Team 6 – Jonathan Zaremba, Tyler Kramer, Alex Holland, Maura Glynn

April 29, 2020

Final Project

SSC442 – Bushong

Analysis and Prediction of Factors influencing the 2020 Presidential Election

Abstract

Presented with the ultimate goal of determining the essential factors that lead to a given result of a presidential scale election, we chose to explore two different sub-sections of classifying data. First, we explored the influence of demographic factors at a state by state level and narrowed these features down to religious orientation, race, age, and level of education. Utilizing the most recent polling data available, we created visualizations and conducted an analysis that would aid in determining the extent to which demographics were present in each state and how the typical orientation of the features in regard to political party adherence would be predictable. Ultimately, it was found that religion, race, and education may be polarizing demographic effects as Christian, white, and low educated population tend to swing right where highly educated people and groups of other races and religions are more likely to vote Democratic. However, while age seems to be a large factor in voter turnout, it does not produce a significant trend for state orientation. At the national level, we aimed to inspect which features of national voting and candidates are most influential in predicting an election and aimed to do so. Utilizing a hand-crafted model of specific factors like incumbency and voter turnout, we determined that incumbency of the candidate and party, voter turnout, war, and stock market levels were significant features that determine the outcome of a national election and predicted the reelection of President Trump in the upcoming 2020 election.

State Demographic Analysis

Context

In the progression of analyzing possible election outcomes of every political level it is necessary to first analyze the demographic layout of potential eligible voters. Over past presidential elections, trends have arisen that give way to certain conclusions being drawn regarding which groups of people would tend to identify and support which of the major two political parties, Republican and Democratic.

The most influential of these demographics that may determine voting patterns are identified religion, race, level of education, and age. Christian denomination adherents have an 80% probability of voting for a Republican candidate where theist and other religious minorities have a 90% probability of voting towards democratic. Additionally, white identifying voters have approximately a 62% probability of leaning red where black voters (as well as Hispanic) will vote blue with a probability of over 90%. People with higher education levels, like a post graduate degree are much more likely to vote blue where non-college educated voters tend to vote red and older voters lean red where millennials and younger eligible voters will lean more towards liberal ideals.

In addition to baseline trends like these, one must also consider registration likelihood and voting likelihood for different demographic groups. For example, 63.7% of white adults aged 18 years and over were registered to vote by the 2016 election and only 51.1% of that population voted. Of this, for the population of white adults aged 18 to 24, 30.9% voted where 67.8% of white voters between the ages of

65 to 74 went to the voting booths. For black voters, 60.2% of the adult population were registered to vote and 48% voted. Age wise, black and white voter groups exhibited similar trends in registration and voting regarding age group where older age groups were more likely to not only register, but vote. For the Hispanic population in the 2016 election, the population registration and voting percentage was considerably lower than that of the white and black population. Only 37.9% of adults 18 and over were registered to vote and 28.5% of the overall population participated in voting. That being said, the same trends in age groups persisted that were present in the white and black age groups.

While these demographics outline a small aspect of what goes through each individual voter's mind while determining political support, they provide a useful insight to state and national trends that may aid in predicting state level party affiliation that accumulates to the national election level.

Tasks

Given the trends identified in section (*a*), we deemed it evident to respond to several questions and identify the state by state trends that these aforementioned demographics outline. Primarily, we aim to observe which demographics are most likely to vote for which demographic party and how is this observed in each state as well as how voter turnout is therefore influenced by these demographics. Ultimately, we gathered state level data regarding the most recent polling on religion, race, age, and education percentages for each of the states in the country and this data is then applied to answer these questions.

With what was gathered based on research regarding these main demographic areas, we aimed to gather visualizations that compact the data in a manner by which these areas of interest and questions may be responded to and hopefully answered.

Analysis procedure

After formulating the demographics of interest that we believe would lead the most useful insight to overall state voting trends, the next step that we took was gathering the useful data. A challenge that was faced, however, was that much of the data that was immediately found was either extremely outdated, or not suitable to the analysis which we were aiming to conduct. After extensive research, data was found from sources such as the most recent census in 2010 as well as more recent survey data that was conducted within the last 3 to 4 years. These data sets gave us information regarding population sizes for each demographic as well as relative proportions within each state and comparative numbers between the different subsections of the relevant demographic groups.

