Predicting the 2020 Election and Analysing the Key Factors

Victor ChanYoung Cho, Usman Sadiq

Nov 2nd, 2020

Abstract

In the midst of the 2020 United States Election issue, we took this opportunity to investigate the leading factors and predict its outcome using the Daemocracy Fund + UCLA Nationscape 'Full Data Set' data (Survey data) and American Community Survey 2018 5 year data (Census Data). After filtering out the approrpriate variables and categorising the varying responses, we modelled a logistic regression equation with the survey data, and analysed the significance of the variables. After that, with the logisitic equation model that we have built, we used the census data to conduct a post-stratification process to finally predict the probability that Donald Trump would win the election. Results showed that the variables we have selected (age, gender, household income to name a few) were all significant key factors that may impact the 2020 US election, and we consequently yielded a result where Donald Trump is likely to not win the US Election. Some limitations exist with our analyses however, mainly such that our varying categorical responses were filtered and simplified based on their relative similarity. We address this bias may have impacted our results, and we suggest selecting numerical variables for the regression model for a better result for future work.

Introduction

An overarching issue that rises an immense political controvery is the United States presidential election. In 2016, Donald Trump impressed the political scene by being elected as the Republican nominee for president. He also continuously won the white house votes, which eventually lead to a surprise of being elected as the president, while many observers of the political scene predicted the election of Hillary Clinton. (Jérôme et. al, 2020) Now, with the election being an ongoing event, Trump has already checked on the different parties in the previous years. (Kim & Kim, 2020) With the ongoing presidential debate as well, Trump has also been having his firm stance against Biden, for his another successful election. The US presidential election as we observe is mainly a competition of the current and former president versus the most popular candidate.

Data

We will be using the Democracy Fund + UCLA Nationscape 'Full Data Set' for the multi-level logistic regression modelling and the ACS 2018 5 year data for post-stratification in order to perform our analysis. From the ACS data, we selected the key factors which we believed would have the strongest effect on the prediction of the outcome. The variables are the sex, age, race, education and total income of potential voters.

Model

Our goal for the study as mentioned above is to predict the outcome of the US election, particularly in the interest of Donald Trump's re-election. Using the Democracy Fund + UCLA Nationscape 'Full Data Set' from the Voters Study Group, We will implement a binary logistic model that concerns whether citizens will vote from Trump or Bide, then analyse the proportions of Trump voters and Biden voters.

To supplement the above, we will also be applying the multilevel regression post-stratification technique on the overall census data obtained through IPUMS to yield the expected the overall outcome of the US election. The detailed procedure and outcome with our built model will be described below in the post-stratification section and the results section.

Model Specifics

We have implemented logistic regression to model the proportion of the voters who will vote for Donald Trump. We decided to use this model as we saw the election process as a binomial process. That is, whether people vote for Trump or Biden could be seen as one Bernoulli random variable, and multiple trials of these result in a binomial process. As mentioned in the data section, the variables we have taken into account to model the probability of voting for Donald Trump is gender, race_ethnicity, household_income, education, and age. We present our logistic regression model by the following:

$$log(p/1-p) = \beta_0 + \beta_1 x_{gender} + \beta_2 x_{race_{e}nthnicity} + \beta_3 x_{household_{i}ncome} + \beta_4 x_{education} + \beta_5 x_{age}$$

Where p represents the proportion of voters who will vote for Donald Trump. β_0 represents the intercept of the model. The components consist of β_1 representing the log odds based on gender, β_2 based on race ethnicity, β_3 based on household income, β_4 based on age, and β_5 based on education.

Post-Stratification

In order to estimate the proportion of voters who will vote for Donald Trump I need to perform a post-stratification analysis. The post-stratification refers to the process of adjusting the estimates, essentially a weighted average of estimates from all possible combinations of attributes. This technique is useful as it increases the representativeness of the sample so we have greater confidence in the validity of our inferences about population parameters of interest. Here I create cells based off different ages, gender, education, household income and race. These factors were chosen as they would have a rather large impact on our prediction of the votes. Using the model described in the previous sub-section I will estimate the proportion of voters in each bin. I will then weight each proportion estimate (within each bin) by the respective population size of that bin and sum those values and divide that by the entire population size.

