provide much needed context to bring a more holistic approach to research.

To improve upon our model given a larger scope, we would look into expanding the data that we used. In terms of elections, we could go beyond the two that we used, and look into more historical presidential elections to build more robust models. We could also expand our analysis to include elections outside of presidential elections, including local elections, state elections, and other federal elections. Further, we could include more demographic information, including career, infrastructure development, and much more. Finally, we could continue to expand upon the model by including more qualitative measures of election outcomes, including social media sentiment analysis to provide a more complete picture of how elections are determined.

Works Cited

Chauhan, P., Sharma, N. & Sikka, G. The emergence of social media data and sentiment analysis in election prediction. *J Ambient Intell Human Comput* **12**, 2601–2627 (2021). https://doi.org/10.1007/s12652-020-02423-y

Durnev, A. The Real Effects of Political Uncertainty: Elections and Investment Sensitivity to Stock Prices (2010). http://dx.doi.org/10.2139/ssrn.1549714

Zhou, Z., Serafino, M., Cohan, L. *et al.* Why polls fail to predict elections. *J Big Data* **8**, 137 (2021). https://doi.org/10.1186/s40537-021-00525-8

Data Source:

MIT Election Lab

(https://raw.githubusercontent.com/MEDSL/2018-elections-unoffical/master/election-context-2018.csv)