PPOL564 Final Project: Predicting Election Results

Word Count:

Colette Yeager

Introduction

For this project, I chose to look at all of the features that play a role in the results of presidential elections, to see which are the most influential. In this report, I will show how I pull data from several different sources to see election results, economic growth, state results, turnout rates, and information about each president. I will go through the process of how I organized, cleaned, and combined my data to end up with a larger data set and database to be able to model. Ultimately, the statistical model I am looking at is how each of these factors predict the vote share a President will win by. I will then look at some of my results from the statistical models and analyze which features ended up being the best predictors.

Problem Statement and Background

The goal of my project was to determine which factors are the most important in predicting future election results. There have been many studies related to this topic, both by campaigns and by forecasters such as those at FiveThirtyEight. Doing these types of studies can very helpful in determining the best campaigning strategies. These studies focus mostly on results from polls[1]; I wanted to see if I could find other key useful predictors other than this.

One famous system for predicting outcome is The Keys to the White House, a checklist of 13 statements about the circumstances surrounding a presidential election[2]. Many of these are difficult to measure quantitatively, such as whether there was social unrest during the previous term whether the candidate is charismatic of a national hero. Two of these that I decided to use in similar ways in my study were whether the incumbent-party candidate is the sitting president and the level of economic growth in the previous term.

Economic growth in particular has been widely studied all over the world, especially when one of the candidates is the incumbent. When the economy has not grown or has gotten worse during an administration, voters might be likely to want to vote for the party or same person that was in office previously. One discussion by shows that the extent that economic performance is reflected in vote share is reflected by citizen's evaluations of the

[1] https://www.sas.upenn.edu/~baron/journal/18/18124/jdm18124.html

[2] https://pollyvote.com/en/components/models/mixed/keys-to-the-white-house/

overall economy[3]. In this study, they similarly pulled from several years of elections, and used results from many countries to make their analysis.

While there are certainly many other factors that play similar role, such as some discussed in the above studies like unemployment, I wanted to use a larger variety of predictors, and focus on some key ones from multiple categories of factors rather than every possible measure. Turnout is a more general factor that has a variety effects on election results. Originally, when I began this project, I wanted to look at turnout from the primary election by party, and compare this to the vote share and winning party. However, this became too difficult to obtain, so I used overall turnout, which would still have an effect on vote share, as we can see below in my analysis.

I also chose to look at some results from key states as a slightly separate section of this study. I was interested to see if I could determine which states are the most predictive of election results based on previous voting patterns. This would be helpful for campaigns, as they would want to spend the most amount of time talking with voters from the states that play the biggest role in determining election results.

I had to make some assumptions going into the project to make my data and analysis a little simpler. First, in comparing the political party of the Presidential candidate to that of the previous President and whether or not the candidate was an incumbent, I got rid of all Presidents that came into office not through an election. For example, there were several Presidents that came into power when they had been a Vice President and their President died. This means that some of the points where I assess whether a President had the same party as the previous President are incorrect, but there are not too many and it would have been significantly more difficult to find the true results. Additionally, in calculating swing states for comparison, I measured which states have been the most accurate over time, even though they might not accurately reflect the states considered to be swing states today. I discuss this further in the analysis section.

Data

Election Results

The first dataset that I pulled was a history of election results since the first presidential "election" in 1789. This data came from Statista[4] in a downloadable form, so I was able to just use the `read_excel` function to read in this data. This data was generally pretty tidy - I split up the column that contained both the President name and year into two columns, took out the percentage signs from the Vote Share columns, and made sure the Year and Vote Share columns were all treated as integers and floats. These Vote Share results would then be treated as my dependent variable.

