American Voting Trends

Ali Amini John Burzawa Adam Weber

2022-01-12

Abstract

Voting issues are a key part of the decision to select a Presidential candidate. Where Americans stand on key issues and how much weight that holds in terms of who they choose to vote for are crucial to understanding future election trends. Our project aims to focus on how public opinion in the United States shifted from 2016 to 2020. The trend of polarization in the U.S. has been long documented for the last 20 years and this is closely tied to election cycles. Using data from the American National Election Studies, we analyze variables related to approval rating on voter issues and compared that against voting trends for Republicans and Democrats. Our research suggests two main trends. First, behaviors learned from the 2016 election may not be an accurate depiction of voter outcomes in the 2020 election. Second, Americans may have had more polarized opinions from 2016 to 2020.

Our research aims to answer the following question: How do public opinion trends from the 2016 election predict voter outcomes in the 2020 election?

1 Introduction

From 2016 to 2020 there have been a number of unexpected events that shaped the U.S. and political views. COVID-19, devasting wildfires and supply chain problems are only some of the major global events that occurred between Presidential elections but there is one event that seems to eclipse all others in the U.S.. The rise of Trump and far-right Republicans that appear to thrive on extreme views. The shift in Republican views may be part of a larger shift in politics and indicate permanent changes in the political field. Political candidates are only one component of how individuals vote. Politics is a dynamic field that can be hard to determine what is playing the biggest impact of voting trends.

Understanding these trends may inform the kinds of political messaging and strategies useful for politicians in future elections. Additionally, trends in the U.S. electorate are a good indication of how policy outcomes affect public opinion. For example, as voter preferences change, so changes the predictability of election outcomes based on prior trends. If the U.S. is becoming polarized, it may mean we need to allocate resources to this problem to ensure government still works, a more harmonious country. We will explore how much weight voting

issues played in the 2016 election versus the 2020 election and whether the greater trends of polarization in voters is continuing.

2 Trumps effect on the Republican Party

A 2022 survey from the Pew Research Center found that 61% respondents characterized the Republican party as “too often mak[ing] excuses for party members with hateful views” (Pew Research Center, 2022). The former President Trump is known for peddling aggressive rhetoric, lies of election fraud, and other extreme views in the political sphere. These extreme views can be characterized as polarization, meaning support and opposition for key voting issues, and moving away from the middle towards the ends of the political spectrum. This is an issue because polarization affects the ability of government to cooperate and address challenges facing society (Levin, Milner, & Perrings, 2021). A more polarized United States means government that is not serving the people and progress comes to a halt. The government shut down during President Trumps term was the longest in history at 35 days with an estimated cost of $5 billion to the government (Congressional Budget OFfice, 2019). A previous Pew Research survey of 10,000 adults had shown a there was already a trend of increasing divisions between parties even prior to Trump’s rise in politics (Pew Research Center, 2014). This trend of division only seems to have been exacerbated by Trump especially with his endorsements of Republican candidates that mimicked his rhetoric.

Despite Trump’s success in the 2016 election, he was not able to secure a second term. The Republican party followed suit in their political talking points but that was not enough to secure the next election. This begs the question “what changed?”. Our research aim is to examine polarization between 2016 to 2020 and how political opinion on key voting issues may have shaped voting patterns.

3 Data and Methods

Our data was taken from the American Nation Election Study for 2016 and the American Nation Election Study for 2016: Pre- and Post-Election Survey, and the American Nation Election Study, 2020: Pre- and Post-Election Survey (ANES, University of Michigan, Stanford University, 2017) (ANES, University of Michigan, Stanford University, 2021). The studies are a part of the American National Election Study (ANES) and is a time-series collection of nation surveys done continuously since 1948. The 2016 study has a sample size of 4,271 and the 2020 survey has 8,280. Individuals were surveyed two mores prior to the election and then again re-interviewed two months after the election. The survey collects data on demographics, Americans’ basic political beliefs, allegiances, personality trains and other information related to political views.

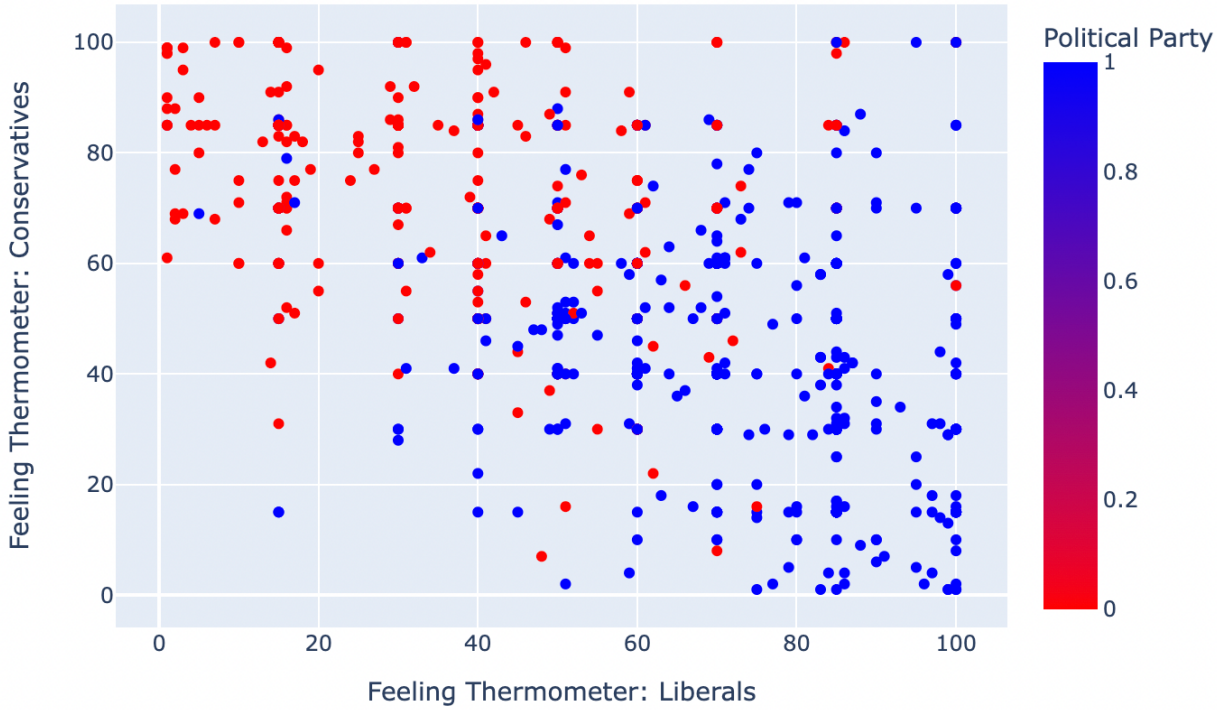
By comparing ANES data from 2016 to 2020 we looked to see if polarization increased. There is no specific survey question that explicitly looks at polarization, but we used opinions on certain political issues as a proxy, some of which are listed bellow:

