Emily Macris, Hiba Khan, Jesus Zarate, Kiduse Gezehagne, Rix Prakash, Shaveen Saadee, Vaibhav Jha
ML Pre-Analysis Plan
DS 3001
4 November 2024

Question: Which candidate is most likely to win the battleground states in the 2024 U.S. presidential election?

I. What is an observation in your study?

Our study leverages a machine learning model that focuses on seven key battleground states – Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, Wisconsin – to predict the outcome of the 2024 U.S. presidential election. Each row in the data represents a county's measures in a prior election. More specifically, an observation contains information on six demographic-based variables including race, gender, education level, economic status, age, and poverty rate. Each county is represented in the data three times, one for each of the prior election cycles (2012, 2016, 2020).

II. Are you doing supervised or unsupervised learning? Classification or regression?

We are using supervised learning in this project because we have historical voting data labeled by county as either Democrat or Republican. Supervised learning allows us to train a model using these known labels, so it learns patterns in the data - such as income levels, age distributions, education rates, and other demographic factors - and then apply what it learns to predict the voting outcome in each county for future elections. Specifically, we are performing classification using a decision tree-based model, which predicts discrete categories (Democrat or Republican) rather than a continuous value. In decision tree-based classifications, the decision tree makes predictions by asking a series of yes/no questions about the features.

III. What models or algorithms do you plan to use in your analysis? How? - How will you know if your approach "works"? What does success mean?

Our group plans to use decision trees as the basis for our analysis. Decision trees will allow us to find patterns of voting behavior based on the variables we selected in the exploratory data analysis. All explanatory variables are numeric, therefore we hope to find splits between different variables to optimize classification of party voted for by county. With decision trees, we envision a series of decision nodes that distinguish Democratic counties from Republican counties. Outcomes are binary, therefore one terminal node will be set as Democratic and the other Republican. The goal is to find the best decision nodes that effectively split the various numeric explanatory variables at measures that distinguish counties' voting outcomes.

Our approach will be bound to "work" if the splits used to make decisions at each node for prior elections are similar to 2024 patterns. For instance, if a decision is made that counties above 0.55 proportion of White-identifying individuals compose a county vote Republican, then we would hope that the 2024 voting trends warrant a similar split. A drastic change in demographics could jeopardize predictions for the 2024 election. A successful model – accurate voting predictions for battleground states – will involve specification of decision nodes in decision trees.

A Random Forest model improves this approach by building multiple decision trees, each trained on a random subset of the data and features, and then combining their predictions. By taking a majority vote across all the trees, the Random Forest reduces errors from any individual tree and achieves a more accurate, stable prediction. This technique is particularly effective for our project because it handles complex, nonlinear relationships in the data and can work well even when features interact in intricate ways, such as how age and income together might influence voting patterns. The result is a robust prediction for each county based on learned patterns from the training data.

IV. What are weaknesses that you anticipate being an issue? How will you deal with them if they come up? If your approach fails, what might you learn from this unfortunate outcome?

Decision trees have a high possibility of overfitting, especially if the model is deep with a large number of splits. This may rely heavily on the training set and fail at generalizing to unfamiliar data with future election county demographics. To deal with this, we plan to limit the maximum depth of the tree to control the model's complexity and reduce overfitting. Another weakness can be the limited performance with categorical variables. The model may struggle with very detailed, granular income levels or the education levels categories, so we plan to group them into larger brackets which can simplify the data and help the model find meaningful splits rather than these narrow categorical splits. Finally, the model has a weakness of failing to capture trends. Over time and especially in elections, there are changes in political, economic, and social voter sentiments that are not being accounted for in the data.

If the decision trees fail to predict the election, it would bring attention to issues that we could learn from. First, voting decisions may not be entirely related to demographic data or generalizations here but rather be influenced by non-demographic factors such as current political events or party specific policies that the decision based model won’t be able to capture. Secondly, if it fails then we would believe our demographic variables may not be the key drivers of voting behaviors and could be misleading features. This would call for an analysis of other influential factors that could be specific to the current election. Finally, we would learn that there was an overemphasis on historical data and that this isn’t sufficient. The constantly changing political and social changes may need a model that is more adaptive and can incorporate real time inputs.

V. Feature Engineering: How will you prepare the data specifically for your analysis? For example, are there many variables that should be one-hot encoded? Do you have many correlated numeric variables, for which PCA might be a useful tool?

The first step in this stage will be data preparation, where we will handle the peculiar characteristics of numeric data to ensure that each feature best fits the requirements of decision tree-based algorithms. Given the demographic variables involved in this project, the structure and representation of data will be critical to model performance and interpretability.

Variables such as race and level of education, which can have multiple categories, will be transformed into binary columns representing each category – separate columns for White, Black, Hispanic, etc., for race. It therefore gives equal representation to each category in the model, not creating an artificial hierarchy that could happen with ordinal encoding. Similarly, for binary features, such as sex, a simpler binary encoding is going to be used in order to encode the two possible values without wasting the expansion of the feature space and an incomprehensible dataset.

We will perform correlation analysis for numeric variables, like median income, poverty rate, unemployment rate, and education levels, to understand the relationships in these features. Although decision trees and random forests can handle multicollinearity, this step will show some insight into potential interactions and dependencies among the demographic and economic indicators that might influence election outcomes. Moreover, looking at correlations will help to see the underlying patterns in each state, which will enable us to interpret model results more concisely. For instance, changes in median income, education attainment, or population changes could be represented as new features, providing a dynamic perspective of evolving county-level characteristics. This helps in the capturing of any socio-economic trends or demographic shifts that might correlate with election results across the years.

VI. Results: How will you communicate or present your results?

In the presentation of the results, feature importance will be a focus point because decision tree-based models are interpretable. We will generate feature importance plots for each state with the aim of projecting which variables carry the most weight in determining election predictions. Such visualizations will reveal patterns across demographic and economic indicators, thus helping to bring an intuitive understanding of the main predictive factors and how these differ across states.

There will also be side-by-side state-level comparisons showing the differences in predictive factors, thus gaining insight into which demographic and economic factors most influence each swing state. It will enable us to make a presentation showing a panoramic view of county-specific and statewide drivers of election outcomes by capturing the unique characteristics of each state, hence contributing to a more nuanced understanding of the predictive factors that may continue to shape future elections.