Previous Presidents

[3] Becher, M., & Donnelly, M. (2013). Economic Performance, Individual Evaluations, and the Vote: Investigating the Causal Mechanism. The Journal of Politics, 75(4), 968–979. https://doi.org/10.1017/s0022381613000959

[4] Citation for data

My next dataset was pulled from Wikipedia[5], so I was able to use the `read_html` package to scrape the table, which included factors about each President. The elements that I ultimately ended up including were just Year, Name, and Party, which I was then able to transform later on to analyze whether the previous President was the same President running and whether the previous President had the same party. To clean this data, I created a function to change the political party name to a shorter version to match that in the state results data. To get rid of Presidents that hadn't been elected into office, I dropped all duplicates by Election Year, and also had to drop duplicates by Name and Election Year to get rid of rows where the President changed party midway through their term - for simplicity, I kept their first party.

Economic Growth

I pulled data on economic growth from Statista[6], so used `read_excel` after downloading. This data was for every year since 1932, not just the presidential years, so I selected the election years by only selecting years that were dividable by 4.

State Votes

For data on which party each state has voted for, I used `read_html` to pull from Wikipedia[7]. This included some additional label columns and rows, which I dropped, and then transposed to have a column for Year and columns for each state. Then, I merged the party column from the Presidential data into this dataset to be able to compare the party each state voted for with the party of the winning President. I then wanted to determine which states had the most accuracy - i.e., which states had voted with the winning party the largest number of times. This would give me a general sense of which states, over time, could be the most predictive of a party winning. I calculated overall accuracy by summing the number of times a state voted for the winning party and dividing by the total number of times a state has been able to vote. I found that New Mexico, Illinois, Ohio, California, Pennsylvania, New York, and Nevada has the highest overall accuracy.

I calculated this a second time as well, this time weighting the 8 most recent elections as 10 times higher to account for the fact that some states began voting a certain way in more recent years. From this, I found that Ohio, Nevada, Pennsylvania, New Mexico, Wisconsin, New Hampshire, and Michigan could be seen as the most accurate current swing states. I then added all of the states, both overall and current to a new data frame as dummy variables, where the value is 1 if in a given year the state voted for the winning party.

Turnout

Finally, I pulled some data on turnout from each year, using `read_html` again[8]. This data was only through 2016, so I manually found the turnout rate from 2020 and added it into

[5] Citation

[6] citation

[7] citation

[8] cite

the dataset.

Total dataset

After all of my data had been pulled and organized, I merged them into one larger data frame. For combining the information on previous Presidents, I had to add a current year column to the previous Presidents data frame, as Year + 4, and then merged based on this new column. Finally, I created some new columns with dummy variables for incumbent status and whether the candidate was the same party as the previous President. In the end, I stored all of these datasets into a SQL database so that I would be able to access them in multiple files.

Analysis

Looking at the data, you can see that there is some missing data from the Popular Vote and Change in GDP columns:

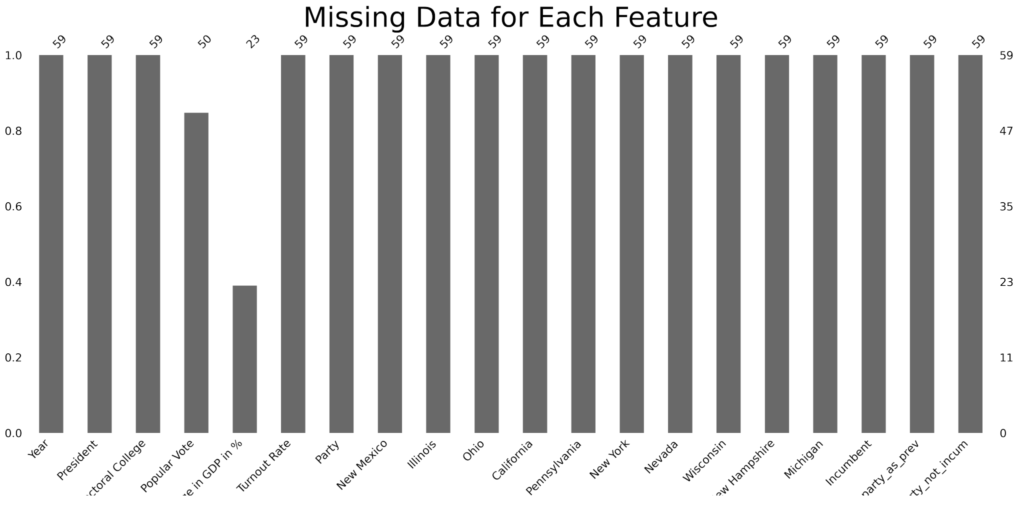


Figure 1: Missing Data

This is the case as these two features weren't really measured until a certain point in time. Since these two variables were pretty important to my analysis, I didn't want to just get rid of them, but also didn't want to only look at rows that had all data present. Therefore, I decided to create a couple different versions of my data to compare results of my models. In general, I used vote share as my dependent variable, but switched off between using Electoral College and Popular Vote as I wanted to be able to look at them both individually. In my opinion, Popular Vote is a better measure of vote share in this situation as it more closely reflects the patterns and decisions of the actual voters, which ties in more closely to my independent variables. However, this wasn't measured until 1824, so I wanted to look at Electoral College as well to have more data points for my assessment. You can see from Figure x below that there were some differences in the trends between the two measures of vote share.

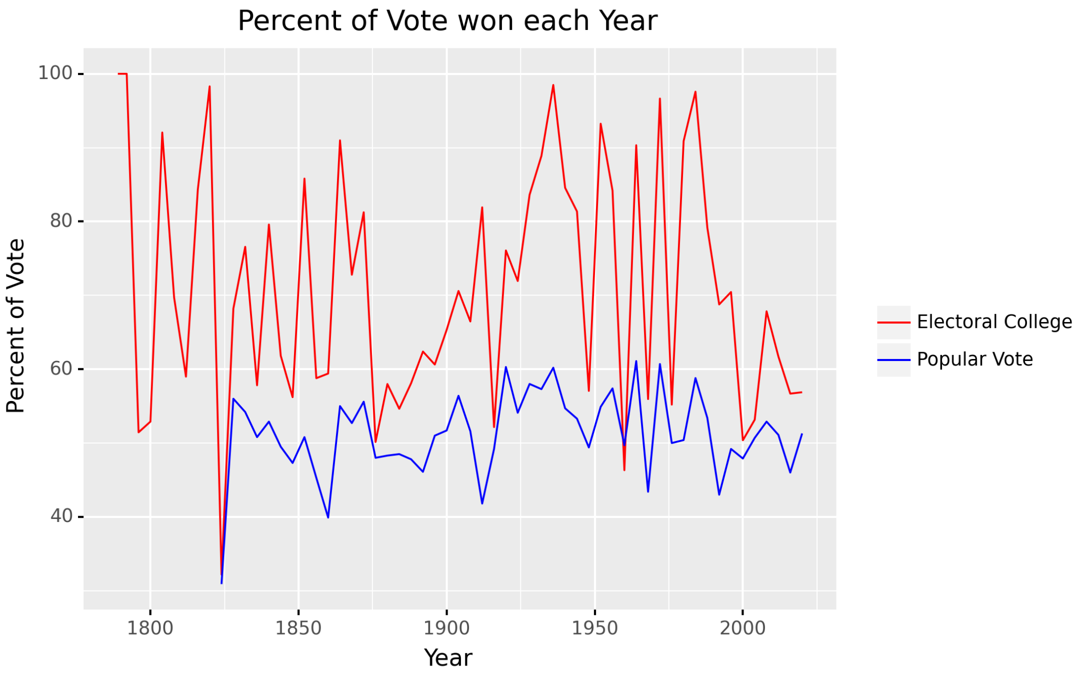


Figure 2: Election Results

I also created a subset that only included the points where change in GDP was measured - I knew that it would be a more limited dataset, but as economic growth has previously been a very significant measure of election results, I wanted to be able to look at it as well.