- Political Party
  Feeling thermometers towards political parties
  Opinions on
  Mexico
  Government health care
  Free trade
- Immigration
- Approval for
  - Congress
  - The President
  - Economy
  - Foreign policy
  - Healthcare
  - Immigration
  - Covid-19 response
- Party affiliation

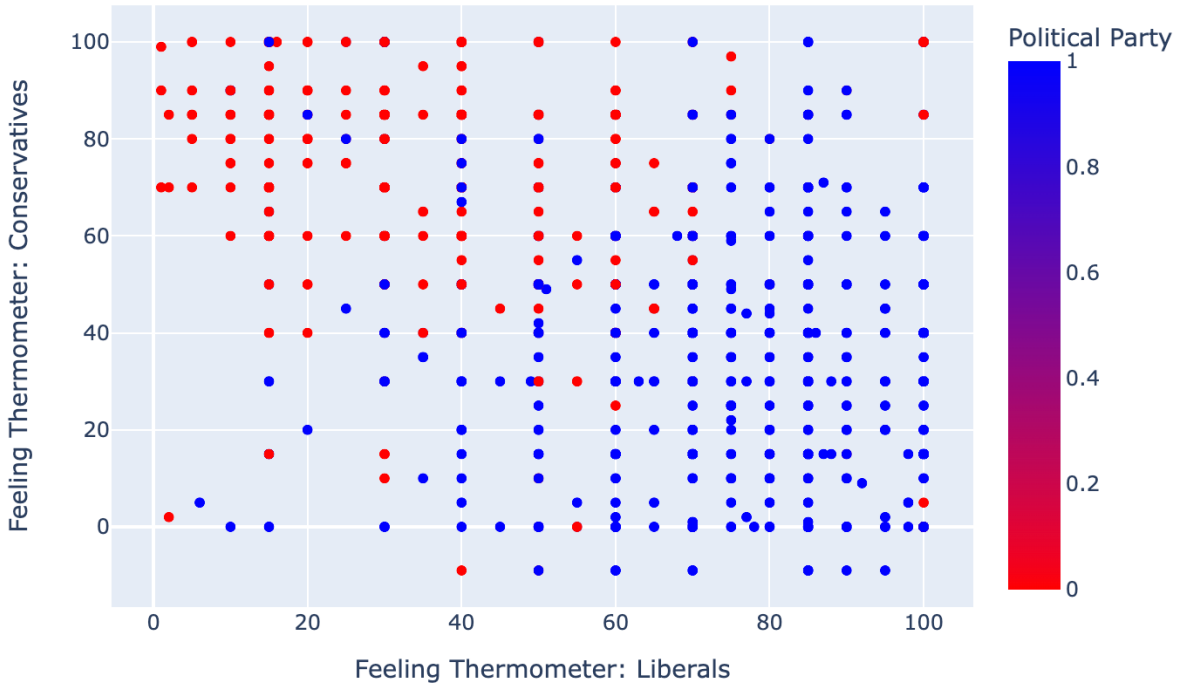
Liberal Vs Conservative Feeling Thermometer

Figures 1 and 2 chart "feeling thermometers" of survey respondents toward liberals and conservatives according to party affiliation. Feeling thermometer scores range from 0-100, with higher ratings indicative of more amicable feelings toward the party, institution, or ideology in question. In 2016, the respondents were prompted to input a number within their range as their response, whereas in 2020, feeling thermometer scores fall in multiples of 5. As you would expect, in both years, democrat-identifying respondents feel more "warmly" toward liberals, and less so toward conservatives. By contrast, republican-identifying respondents most often indicated preference for conservatives as opposed to liberals. Note that these two groups are not mutually exclusive, as respondents are free to assign warm values to both parties. As one would expect, however, this is not how most people chose to respond.

Liberal vs Conservative Feeling Thermometers, 2016



Liberal vs Conservative Feeling Thermometers, 2020



If between 2016 and 2020, the feeling thermometers shift substantially away from one another, then it may indicate that polarization has generally increased. If the points moved closer, it may indicate a reduction of polarization. Unfortunately, due to the shift in scale from continuous in 2016 to ordinal multiples of 5 in 2020, these feeling thermometers were unable to serve the purpose we had hoped in being able to visualize polarization on political party, as there is no clear difference between the two election year surveys.

Supervised Learning Model Random Forest

This project utilizes a supervised machine learning model to predict voter outcomes. Supervised machine learning describes one method of predictive analytics, a subfield of data science which aims to leverage statistical modeling for the purpose of making predictions, usually informed by some empirical intuition. Supervised machine learning models utilize a labeled dataset, in which observations have some set of known features. For example, the ANES dataset used for this project (both in 2016 and in 2020) contains information on how voters answered questions in the survey, such as "for whom did you vote for president?" This survey data has "labels" or distinct outcomes which can be used for prediction. For the sake of simplicity, our analysis only looks at whether someone voted for the Democrat or the Republican candidate in both 2016 and 2020.

Because the ANES datasets have coded variable names, and coded responses, we spent a good amount of time recoding the variable names and the data. Unfortunately, the codebooks for both datasets take quite a bit of time to look through. Thus, we opted to use a semi-cleaned version from the University of Michigan. The offerings from the University of Michigan included Stata data (.dta), which came pre-loaded with all the variable labels. This was particularly helpful as we searched to recode our data for analysis in Python. We were able to use the tabulate command in Stata as a reference in place of the codebook, as it facilitated the recoding process, letting us copy-paste variable codes, and quickly see an interpretable summary of what each value in the dataset indicated. Because our interest in the data involved machine learning, we thought Python may be a stronger tool than R, as the Sci-Kit Learn package is an extremely helpful tool for many different models. We created two lists, one of variable names, and one of variable codes in Python. These lists had consistent indices such that the first item in the variable code list corresponded to the newly created variable name. We then combined the two lists into a dictionary using the variable codes as the keys, and the variable names as the values. This allowed us to rename the columns of interest in Pandas using the df.rename() function, making them easier to work with than the coded feature names. We also subset the data and recoded some of the values in a way that would be both intuitive to interpret as human beings, and appropriate for the machine learning process. We dropped several observations, including those which did not have a vote for either Donald Trump or Hillary Clinton (Joe Biden for 2020), those observations which did not identify with a particular political party, and those observations which did not answer all of our policy questions.