Results

In the analysis, we estimate the proportion of voters in favour of voting for the republicans modelled by a multilevel logistic regression model, which accounted for the age, gender, race, household income and education of the voter.. The results from the analysis are as follows:

```
## # A tibble: 1 x 1
##   alp_predict
##         <dbl>
## 1       0.366
```

We can see that the value we get is 0.366. This means that 36.6% of the voters are likely to vote for the Republican party while 63.4% voters are likely to pick the Democratic party. In order to see if our results are statistically significant, we take a look at the p-values for each of the variables used in our logistic regression model.

```
## # A tibble: 16 x 5
##    term                                  estimate std.error statistic  p.value
##    <chr>                                    <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                             -0.599    0.218      -2.75  5.99e- 3
##  2 as.factor(gender)male                    0.412    0.0617      6.68  2.47e-11
##  3 as.factor(race_ethnicity)black/african~ -1.77     0.225      -7.88  3.22e-15
##  4 as.factor(race_ethnicity)chinese        -0.963    0.380      -2.54  1.12e- 2
##  5 as.factor(race_ethnicity)japanese       -1.07      0.606      -1.76  7.78e- 2
##  6 as.factor(race_ethnicity)other asian o~ -0.204     0.281      -0.725 4.68e- 1
##  7 as.factor(race_ethnicity)other race, n~ -0.405     0.220      -1.84  6.55e- 2
##  8 as.factor(race_ethnicity)white           0.333     0.186       1.79  7.34e- 2
##  9 as.factor(household_income)150,000 or ~  0.193     0.125       1.54  1.23e- 1
## 10 as.factor(household_income)25,000 to 4~ -0.369     0.106      -3.49  4.80e- 4
## 11 as.factor(household_income)50,000 to 9~ -0.314     0.0998     -3.15  1.66e- 3
## 12 as.factor(household_income)less than 2~ -0.428     0.111      -3.87  1.11e- 4
## 13 as.factor(education)bachelors degree    -0.136     0.0825     -1.66  9.79e- 2
## 14 as.factor(education)highschool or less   0.414     0.0816      5.08  3.86e- 7
## 15 as.factor(education)masters or above    -0.127     0.107      -1.19  2.35e- 1
## 16 age                                      0.00913   0.00191     4.77  1.83e- 6
```

As the p-value of each of our estimated coefficient is less than 0.05, the results are statistically significant which means it is very unlikely that the model is incorrect based on these variables.

Discussion

We implemented a multi-level logistic regression model on the UCLA data and then performed post-stratification analysis on the model to predict the proportion of voters likely to vote for the Republican party. We included the factors in our logistic regression which we believed would have the strongest effect on the prediction. These factors are age, gender, race, household income and education. Census data was used from ACS 5 year data for the post-stratification analysis.

The results of our model indicate that the Democratic party would have the more popular vote as their proportion is 0.634 compared to the Republican's 0.366. As the Democratic party is predicted to win 63.4% of the total votes, we conclude that it is very likely that the Democratic party would win the 2020 US Elections

Weaknesses

While our analyses shows plausible results, it is important to understand that the prediction must be taken as a rough estimate. Preparing the data was not easy as there were so many categorical variables. We had to immensely simplify the varying responses of the categorical variables into constrained groupings. For instance, relatively similar responses were grouped into their own categories and data given with a range of values were mutated so the census data and the survey data match up. Furthermore, our model predicts the overall popular vote. However, the US president is not elected by a majority of popular vote. Under the Constitution, the candidate who wins the majority of 538 electors, known as the Electoral College, becomes the next president. (Reuters & The Canadian Press, 2020)

Next Steps

A possible next step in order to improve the model could be to perform a more detailed analysis by analyzing the electoral college votes rather than overall popular vote. Furthermore, it is imperative to compare our analysis results to the actual outcome of the election. Analyzing the difference in actual votes and our estimation would point out the caveats in our model and give us a better understanding for future predictions. Moreover, it would be better if the variables used for the dataset would be selected differently. Due to varying categorical responses, some categorical groupings may be subjective. If the model contained more non-categorical variables, the results would not be biased towards the categories based on relative judgments while simplifying the data.

References

Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from [https://www.voterstudygroup.org/downloads?key=9e0589c5-5ecc-4575-be2a-8285262f97fe.

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

Jérôme, B., Jérôme, V., Mongrain, P., & Nadeau, R. (2020). State-Level Forecasts for the 2020 US Presidential Election: Tough Victory Ahead for Biden. PS: Political Science & Politics, 1-4. doi:10.1017/S1049096520001377

Kim, A., & Kim, P. (2019). Estimation of the 2020 US Presidential Election Competition and Election Stratagies. 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). doi:10.1109/uemcon47517.2019.8992973