Finally, I created two subsets where one included the swing states found from a historical view and the other with more current swing states, to be able to see which model was more accurate. As seen from the heatmap below (Figure x), it appeared that the historical states had a larger number of accurate votes by party, but I wanted to be able to run the model and determine if that was correct.

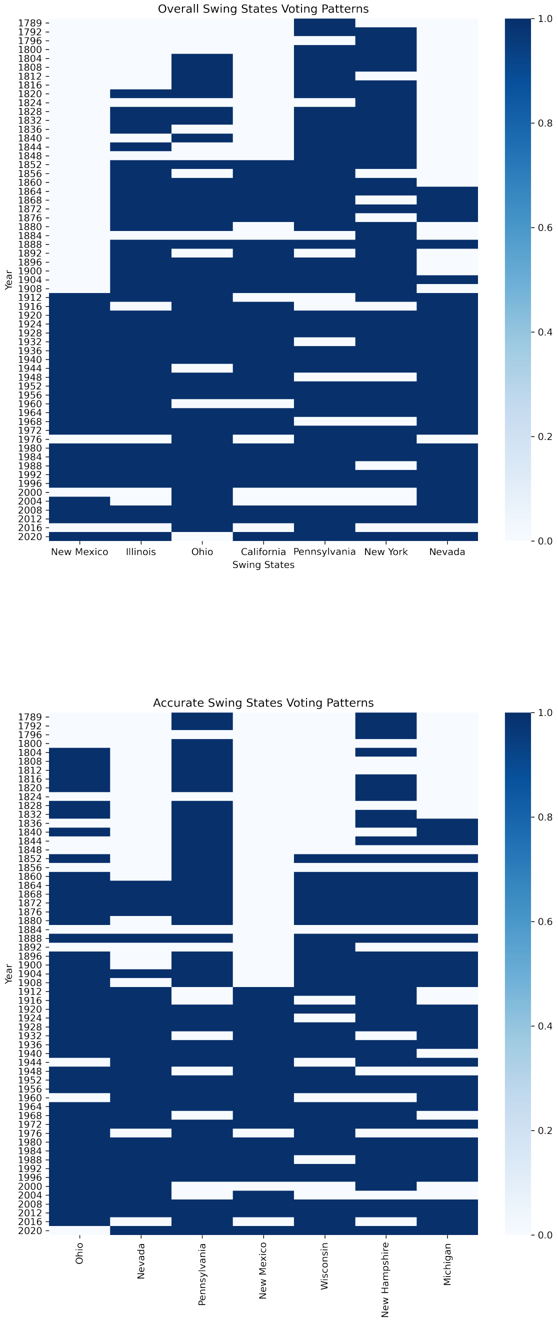


Figure 3: Overall and Accurate Swing States

I then took a look at some of the trends in the variables to get a sense of what my results might look like. The only real continuous independent variables were Change in GDP and Turnout Rate, so I looked to see whether there were any clear patterns there.

It did look like there were some slight trends from these graphs, with Figure (2?) having a negative correlation and Figure (3?) having a positive correlation. I also took a look at some