The process of machine learning aims to predict future outcomes using an algorithm which is optimized based on statistical inference. For our case, we aim to predict who Americans will vote for on election day based on their responses to a variety of political questions asked in both of the ANES datasets. In our first model, we used party affiliation (which in this case we only looked at democrats and republicans), as well as approval variables for political outcomes, which take on a value of 1 if the respondent indicates they approve, and 0 if they do not. These particular values include whether or not a respondent approves of "congress handling their job," "the president handling their job," "the president's handling of the economy," "the president's handling of foreign relations," and "the president's handling of health care." These questions were selected because they were worded consistently in both 2016 and 2020, and appear to capture some breadth in different policy areas.

Two models were tested to determine which would be a better predictive tool given our data- the logistic regression model and the random forest classifier. Both models are often used to make predictions for what's seen as a classification question in our research- being able to predict who an individual will vote for can be thought of as asking the question, how do we classify (or predict) voters into groups depending on their likelihood to vote for one candidate over the other. Classification problems can be thought of in terms of those questions which have a finite set of answers, whereas regression problems have a continuous outcome that the model attempts to predict. Since we aim to assess which of two options an individual voted for, we utilize a random forest classifier and a logistic regression. The logistic regression model looks at the probability of an outcome by trying to fit an s-shaped sigmoid function to the data based on the features selected for the model. This algorithm is one of the most used for classification problems, and it outputs the conditional probability of an observation having such an outcome. Then, given the probabilities calculated using the logistic regression method, the algorithm assigns individuals to groups based on that group which has a higher chance of being the correct one. However, the logistic regression model tends to work better using features which are continuous. In the case of our survey data, we have many categorical features which may be better suited for the random forest. The random forest is a type of ensemble method (meaning a combination of several algorithms) which makes predictions through a random subset of the data, fitting several decision trees based on a random selection of features, which aims to prevent the overfitting problem of traditional decision trees. Each "tree" is comprised of junctures at which certain cutoffs within the data are made based on if a value is less than, greater than, or equal to a threshold value. This is a particularly helpful algorithm for modeling nonlinear data. By contrast, one of the assumptions of the logit model is that the features and the outcomes have some linear relationship.

We estimate two models for 2016, which are optimized or "trained" based on a random sample containing 90% of the full ANES 2016 data, and then tested on the remaining 10% to assess how well the model fits. This analysis is conducted in Python using the train_test_stplit() function in the sci-kit learn package. We utilize cross-validation to determine the accuracy score of both models. The accuracy score represents the number of correct predictions (i.e. the

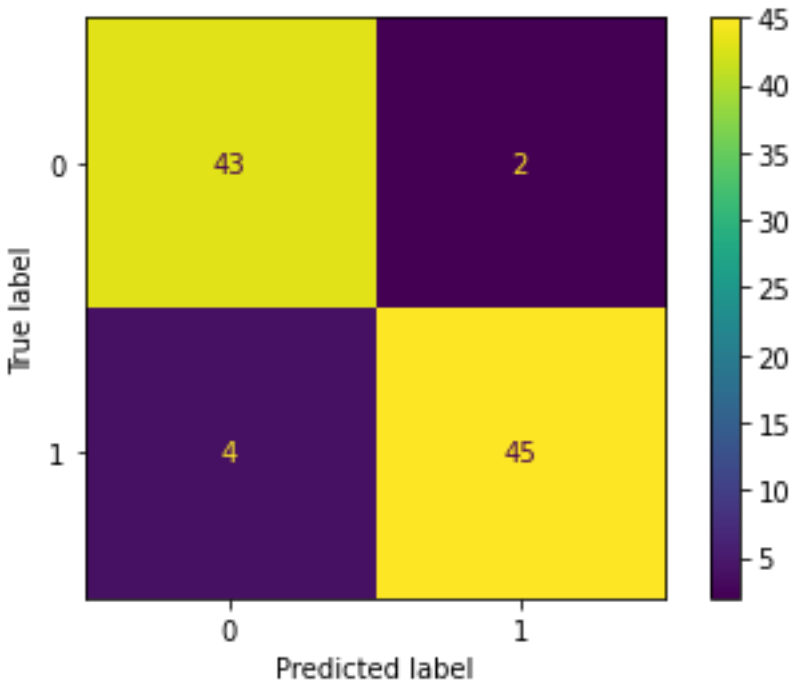
model predicted someone would vote for trump, and they did indeed end up voting for trump) divided by the number of total predictions. Although some scholars have argued that accuracy scores are not the best measures of model validity when focusing on either positive outcomes (for which precision scores are more appropriate) or negative outcomes (for which recall scores are more appropriate), we find that for the purposes of this analysis, we are equally interested in seeing how positive outcomes are affected as we are negative outcomes. The accuracy scores for both models are shown in Table 1 after cross validation:

Table 1

Random Forest	Logistic Regression
.941294	.934946

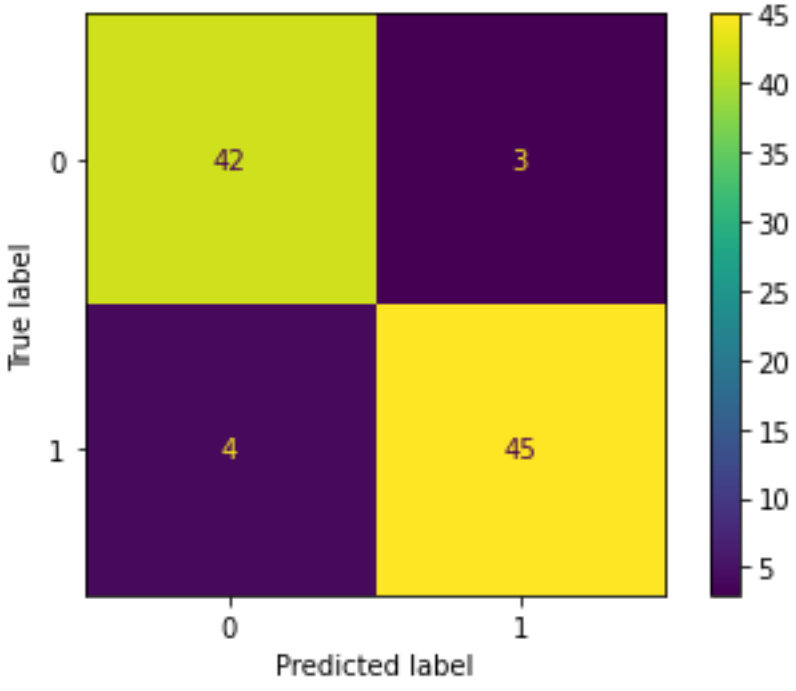
As shown in the table, the two models perform similarly to one another, and with generally a good deal of accuracy. Thus, we move forward in estimating the models for 2016 and 2020 using the same set of features. The confusion matrices for which are seen in figures 3, 4, 5, and 6:

