Prediction of the 2020 US Election Using Post Stratification (Donald Trump Wins)

Wenxi Li, Jiatong Li, Tianyi Jiang

11/02/2020

Model

Here we are interested in predicting the upcoming vote outcome of the 2020 American presidential election. To do this we are going to employ a post-stratification technique. In the following sub-sections, I will discuss the variable selection process and the model, as well as the post-stratification calculation.

Model Specifics

A logistic regression model using glm() is used here to predict the proportion of voters who will vote for Donald Trump. Logistic regression is appropriate here as the outcome of my interest is binary. We decided to include five predictors in the model: gender, age, race, household income and state since these predictor variables could have a significant impact on predicting the vote.

Racial justice has become a very controversial topic for politicians of the United States. For Trump, the accusation of his racism has surrounded him for quite a while and it is expected that more African Americans tend to vote for Joe Biden.

The state is also a good predictor of the model. History data has shown that some states have more preference for a particular party. For example, citizens in Texas, Oklahoma, Utah, and some other states have been big supporters of the Republican Party while citizens in the state of New York, New Jersey, and California tend to vote for the Democratic party.

Household_income is also a relatively strong predictor in my opinion. Many Trump supporters are known to be "red necks". It features white Americans with blue-collar jobs, earning lower than average/ average income. Individuals with gaps in household income or social class might have a different political preference. We are also interested in how people within different age groups and gender differences in their presidential choice. The following part is the description of the model.

$$log(y) = \beta_0 + \beta_1 x_{agegroup} + \beta_2 x_{gender} + \beta_3 x_{race} + \beta_4 x_{state} + \beta_5 x_{householdincome}$$

Here, y represents the proportion of voters who will vote for Donald Trump. β_0 represents the intercept of the model. All of the predictors are categorical variables. β_1 represents the change in log odds with different age groups; β_2 is the change in log odds with a different gender. Similarly, for a different race, we expect the log odds to change by β_3; we expect log odds to change by β_4 for 50 states and DC. Last but not least, β_5 represents the change in log odds with different (household) income groups.

Additional Information about the model

We further plotted the ROC curve and tested for AUC to check the goodness of fit for the model. ROC curve is the plot of sensitivity (true positive rate) against the false positive rate. We can get a curve with an area under the curve(AUC) between 0.5 and 1. If the AUC is close to 1, we can say the model has a very good discrimination ability. AUC for this model equals 0.7031 and we can conclude that the model has a fairly good discrimination ability.

Post-Stratification

To estimate the proportion of voters who will vote for Donald Trump, we performed a post-stratification analysis. Post-stratification is the process of partition the data into many demographic cells and estimate the response variable for each cell. Then, it sums the cell-level estimates to a population-level estimate by weighting cells by their own proportion in the population. Here, we have a combination of gender (2 categories), race (4 categories), age_group(7 categories), state(51 categories), and household_income(24 categories). Thus it partitions data into 68544 cells. Using the model described previously, we will estimate the proportion of voters in each cell. After weighing each proportion estimate by the respective size of the cell, we will sum those numbers and divide it by the population size.

Results

We calculate the y_hat from

$$y^{\hat{ps}} = \frac{\sum Nj\hat{Y_j}}{\sum N_j}$$

, and we estimate that the proportion of voters in favor of voting for Republicans to be 0.6065. This is based on our post-stratification analysis of the proportion of voters in favor of Republican modeled by a logistic regression model, which accounted for age, race, household income, state, and gender.

Discussion

The overall objective of this study is to predict who will win the 2020 US. election. In the previous part, we first used an Individual-level survey data from Democracy Fund + UCLA Nationscape 'Full Data Set' to carry out a logistic regression. The model is based on their decisions over the two candidates in the previous vote versus some of the variables in the survey data.

Then we used the model we got to further predict the individuals' selection over the two candidates based on a census provided by American Community Surveysamoung. After we did the analysis, we found out that Donald Trump is going to have about 60 percent of the total population voting for him and will be more likely to become the president of the USA by calculating the average likelihood of people in each cell voting for him.

Weaknesses

There are weaknesses in our model. Firstly, during the cleaning process of the survey data, we chose to omit those who said they were not sure whom to vote for in the 2020 selection. But clearly, those people might end up voting for one of the two candidates, and therefore, the final predictions may vary based on these people's decisions.

In addition, there are clearly more variables that might influence one's decision than those we included in our model. These factors may include one's educational background, health situation, etc.

What's more, the number of data sample in the survey data are limited, especially if we are dividing it into a large cell; there may not be enough data to study in each stratum. As a result, the final prediction is not likely to be accurate if we do not have enough data to be modeled.