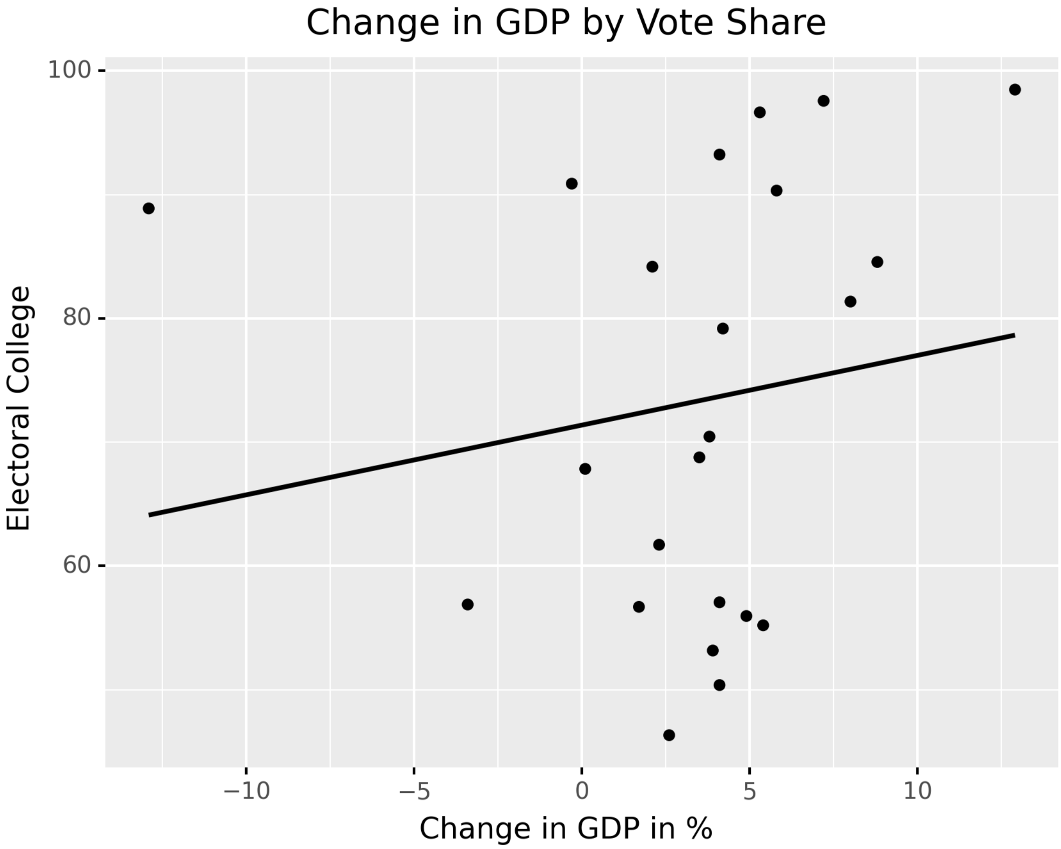


Figure 4: GDP by Vote Share

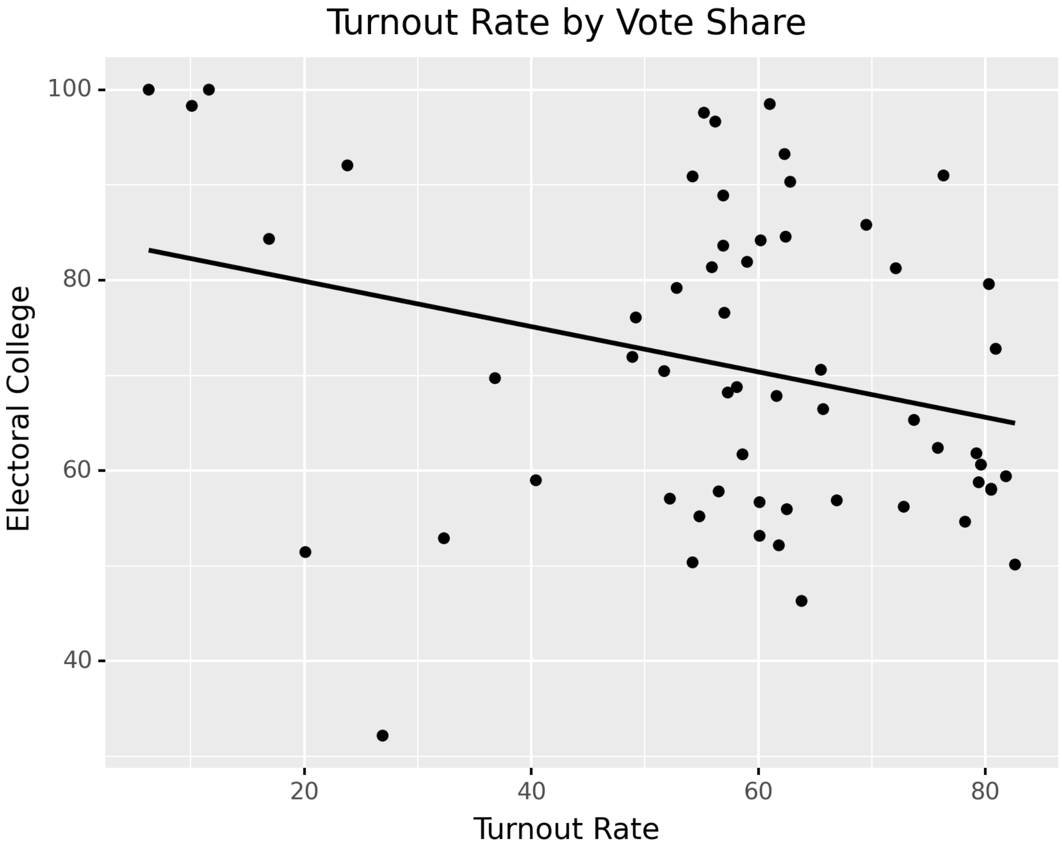


Figure 5: Turnout Rate by Vote Share

of the interactions with the dummy variables, such as how political party and incumbent status interact with the relationship between change in GDP and vote share as seen below in Figures (4?) and (5?).

