Pre-Analysis Plan:

1. What is an observation in your study?

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.

For each observation, the target variable “rdummy” is a one-hot encoded dummy variable, where 1 represents a Republican win for that county, year combination, and 0 represents another (always Democratic, but could technically be independent) win. 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.

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

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, winner, and more. Additionally, we will utilize classification in our study, such as “democrat” or “republican.” This is salient because classification models assign the labels to a case based on our inputs. The purpose of this is to break the subject into a more manageable part. Regression analysis is not useful because it predicts continuous values like prices.

1. What models or algorithms do you plan to use in your analysis? How?

For our analysis, we’ll need a model that works well with supervised classification tasks; we want our output variable to be binary (win/loss for each party), and we want our model to consider historical trends to predict classifications for 2024. For our Dataset Logistic classification makes the most sense given our binary target variable. Further, the logistic model, in producing coefficients for the relevant variables of analysis, will provide transparency into which factors are the strongest drivers in predicting election performance. Finally, logistic classification makes the most sense for our project because it’s a relatively simple algorithm that will run quickly on a large dataset.

Using data resampling/bootstrapping we can also verify the validity of trends among certain demographic variables giving a greater level of confidence in the observed result. (ex: comparing white and non-white county voting we can resample the data and graph a density plot and ECDF, by using a confidence interval of 10% we can be assured in drawing conclusions from our data)

1. How will you know if your approach "works"? What does success mean?

Success of the model can be measured both qualitatively and quantitatively. Quantitatively, we’ll split historical data into training and testing sets to evaluate our model’s performance on county, year combinations where we know the actual result. To measure the strength of our model, we’ll evaluate the accuracy of the model on the test set (how many predictions did the model guess correctly?) We’ll also evaluate the precision of the model on the test set (how many predicted wins did the model get correct?) Because we are using classification instead of regression, quantitative measures like r-squared or root mean squared error are not applicable.

We also plan to create a confusion matrix based on the performance of the model on the test set to best understand which cases cause the model to make errors. For example, if false positive results are particularly common for a certain set of counties, we might investigate what variables those counties share to understand how our model could be improved. A strong model for our purposes will minimize false negative results to reduce missed “true wins.” Underpredicting in these counties could overlook relevant shifts where voter behavior affects the overall outcome.

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.

1. 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?

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 from 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. 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.