Figure 3: Supervised learning using a random forest to predict votes in 2016



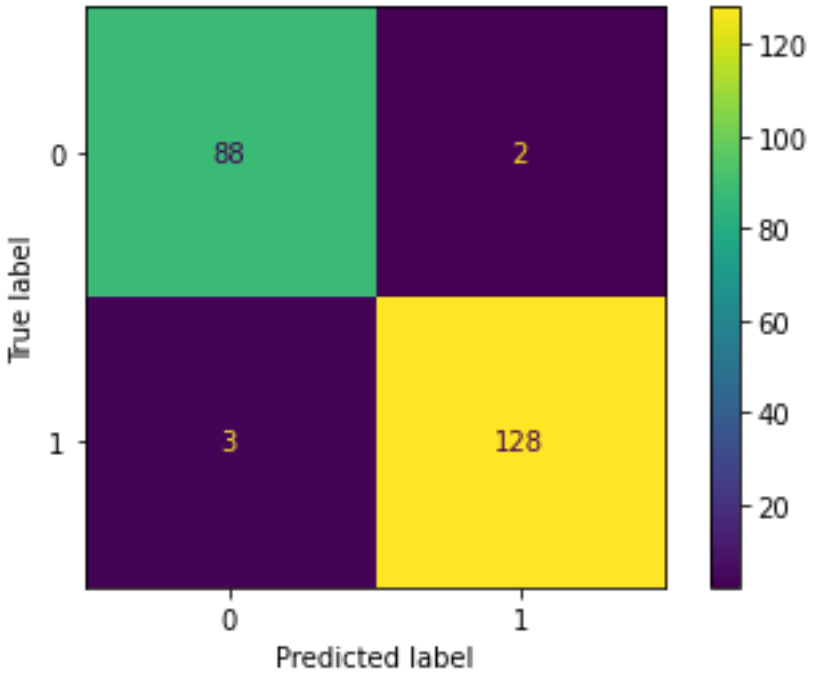
Source: created by authors based on ANES (2016)

Figure 4: Supervised learning using a logistic regression to predict votes in 2016



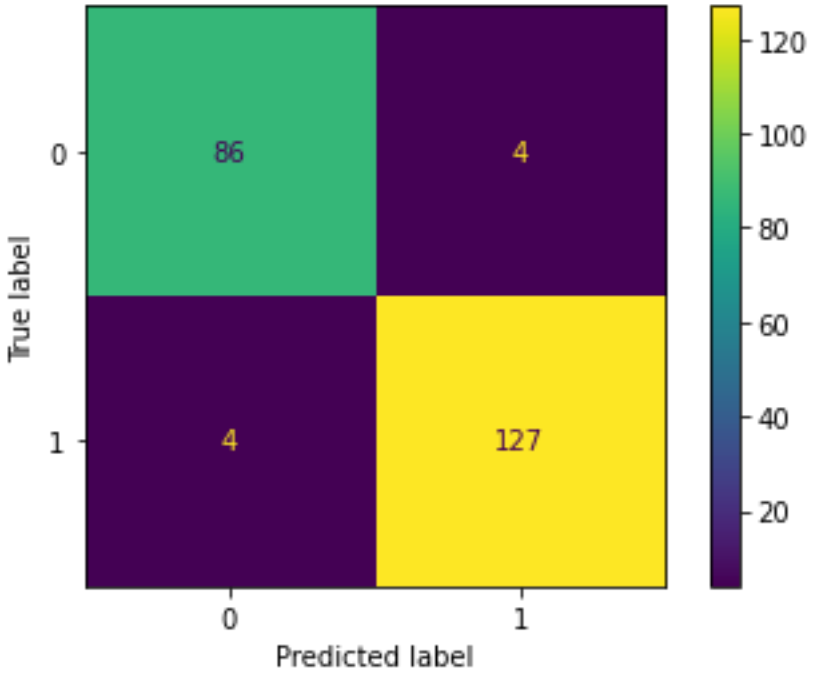
Source: Created by authors based on ANES (2016)

Figure 5: Supervised learning using a random forest to predict votes in 2020



Source: Created by authors based on ANES (2020)

Figure 6: Supervised learning using a logistic regression to predict votes in 2020

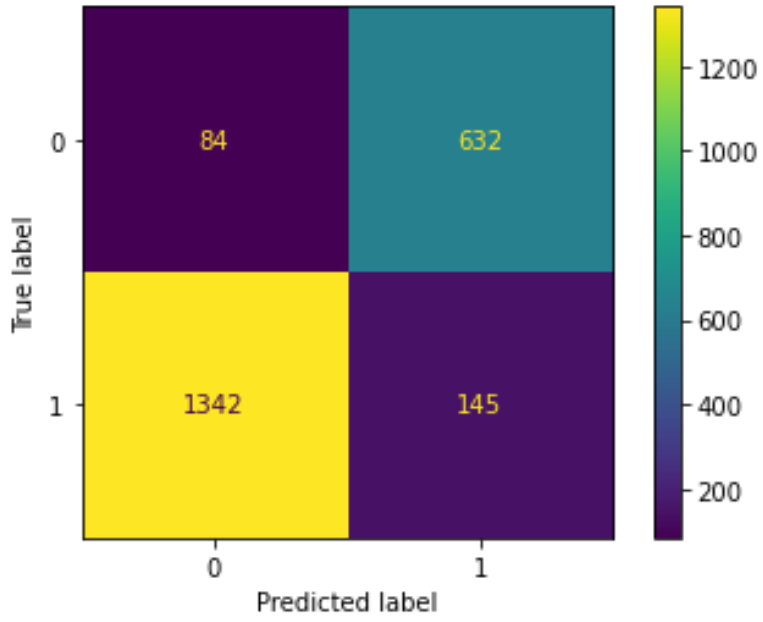


Source: Created by authors based on ANES (2020)

Each matrix is comprised of two axes- one for predicted values and one for the true values of who someone voted for. In both cases, a value of 0 indicates the respondent voted for Donald Trump to become president, whereas a value of 1 indicates the respondent voted for the Democratic candidate (Hillary Clinton in 2016 and Joe Biden in 2020). The number in each quadrant of the matrix indicates how many observations fall into that category based on who the model predicted they would vote for and who they actually voted for. The upper left-hand corners and lower right-hand corners of each chart contain far more observations than the other two, indicating that when the model predicts 1 or 0, far more often than not this prediction is correct, and the true value is indeed the same.