Last but not least, the final result of the US election essentially depends on who wins more support in different states, and due to this special feature or policy of the election, the model we built and the prediction we made are not perfect at all.

Next Steps

To make the model and the afterward prediction more precise, several next steps could be made in the future.

First of all, as mentioned, we could include more sample data in the survey to build a stronger model at the very first step; even though the cost on both a time basis and a monetary basis would increase.

Second of all, we could include more variables in the model to make the model predict better. To do so, we would have a larger cell, and therefore, to achieve this, once again, we will need more data in the survey to give enough sample data for each cell, as explained above.

Besides, in the future study, instead of dividing it into too many cells, we could focus more on predicting the election in each state, since after all, the final decision is based on the support rate in the 51 states. We could also further investigate what happens to those safe states and swing states this year, i.e, Alabama has always been supporting the Republicans and Michigan supporting the Democrats, etc.

References

1.RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

2.Hadley Wickham [aut, cre],Evan Miller [aut, cph] (Author of included ReadStat code),RStudio [cph, fnd]. Import and Export 'SPSS', 'Stata' and 'SAS' File. https://github.com/WizardMac/ReadStat

3.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). "Welcome to the tidyverse." Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686.

4.Joseph Larmarange [aut, cre] (https://orcid.org/0000-0001-7097-700X),Daniel Ludecke [ctb],Hadley Wickham [ctb],Michal Bojanowski [ctb],François Briatte [ctb]. Manipulating Labelled Data. http://larmarange.github.io/labelled/

5.Matthew Kay [aut, cre],Timothy Mastny [ctb]. Tidy Data and 'Geoms' for Bayesian Models. https://mjskay.github.io/tidybayes/

6.Tobias Sing [aut],Oliver Sander [aut],Niko Beerenwinkel [aut],Thomas Lengauer [aut],Thomas Unterthiner [ctb],Felix G.M. Ernst [cre] (https://orcid.org/0000-0001-5064-0928). Visualizing the Performance of Scoring Classifiers. http://ipa-tys.github.io/ROCR/

7.Paul-Christian Bürkner [aut, cre],Jonah Gabry [ctb],Sebastian Weber [ctb]. Bayesian Regression Models using 'Stan'. https://github.com/paul-buerkner/brms

8.Max Kuhn [aut, cre],Jed Wing [ctb],Steve Weston [ctb],Andre Williams [ctb],Chris Keefer [ctb],Allan Engelhardt [ctb],Tony Cooper [ctb],Zachary Mayer [ctb],Brenton Kenkel [ctb],R Core Team [ctb],Michael Benesty [ctb],Reynald Lescarbeau [ctb],Andrew Ziem [ctb],Luca Scrucca [ctb],Yuan Tang [ctb],Can Candan [ctb],Tyler Hunt [ctb]. Classification and Regression Training. https://github.com/topepo/caret/

9.Douglas Bates [aut] (https://orcid.org/0000-0001-8316-9503),Martin Maechler [aut] (https://orcid.org/0000-0002-8685-9910),Ben Bolker [aut, cre] (https://orcid.org/0000-0002-2127-0443),Steven Walker [aut] (https://orcid.org/0000-0002-4394-9078),Rune Haubo Bojesen Christensen ctb,Henrik Singmann [ctb] (https://orcid.org/0000-0002-4842-3657),Bin Dai [ctb],Fabian Scheipl [ctb] (https://orcid.org/0000-0001-8172-3603),Gabor Grothendieck [ctb],Peter Green [ctb] (https://orcid.org/0000-0002-0238-9852),John Fox [ctb],Alexander Bauer [ctb],Pavel N. Krivitsky [ctb, cph] (https://orcid.org/0000-0002-9101-3362,shared copyright on simulate.formula). Linear Mixed-Effects Models using 'Eigen' and S4. https://github.com/lme4/lme4/

10.Xavier Robin [cre, aut] (https://orcid.org/0000-0002-6813-3200),Natacha Turck [aut],Alexandre Hainard [aut],Natalia Tiberti [aut],Frédérique Lisacek [aut],Jean-Charles Sanchez [aut],Markus Müller [aut],Stefan Siegert [ctb] (Fast DeLong code),Matthias Doering [ctb] (Hand & Till Multiclass). Display and Analyze ROC Curves. https://www.expasy.org/resources/proc

11.Voter Study Group. Nationscape Data Set. https://www.voterstudygroup.org/publication/nationscape-data-set

12.IPUMS USA. U.S. CENSUS DATA FOR SOCIAL, ECONOMIC, AND HEALTH RESEARCH. https://usa.ipums.org/usa/index.shtml

Appendix

all the code used could be found from https://github.com/wethanl/STA304A3/blob/main/A3%20all%20code.R