Predicting the outcome of the 2020 American election

Smriti Kannan

November 2nd 2020

Abstract

In this paper, we will try to predict the outcome of the 2020 election by constructing a logistic regression model that calculated the probability of individuals voting for Donald Trump. The data used to make the model is from the voters' study group. The model seems to predict that factors such as gender, race and birthplace do have a substantial effect on whether a person votes for Trump. The model seems to predict that it is unlikely for Trump to get the majority of the population votes.

Introduction

2020 is a significant year for U.S politics being an election year. Americans elect a president every four years on the day after the first Monday of November (Flanagan). In the US, there are only two main political parties considered by most voters (BBC News). They are the Democrats - the liberal, left-of-centre party and the Republicans - the conservative, right-of-centre part (BBC News). The Republican candidate this year is the current president, Donald Trump, who won the previous 2016 election (Flanagan). The Democratic candidate is Joe Biden, who was the vice president from 2008 to 2016 under Barack Obama (Flanagan).

A candidate wins the election if they have a simple majority of 270 out of 538 electoral college votes (BBC News). This makes some states more important, as more populous states have a larger number of electoral votes (BBC News). The overall popular vote can be an indication of who wins the election but it is the electoral college votes that determine the election winner (BBC News). It is possible for a candidate to win the election despite receiving fewer votes overall than their main opponent (Flanagan). This is quite uncommon, as it has occurred only four times in fifty-eight presidential elections - but two of them have been quite recent occurring in 2000 and 2016 (Flanagan). The winner of the election is sworn in at the inauguration that is set to take place on the 20th of January 2021 (BBC News).

Data

The data for building the statistical model was taken from the Voter study group, Nationscape data set. The data was filtered so that only the variables of interest were selected. I also created a new variable for race, which just simplified the possible values for that column to only be two. In the data, the race variable has many values - I simplified them into two categories: white and not white. This will make the model a bit easier to understand and more digestible. Another method used to clean the data set was to filter out data points which said "I would not vote" and "No, I am not eligible to vote" for the vote 2020 and vote intention variables. These data values were removed in order to make the model more accurate and also because if an individual cannot vote, we should not include them in our model which looks at the probability of an individual voting for Trump.

Model

I am interested in predicting the outcome of the 2020 American election and will be using a logistic regression model to model the log of the probability of people voting for Donald Trump. The variables being used in this model are race, gender and whether the person was born in a foreign country or in America. The equation for my logistic regression model is:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 X_{race} + \beta_2 X_{gender} + \beta_3 X_{foreign\_born}$$

where: race = whether the person is White or not; gender = whether the person is a male or female; foreign_born = whether the person was born in the United States or not;

$$\hat{p}$$

= the probability that they vote for Trump

Results

Results from the logistic regression model:

Figure 1

```
## # A tibble: 4 x 5
##   term                                 estimate std.error statistic   p.value
##   <chr>                                   <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)                            -1.61     0.0736   -21.9   7.07e-106
## 2 as.factor(gender)Male                   0.455    0.0552     8.24  1.71e- 16
## 3 as.factor(race)White                    1.29     0.0750    17.1   6.34e- 66
## 4 as.factor(foreign_born)Another country -0.110    0.127     -0.870 3.84e-  1
```

Probability calculations from the logistic regression model:

Figure 2

Value of variables selected	Probability of voting for Trump
Female, white, born in US	0.4204
Male, white, born in US	0.5335
Female, not white, born in US	0.1668
Female, white, not born in US	0.1521

Discussion

From Figure 1 in our results, we can see that the factors chosen for our model do change the log odds of a person voting for Trump or not. Gender seems to play a role, with males having a greater probability of voting for Trump than females (0.5335 versus 0.4204 as seen in figure 2). From Figure 2, we also notice that race seems to have a larger influence on whether an individual votes for Trump or not. A female who is white has a probability of 0.4204 of voting for Trump but this is much larger when compared to a female who is not white, who only has a probability of 0.1668 of voting for Trump. The final factor we looked at also seems to largely influence the probability of voting for Trump is whether or not the person was born in the US. A white female born in the US has a probability of 0.4204 voting for Trump. A white female born outside of the US only has a probability of 0.1521 of voting for Trump. Therefore it seems like race and where an individual is born, have a large influence on whether or not the person will vote for Trump. Also by looking at the probabilities in Figure 2, it seems that we can conclude that it is unlikely that Trump will get the majority vote from the population. This means that he is likely to lose the election but as mentioned in the introduction, it is possible to not have the majority vote but still win due to the electoral college votes.

Weaknesses

The model that was built above only looks at the probability of an individual voting for Trump. A stronger analysis would also build a model that models the probability of individuals voting for Biden. The logistic regression model created only taken into account a few factors such as gender, race and whether or not the person was born in the US. A stronger model would take into account a few more factors, such as age, how the individual voted at the previous 2016 American election.

Next Steps

Some statistical analysis that can further improve our analysis would also take into account the State the individual is voting in. As mentioned in the introduction, when it comes to winning the election, it is the "electoral college" vote that counts. Republican vote dominated states are known as "red states" and Democrat vote dominated states are known as "blue states" (BBC News). There are some states known as "swing states", that change hands depending on the candidate (BBC News). With a statistical model looking at states, it would be interesting to look at the data for "swing states" and maybe even compare the model predictions to the previous year votes. This can be done by finding a source for the voting data for the 2016 American election. Therefore maybe additional analysis including states would provide a more accurate and interesting analysis and report.

References

"New: Second Nationscape Data Set Release." Democracy Fund Voter Study Group, 2020, www.voterstudygroup.org/publication data-set. Accessed 29 Oct. 2020.

"IPUMS USA." Ipums.Org, 2010, usa.ipums.org/usa/. Accessed 31 Oct. 2020.

Flanagan, Ryan. "U.S. Election 101: A Last-Minute Cram Session for Interested Canadians." America Votes, CTV News, Nov. 2020, www.ctvnews.ca/world/america-votes/u-s-election-101-a-last-minute-cram-session-for-interested-canadians-1.5169948. Accessed 1 Nov. 2020.

BBC News. "US Election 2020: All You Need to Know about the Presidential Race." BBC News, BBC News, 19 May 2020, www.bbc.com/news/world-us-canada-51070020. Accessed 1 Nov. 2020.