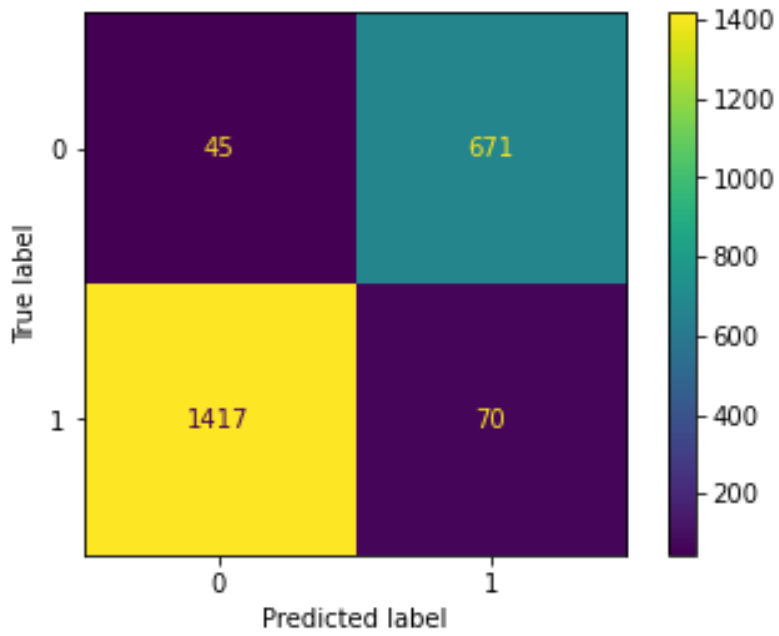
As we are interested in determining the way voter patterns in the United States change over time, we move to assess whether or not the model which was estimated based on 2016 data can be used to predict outcomes of the 2020 election. This is made possible by both surveys containing the same or similar questions, and Donald Trump being coded as 0 in both cases. If the 2016 model fails to accurately predict 2020 voter outcomes, it may be the case that voter preferences or attitudes have changed significantly over the course of the Trump presidency (although there are some limitations to this claim, and other possible explanations which are outlined later). We use the model that was estimated for 2016 to make predictions using data from 2020, and receive the following confusion matrices as a result:

Figure 7: Supervised learning using a random forest classifier to predict votes in 2020 with 2016 model, initial



Source: Created by authors based on ANES (2020) and ANES (2016)

Figure 8: Supervised learning using a logistic regression to predict votes in 2020 with 2016 model, initial

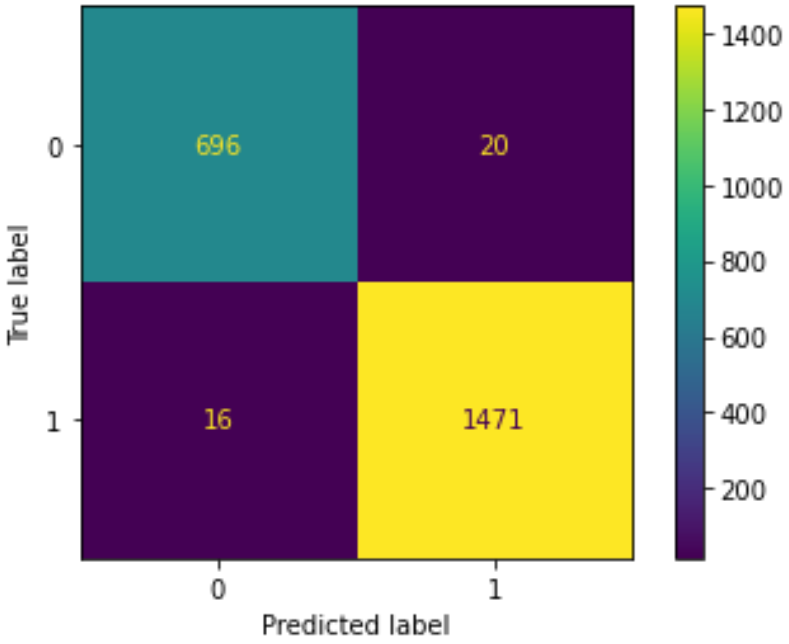


Source: Created by authors based on ANES (2020) and ANES (2016)

As shown in figures 7 and 8, the model performs extremely poorly. In fact, the model performs suspiciously poorly, almost predicting the opposite outcomes for every case. After noticing this almost perfect prediction of the opposite outcome, we review the features of the initial 2016 model. The approval rating variable questions were asked in a way that aims to gauge how voters approved the "president's handling" of X. Since the party of the president switched (from Barack Obama to Donald Trump) from Democrat to Republican in 2016, it is extremely likely that those who did approve of the president's actions in 2016 no longer did so in 2020, and these confusion matrices serve as strong evidence behind this intuition. Thus, we move to select different features from the data which we don't expect to diametrically change from 2016 to 2020. In the case of our second attempt at fitting a model to 2016 voters, we selected opinion questions which gauge voter's attitudes toward immigration policy, party identification, free trade, and government health insurance offerings. We also use feeling thermometers for both liberals and conservatives as features. For simplicity, we only estimate this model using a random forest.

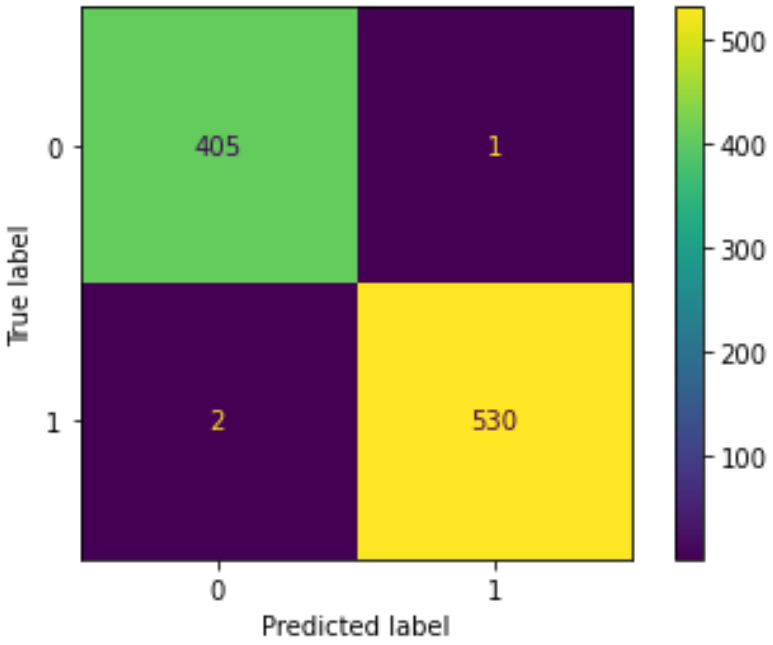
As we have new features, it is important to re-estimate the 2020 model using 2020 data, and the 2016 model using 2016 data as to create a benchmark from which to judge the cross-year model using the same features. The resulting confusion matrices of this procedure are seen in Figures 9 and 10:

Figure 9: Confusion matrix for 2020 predictions using 2020 data and random forest



Source: Created by authors based on ANES (2020)

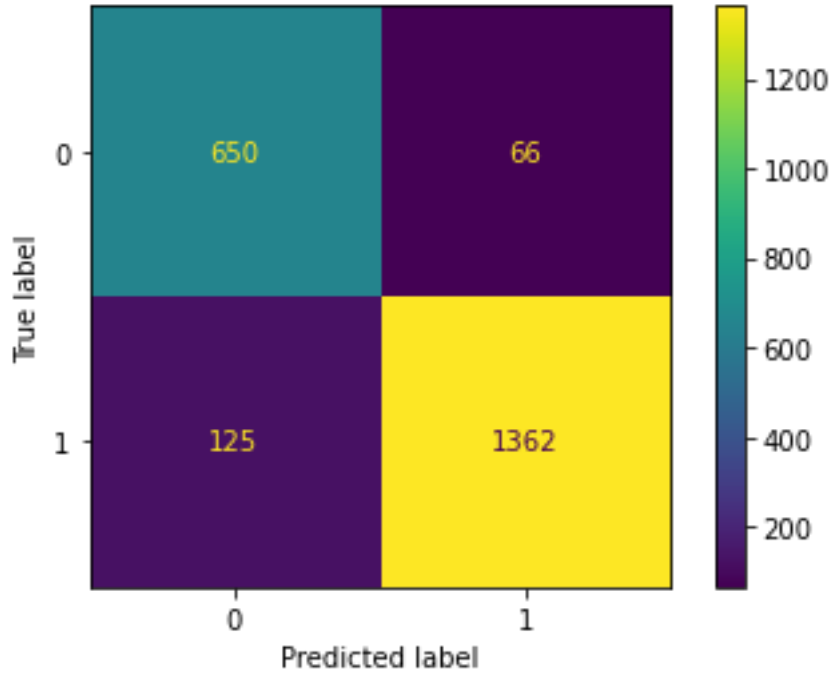
Figure 10: Confusion matrix for 2016 predictions using 2016 data and random forest



Source: Created by authors based on ANES (2016)

Both of the new models do a much better job than the previous estimates, with accuracy scores of 98.3% and 99.6% for 2020 and 2016 respectively. This may be an indication that the features we chose to represent voter outcomes in this second round of estimates were superior predictors of who an individual decided to vote for. Figure 11 shows the confusion matrix for the 2016 model's ability to predict 2020 voter outcomes. As can be seen, the model still predicts fairly well, with an accuracy score of around 91%. However, the same model in the same country with the same questionnaire was not able to predict quite as accurately as it did four years prior.

Figure 11: Confusion matrix of 2016 random forest parameters applied to observations in 2020



Source: Created by authors based on ANES (2020) and ANES (2016)

T-Test and Visualization

We compared political views on free trade, healthcare, Global warming, "building the wall on borders with Mexico", and the president to see if there was a statistically significant difference in attitudes from 2016 to 2020. As mentioned previously we spent a lot of time to clean the data and create new variables for each data set, each party, and each variable.

All of the t-tests show a significant difference between means, meaning we can reject our null hypotheses that there is no difference in means in favor of the alternative (there does exist a difference in means). The T-test results can be found in the appendix. We also used the ggplot2 packages for visualization that can also be found in our appendix. We only mention those variables here and bring one of the outputs here. We show the distributions, the mean, and the

standard deviation for each data set, and for two major parties and NAs. We used na.rm= TRUE in the arguments of our functions.

Additionally, we compared the standard deviations (SD) of the political view responses as a proxy for polarization. Bigger standard deviation would indicate a larger spread in views and more polarization. Smaller standard deviation would indicate the views on political views are more similar and less polarization.

```
#APPROVE OR DISAPPROVE PRESIDENT HANDLING JOB

## # A tibble: 4 x 2
##   pid_chr sd_opinion_persident_job
##   <chr>                      <dbl>
## 1 dem                         1.07
## 2 ind                         1.73
## 3 rep                         1.45
## 4 <NA>                        1.84

## # A tibble: 4 x 2
##   pid_chr mean_opinion_persident_job
##   <chr>                        <dbl>
## 1 dem                         -1.63
## 2 ind                         -0.618
## 3 rep                          1.18
## 4 <NA>                        -0.290

## # A tibble: 4 x 2
##   pid_chr sd_opinion_persident_job
##   <chr>                      <dbl>
## 1 dem                         1.42
## 2 ind                         1.74
## 3 rep                         1.26
## 4 <NA>                        1.76
```

Scale [-2 , 2]

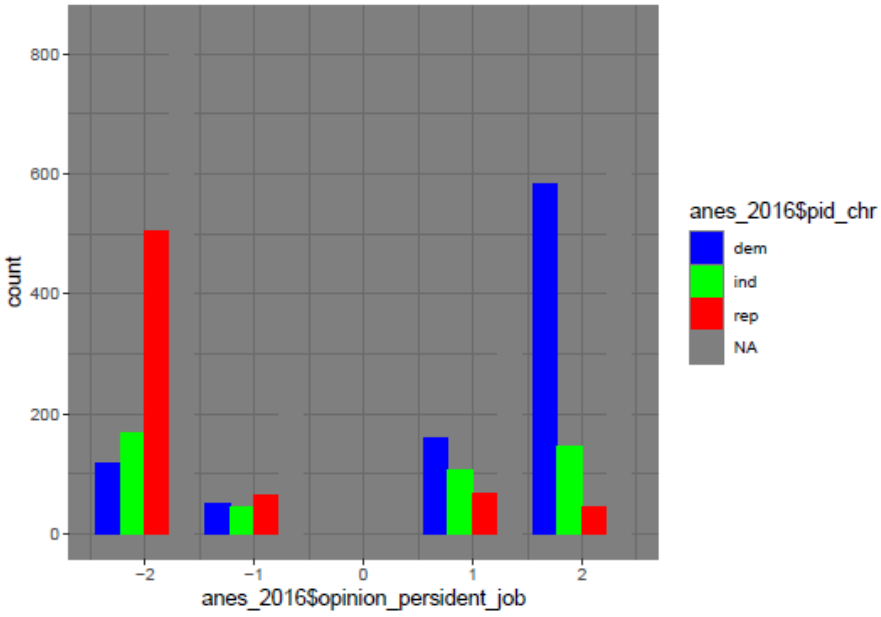
The first tables in this dataset are for 2020, here the first table shows the SD for 2020 which shows a significant SD/polarization in the Republican party, which may mean that Republicans are more polarized for how well did Trump his job in 2020. We can see also a higher polarization in means when we compare the 2020 with 2016 below table between parties.