After the data was collected, the next step became cleaning the data and sub-setting the variables down to the relevant factors with which we were going to conduct our analysis. This process too proved to be difficult in that when dealing with very large datasets of varying sizes and formats, making the files compatible to not only R studio but to each other became rather complex. Utilizing both the R toolkit and manipulating the data by hand, the data sets were worked down to their essential information and made into a format with which surveying, and visualization can be done on.

Once the data sets were transformed into a workable array of information, we utilized aspects of the ggplot2, tidyverse, and usmap libraries to conduct our analysis and produce visualizations. First, the relevant data was separated by variable per each demographic. Ultimately, these variables were then utilized to produce color scale maps of the continental United States which displayed the intensity by which the specified demographic was present in each individual state (as shown in section (*e*)).

Conclusion and further discussion

Given the produced visualizations, several conclusions can be drawn. First, regarding the religion demographic, it can be seen that a large majority of the populations of Utah, North Dakota, South Dakota, and other areas of the Midwest identify as Christian adherents. Given this and the knowledge that roughly 80% of Christian adherents tend to lean Republican, it can be gathered that these states may be red. Conversely, states such as Nevada, Oregon, and Washington have comparatively low proportions of Christian populations and since 90% of these identifiers lean blue, it can be predicted that these states may vote democratic.

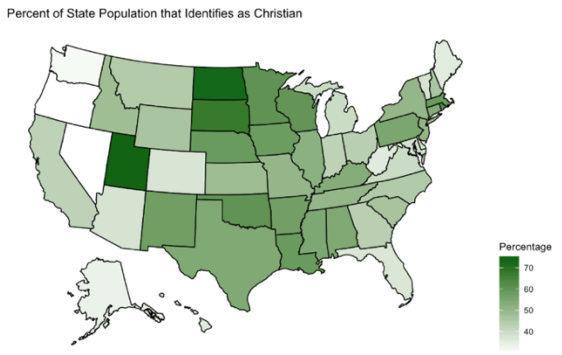
A similar approach may be taken to the race data. States such as Kentucky, West Virginia, and Montana have high proportions of white citizens and this group tends to have a high voter turnout compared to other racial groups and these states are predicted to vote Republican where states such as California and New Mexico with lower white populations may lean more towards Democratic candidates due to more diverse voter populations. Given the information from all three race demographic maps, the states with higher Hispanic and Black populations slightly overlap in New Mexico, Texas, California, and Mississippi, further supporting this prediction.

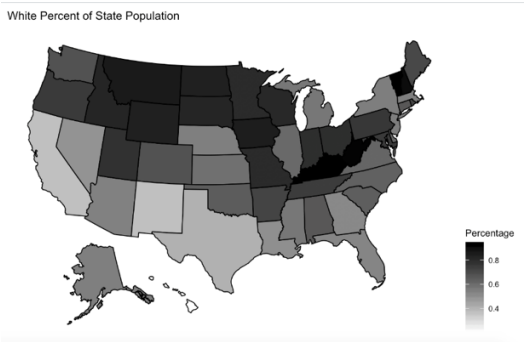
We can see that conferrals of degrees also produce interesting results. Many states who lean both red and blue have low rates of high school degree attainment as seen by our figure below. When comparing this result to the provided map giving the winning candidates' margin of winning, that California and Mississippi had low levels of high school degree attainment – 82.5% and 83.4%, respectively, but relatively margins of winning, with California voting for Clinton by 15% or more and Mississippi voting for Trump by 15% or more. From this we can extrapolate that perhaps the level of education of the population can affect the voting results on both sides of the political spectrum and lead to more polarization among United States voters.