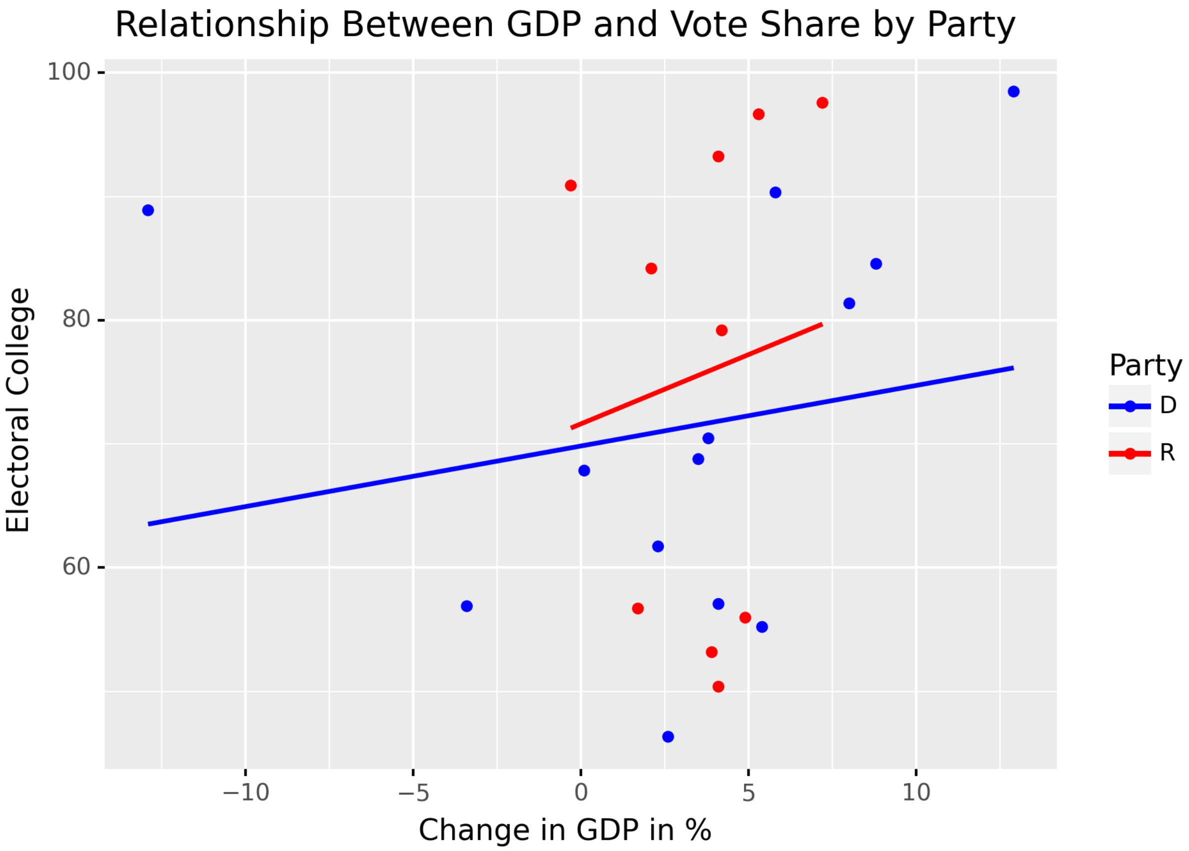


Figure 6: GDP by Vote Share by Party

There were some clear trends here, and you can see how including interaction variables in the model would likely give some more accurate results. In further analyses of this project, I would add in actual interaction variables into my dataset and see how it affects the model.

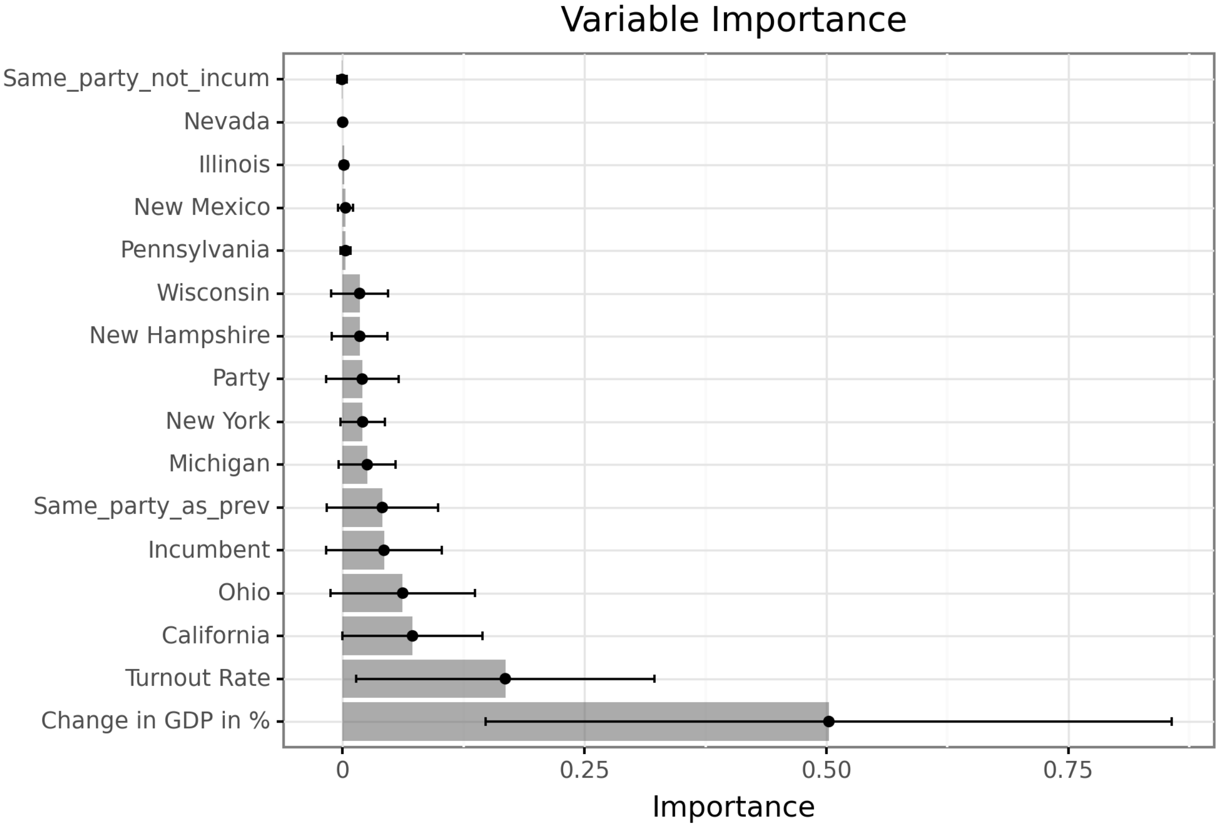
For the actual modeling section of my analysis, as my dependent variables were continuous, I had to choose models that weren't just focused on classification. Therefore, I included a Linear model, a Bagging model, K-Nearest neighbors, Decision Trees, and Random Forests in my search space. The scoring parameter I decided to use is `neg_mean_squared_error` because it is relatively simple to interpret - the lower the score, the better the fit, and the square root of this number will approximately represent the average distance between the predicted values and the actual values of the dependent variable. Because of my relatively low sample size, I split my data into training and testing data with 80% of the data going into the training set and only 20% in the testing set.