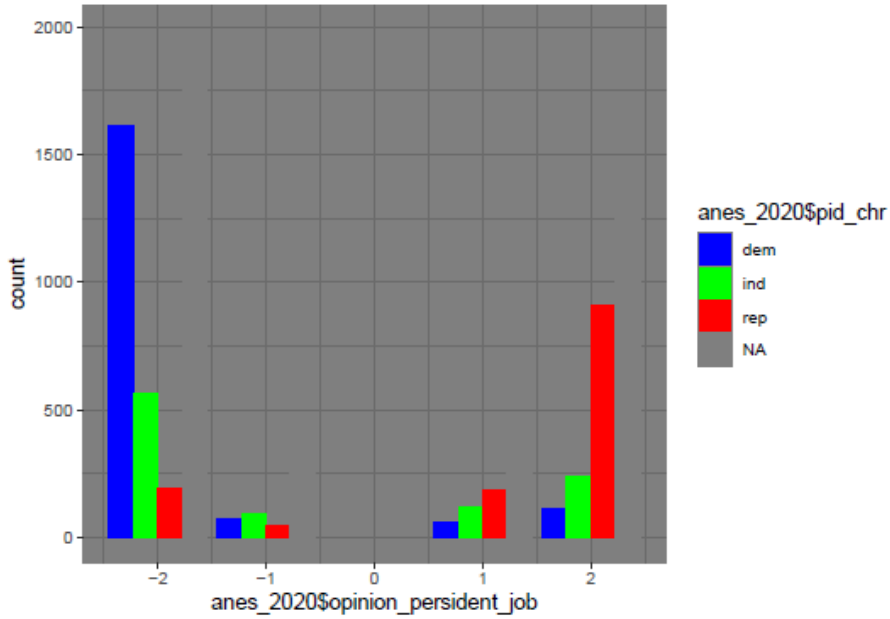
```
## # A tibble: 4 x 2
##   pid_chr mean_opinion_persident_job
##   <chr>                        <dbl>
## 1 dem                         1.14
## 2 ind                         0.0409
## 3 rep                        -1.35
## 4 <NA>                       -0.0718
```

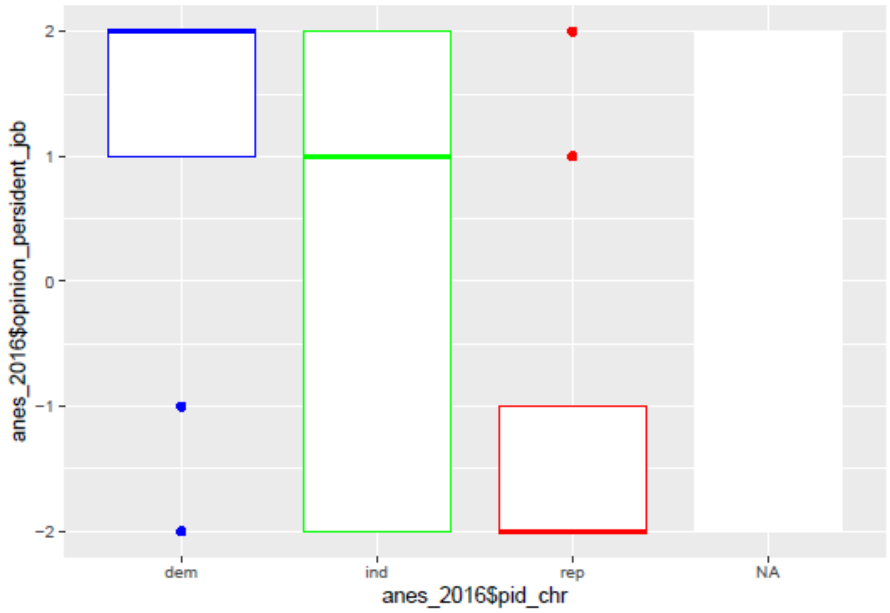
Additionally, we compared the standard deviations of the political view responses as a proxy for polarization. Bigger standard deviation would indicate a larger spread in views and more

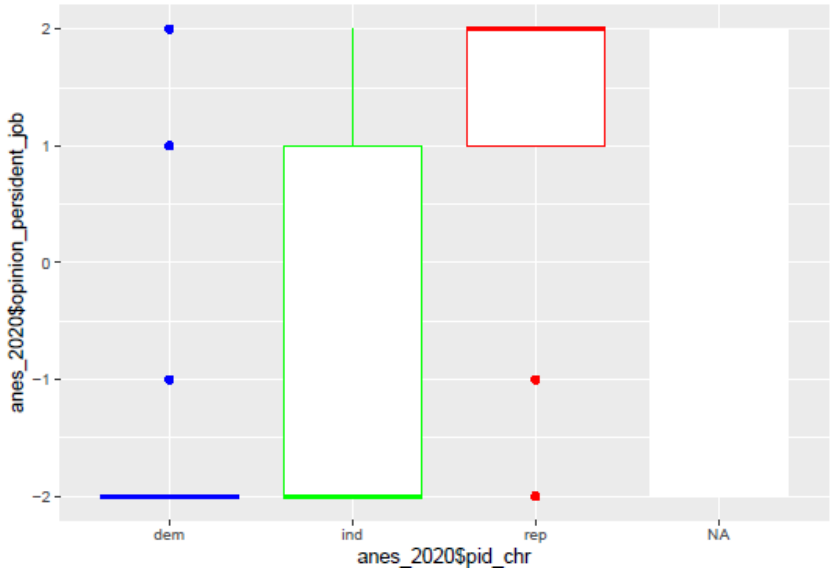
polarization. Smaller standard deviation would indicate the views on political views are more similar and less polarization. This fact can also be seen in the figures.

```
##  Welch Two Sample t-test
##
## data:  anes_2016$opinion_persident_job and anes_2020$opinion_persident_job
## t = 11.482, df = 8712.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3256703 0.4597587
## sample estimates:
##    mean of x    mean of y
## -0.003321471 -0.396035992
```