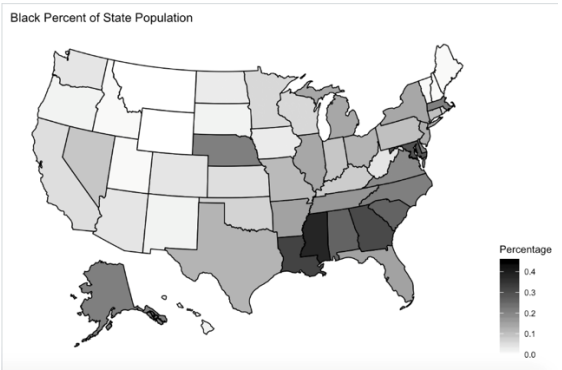
Moving on to attainment of bachelor's and advanced degrees, we can see that many of the states with the highest levels of educational attainment reside in the Northeastern United States. This area also has a very high probability of voting for a Democratic candidate. This makes sense, given that 67% of voters with a postgraduate degree are likely to vote for Democratic candidates. On the other side of the spectrum, the states with the lowest levels of educational attainment reside in much of the Southern United States, an area that has higher rates of voting for a Republican candidate.

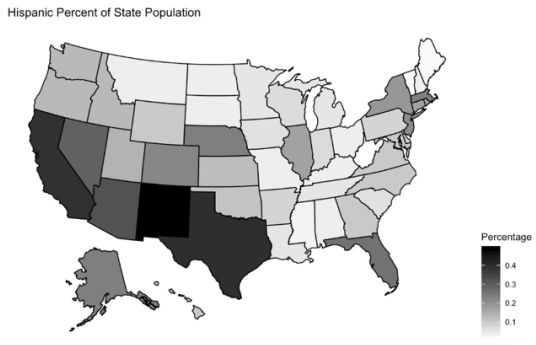
Finally, in an analysis of age in the United States, there was little to be drawn from these results. The lowest average age in the United States was Utah, with a weighted average age of 31.5, which was 3 years lower than the next lowest average age, which was Texas with a weighted average of 34 years, this was followed closely by Alaska and the District of Columbia. All of these areas, with the exception of D.C. are likely to vote for a Republican candidate, although by varying margins. The highest weighted average ages existed in Florida, West Virginia, Vermont, and Maine, ranging from 39.6 to 40.8. This is a much smaller weighted age gap than the younger state counterparts. Additionally, these states are split between voting for primarily Democratic and Republican candidates and have varying margins of winning. This suggests that age may not matter as much as initially thought, or that older populations may have a more distinct political view regardless of the location in which they may reside.