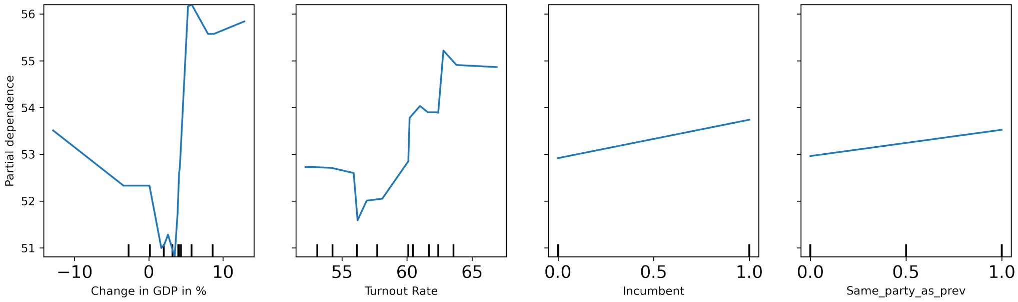
Results

When I ran each of my models using the different subsets I created, I found that there were pretty different levels of accuracy depending on the subset. * Discuss relative scores for each model and how the swing states compared

- talk about the model interpretation, used the gdp data since it was the most accurate

In looking at the features, used permutation_importance with my most accurate model to look at the relative importance of each of my variables





Found that change in gdp, turnout rate, and incumbent were most important, which made sense with previous studies and was pretty cool Looking at the distribution of partial dependence of variables was interesting, had some very non-linear changes

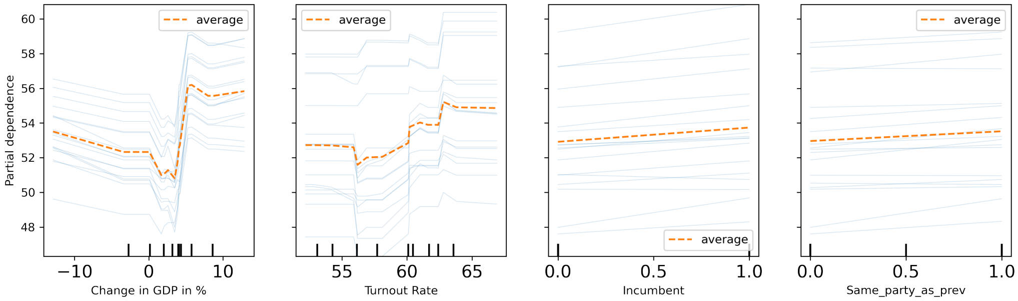


Figure 7: Variable Importance

Discussion

There were definitely some not great values for some of these models, r^2 is pretty low - with more time, I would plot out the relationship between the predicted values and the actual values and see if there are any trends in the years that have the least accurate predicted values - possible that there are some patterns that have to do with the specific features, or maybe current events

What does it mean that gdp data was the most accurate? probably that things have been more consistent in more recent years, and so more discernable patterns

Works Cited