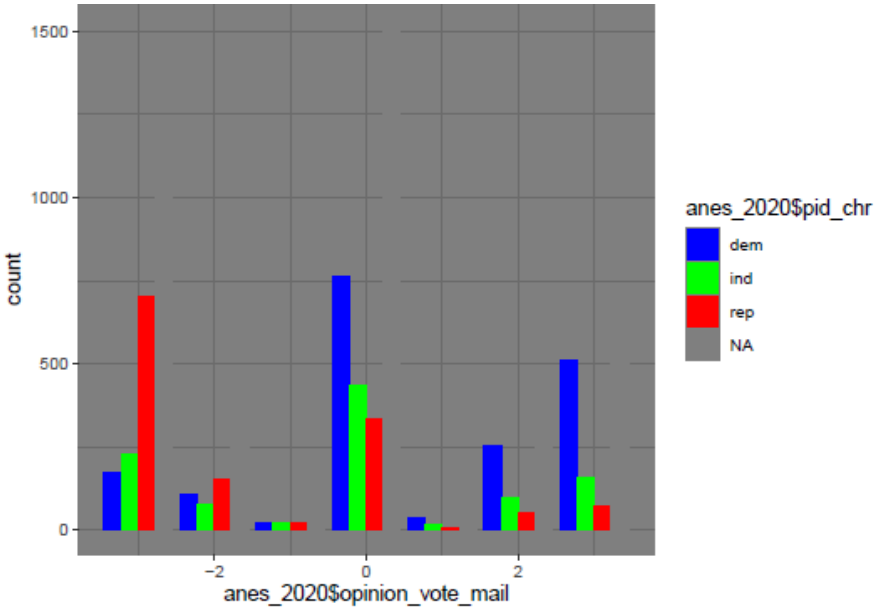



2000
1500
1000
500
0
count
−2
−1
0
1
2
anes_2020$opinion_persident_job
anes_2020$pid_chr
dem
ind
rep
NA


2
1
0
−1
−2
anes_2016$opinion_persident_job
dem
ind
rep
NA
anes_2016$pid_chr



We can also see a big difference between parties in 2020 about voting by mail.



4 Discussion

Our research reviewed the American Nation Election Studies surveys to see how public opinion trends in the 2016 election predict voter outcomes in the 2020 election. Through analyzing voter opinions on key issues, we found that Americans may have become more polarized between elections. We also found that behaviors and opinions in 2016 did not have the same effect on who individuals voted for in the 2020 election. While there are many factors which can affect who someone votes for, due to the hypothesis testing it is evident that 2016 between and 2020 elections, there was a shift in the dynamics of American politics.

One of the major limitations of our research is the sheer complexity and the number of variables that affect public opinion. COVID-19, supply chain issues, geo-political events outside of the United States along with other factors that cannot be measured can all affect political views and the proxies we used to measure polarization. People do not necessarily vote solely on political ideologies and party lines. The candidates themselves may be more, or less appealing to voters. In the 2016 election there may have been sexism against Hilary that did not affect Biden in 2020. While this survey data gives us a snippet of the attitudes towards political candidates, it is only a small piece in a much larger context.

There are some limitations of our analysis with the supervised learning model in being able to justify our claims. For instance, due to time-inconsistency, it may be the case that there is some serial collinearity between election cycles which would lead to variance that is not captured or explained in our model. It also may be the case that some methodological change on the survey level can explain this variability. Imprecise wording or changes in survey question order may impact the way the voters interpret the questions, or feel as they answer the survey, which would change how we expect the data to predict future elections. It may be the case that the way in which public opinion determines voter outcomes has changed- in other words, there is some characteristic of Donald Trump, the Republican Party, the Democratic Party, or the Democratic candidates which changed over the course of the Trump presidency.

For future research on the project the number of surveys included could be expanded. Our research only looked at two surveys, but this data could include all ANES since 1948 to view the larger trends in polarization and voting. This would likely lead to more accurate models that can predict the weight of voting issues on Presidential selection. Additionally, adding more control variables could be used to isolate which voting issues have the most effect on selecting a candidate. We could use LASSO for feature selection, which also might aid in this process. Finally, if there was a way to study similar data for other countries it could give greater insights to how politics work outside of the United States.

References

ANES, University of Michigan, Stanford University. (2017). *American National Election Studies.* Retrieved from https://doi.org/10.3886/ICPSR36824.v2

ANES, University of Michigan, Stanford University. (2021). *ANES 2020 Time Series Study.* Retrieved from https://doi.org/10.3886/ICPSR38034.v1

Congressional Budget OFfice. (2019). *The Effects of the Partrial Shutdown Ending in January 2019.*

Levin, S., Milner, H., & Perrings, C. (2021). The dynamics of political polarization. *Proceedings of the National Academy of Sciences of the United States of America*, e2116950118.

Pew Research Center. (2014, June 12). *Political Polarization in the American Public.* Retrieved from https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/

Pew Research Center. (2022, August 9). *Traits of the parties: Trump and the GOP*. Retrieved from https://www.pewresearch.org/politics/2022/08/09/2-traits-of-the-parties-trump-and-the-gop/

Software

Deepnote (Jurovuch)

Devtools (Wickham, Hester, Chang, Bryan 2022)

Dplyr (Wickham, François, Henry, Müller 2022)

Ggplot2 (Wickham 2016)

Ggalt (Rudis et al. 2017)

Github 3.7.2 (Wanstrath et al. 2022) R version 4.2.2 (R Core Team 2022)

Git 2.39 (Torvalds et al. 2022)

Haven (Wickham, Miller, Smith 2022)

Here (Mulle 2020)

Numpy 1.24.1 (Oliphant 2022)

Pandas 1,5,2 (McKinney 2022)

Plotly Express 4.1 (Johnson et al. 2019)

Python 3.11.1 (Guido van Rossum 2022)

RStudio 2022.12.0+353 (Elsbeth Geranium 2022)

RMarkdown 2.19 (Allaire et al. 2022)

SciKitLearn 1.2 (Cournapeau 2022)

Stata Statistical Software: Release 17. (StataCorp. 2021) College Station, TX: StataCorp LLC.

Tidyverse (Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H 2019)