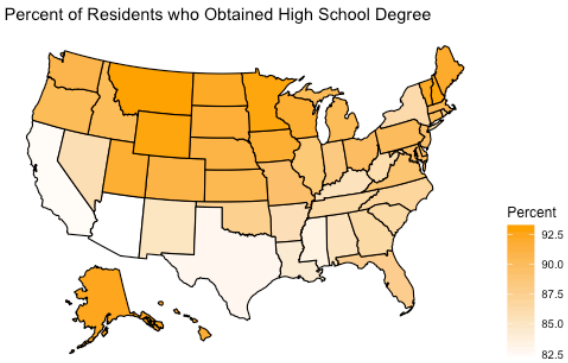
Supporting figures

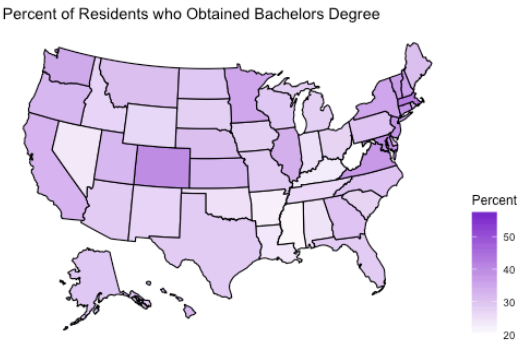


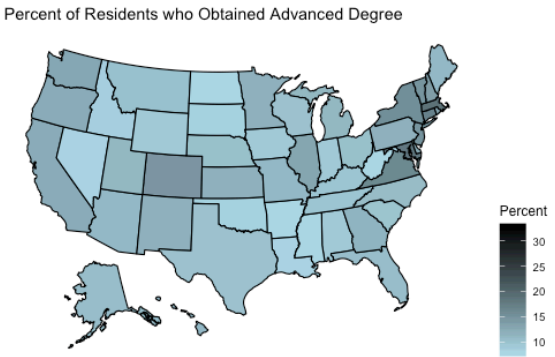


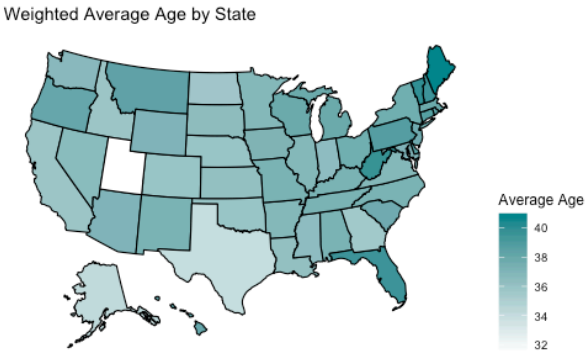


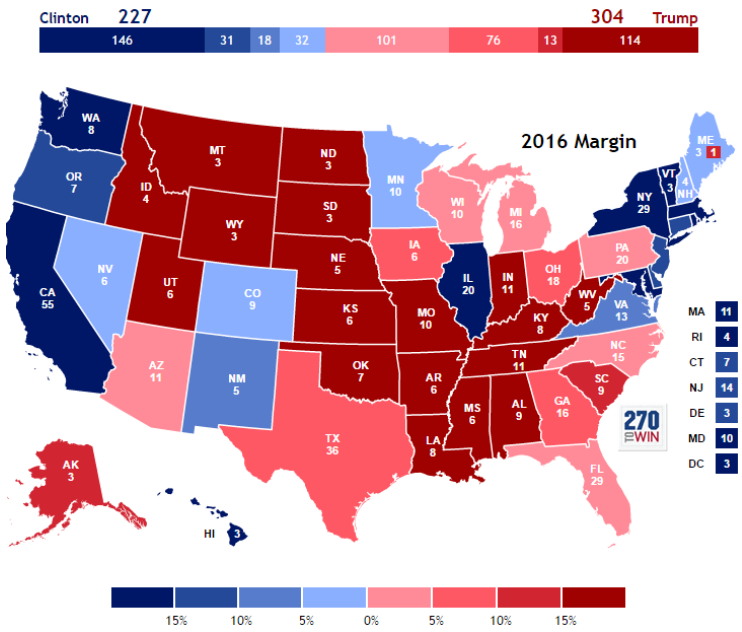












Sources

- https://www.census.gov/library/publications/2011/compendia/statab/131ed/population.html
- https://www.kff.org/other/state-indicator/distribution-by-raceethnicity/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D
- https://www.theatlantic.com/education/archive/2018/11/education-gap-explains-american-politics/575113/
- https://data.census.gov/cedsci/table?q=age&tid=ACSST1Y2018.S0101&t=Age%20and%20Sex%3ACounts,%20Estimates,%20and%20Projections%3AEducational%20Attainment&vintage=2018&layer=VT_2018_040_00_PY_D1&cid=S0101_C01_001E&g=0400000US53,41,16,30,56,01,02,04,08,06,05,09,11,10,12,13,15,17,18,19,20,21,22,23,24,25,26,27,28,29,31,32,33,34,37,35,36,38,39,40,42,72,44,45,46,47,48,49,50,51,54,55_0100000US&hidePreview=true&moe=false&tp=false&d=ACS%201-Year%20Estimates%20Selected%20Population%20Profiles
- https://data.census.gov/cedsci/table?q=Educational%20Attainment&tid=ACSST1Y2018.S1501&vintage=2018&layer=VT_2018_040_00_PY_D1&cid=S1501_C01_001E&hidePreview=false&moe=false&tp=false&t=Educational%20Attainment
- https://www.270towin.com/historical-presidential-elections/timeline/margin-of-victory/

National Factor Analysis

Context

Recently, President Trump likened the COVID-19 crisis to war: "We're at war, in a true sense we're at war, and we are fighting an invisible enemy." It is possible that President Trump is aiming to encourage the public to "rally around the flag" at this moment of national unease. Recent polling has benefited the president, following suit with the historical trend of *war-time* presidents (CNN). This story inspired us to explore whether war or other factors are statistically significant in predicting election outcomes.

Predicting election outcomes has been a subject of interest since elections begin, with actionable insights pursued by those particularly affected by the outcome. Such parties include political party members, who stand to benefit from their party or preferred candidate's occupation of executive or legislative branches of government; business professionals, who might be lobbying for legislation or needing to adjust business plans depending election outcomes; and investors, who often base investment decisions off election outcomes.

Formally studying how factors effect election outcomes gained greater notice with Yale professor Ray Fair's 1978 paper, *The Effect of Economic Events on Votes for President*. Since Fair's paper, many have studied the effect of how various factors (e.g., GDP growth, incumbency) impact election outcomes. In recent years, historian Allan Lichtman's book on determining U.S. presidential election outcomes, *The Keys to the White House*, has become increasingly referenced by political analysts and journalists. In the book, Lichtman discusses thirteen keys: Party Mandate, Contest, Incumbency, Third Party, Short-Term Economy, Long-Term Economy, Policy Change, Social Unrest, Scandal, Foreign/Military Failure, Foreign/Military Success, Incumbent Charisma, and Challenger Charisma. Using Lichtman's model, when fewer than six of the keys are false, the incumbent party is predicted to win the popular vote; the challenging party is predicted to win the popular vote if more than five keys are false.

Our research was inspired by the body of work on predicting election outcomes – particularly the ways in which party incumbency interacts with various *macro* factors. In each election year, journalists and political analysts are often interviewed about the probability of each candidate succeeding in the presidential election. As was recently proven in the 2016 election with President Trump's victory, poll results are not infallible indicators of election outcomes; they are often vulnerable to biases and inconsistencies.

The remainder of this report aims to identify and explore how various factors affect election outcomes. Such factors include the candidate's party, whether the candidate is an incumbent, whether the candidate belongs to the incumbent party, voter turnout, the percentage of the population comprised of various races, whether the United States is at war, the percent change in producer price index commodities, population growth, population density, general strikes, riots, anti-government demonstrations, a weighted conflict index, GDP growth, inflation, and stock market returns.

Tasks

Given the motivations in the context section, we decided to investigate election outcomes by creating a hand-picked model, where we hypothesized a number of variables that we believed would be relevant in predicting the outcomes of elections, ran a regression to determine whether the variables were statistically significant, and assessed how the model performed by training and testing the model on historical election results.

While it can be tempting to use models that over-fit the data because they explain much of the variance in the output, the limited number of observations available to us (US presidential elections only occur once every four years) did not provide enough data to test all the factors we thought could impact the outcome of elections. So we selected the following five factors to use in the regression: incumbent person (dummy variable), incumbent party (dummy variable), the interaction between incumbent party and voter turnout, the interaction between incumbent party and war, and the interaction between the incumbent party and stock returns (using the Dow Jones Industrial Average Index).

We anticipated that a candidate often benefits from being an incumbent (i.e. already occupying the presidency), because most presidents seem to perform well enough in office that the challenger is facing an uphill fight. For a similar reason, we predicted that the candidate membership to the incumbent party (i.e. the prior president's party) will also benefit, on average, as long as people generally approved the incumbent party's handling of domestic and foreign policy. We believe that another indicator of election outcome might be the interaction between the incumbent party and voter turnout. Ostensibly, the most motivated voters are those that are seeking change; we expected them to boost voter turnout, thus hurting a candidate representing the incumbent party. Similar to the *rallying around the flag* during times of national crisis rationale discussed in the CNN article, we predicted that war would benefit the incumbent party candidate, given people's typically higher rates of approval of war-time presidents. Finally, we felt that investors and citizens eying their stock portfolio are likely to be motivated investors if they seek change. We believe that stock market returns will benefit the incumbent party (this variable also likely captures expectations regarding economic growth, as seen in corporate earnings).

Similar to *Lab 4,* we used the cv.glm() function to cross-validate our data from the *boot* library. We used five-fold cross-validation, which randomly sets aside a fifth of the data, trains the model on the remaining data, and evaluates the misclassification ate on the held-out data five times.

In addition to using this method to evaluate the models using the hand-picked variables, we attempted to identify any additional variables with predictive power by looping through all reasonable variables in our dataset (excluding a handful of variables such as the candidates name) and running a logit regression for each one's interaction with interaction with the candidate's party. From there we used our cross-validation technique to identify the variables resulting in the best models.

Analysis procedure

The hand-picked variables in the regression were statistically significant, and signs (e.g. positive) of the coefficients were in line with our expectations. The regression found that the standalone incumbent person and incumbent party variables were statistically significant at the 1% level. The interaction between voter turnout and the incumbent party, the interaction between war and the incumbent party, and the interaction between stock market returns and the incumbent party were also statistically significant at the 1% level.

The incumbent party dummy variable was much more influential than simply being the incumbent candidate – this result was somewhat surprising. It was found that increases in voter turnout hurt the candidate belonging to the incumbent party – this was in line with our expectations and may have been caused by motivated voters increasing voter turnout as they seek a change in the executive branch. The war dummy variable was noted as beneficial to the incumbent party candidate. Finally, the stock market returns were seen to have a large positive effect for the candidate belonging to the incumbent party.

Generating a confusion matrix, we found that our model correctly predicted the outcome in 73% of the testing data. This is encouraging, given the relatively few factors that our model relies on. This can be taken to mean the voter turnout, war, and stock market returns are often indicators of election results. Using five-fold cross-validation, we yielded a prediction error of 0.26 (this is helpful in comparison to the model generated by looping the .

The generalized linear models function – glm() – found that the following variables had the lowest cross-validation prediction errors: population density, general strikes, voter turnout, stock returns, and rise in inflation. In this case, the following variables were found to be statistically significant: incumbent party (positive correlation; 1% level), the interaction between the incumbent party and population density (negative correlation; 5% level), the interaction between incumbent party and voter turnout (negative correlation; 1% level), the interaction between incumbent party and stock market returns (positive correlation; 5% level), and the interaction between incumbent party and inflation (negative correlation; 1% level).

When generating a confusion matrix, we found that this model predicts only 45% of the elections in the testing data. Using five-fold cross-validation, this model yields a prediction error of 0.42 – this value is much higher than found in our hand-picked model indicating a less accurate model.

Conclusion and further discussion

This report found that the following indicators used in the model are statistically significant at the 1% level: incumbent person (a positive effect), incumbent party (a positive effect), voter turnout (a negative effect for the incumbent party candidate), war (a positive effect for the incumbent party candidate), and stock market returns (a positive effect for the incumbent party candidate). We also found using the prediction error looping model that population density, general strikes, voter turnout, stock returns, and rise in inflation explain could also have explanatory power when it comes to presidential elections.

The hand-picked model predicts President Trump's re-election victory in November if things continue on this trajectory. President Trump has a significant advantage given that he is the incumbent person (and party candidate) and the strong performance of the stock market under his tenure (even despite the COVID-19 bear market). If the public tends to agree with his assessment that he is a war-time president, that will be another variable in his favor. Voter turnout will be difficult to assess until election day.

It is important not to mistake correlation with causality. For example, simply seeing the positive correlation between the return of the stock market does not imply that there is a causal relationship between the two. In some cases, presidents inherit economic problems from previous administrations or an increase or decrease of business regulations can be seen as stemming from Congress, not the presidency. A multitude of factors influence stock market movements, and they do not always trace back to the oval office.

A possible avenue for future exploration might be the interaction between specific parties and the variables. For example, it might be found that Democrats benefit from poor economic situations because voters might assume that they will provide more fiscal stimulus than Republicans. Conversely, voters that are looking for more aggressive strategies carried out in foreign wars might prefer a Republican candidate.

There were a number of challenges with predicting presidential election outcomes given how little data there was on prior elections. To find data for variables that were measured consistently throughout the entirety of U.S. election data. We had to settle with an abbreviated time span, 1916-2016. A further area of exploration would include data that is more robust to include additional factors and a longer time period.

Sources

Data:

Cross-National Time-Series Data Archive (1815-2016)

Maddison Project Database 2018

Federal Reserve Economic Data – St. Louis Federal Reserve Bank

Inflationdata.com

Macrotrends.net

https://web.archive.org/web/20111126224208/http://www.aolnews.com/2010/07/12/professors-13-keys-predict-obama-will-get-re-elected/

https://link.springer.com/article/10.1007/s40092-017-0238-2#ref-CR11

https://www.cnn.com/2020/03/24/politics/fault-lines-trump-coronavirus-wartime-president/index.html