Prediction on U.S. Presidential Election 2020 Using Logistic Model

Zhen Xia, Jingyao Wang, Ruolan Zhang

11/02/2020

Model

The U.S. presidential election 2020 is now approaching white-hot. The final result of the election between two candidates Donald Trump and Joe Biden would affect the political and economic regulations in America. Therefore, we separately analyzed the votes of both two candidates and made predictions about the results of the election. In the data cleaning section, observations with missing values are removed, as well as any non-qualified voter. Two indicator variables "vote_Trump" and "vote_Biden" are created to display which candidate they voted. Another essential step is to adjust names of variables in survey data and census data into the same, then data from both two dataset can correspond together.

Model Specific

This report uses logistic models and post-stratification to examine how 6 diverse variables, "age_group", "gender", "race", "education", "household_income" and "state", work as influence factors of voting. Logistic models are created for Trump and Biden respectively.

$$y = \beta_0 + \beta_1 x_{age} + \beta_2 x_{gender} + \beta_3 x_{state} + \beta_4 x_{education} + \beta_5 x_{race} + \beta_6 x_{household\_income} + \epsilon$$

β_0 is the interception which represents when all categorical variables are at baseline, the odds of voting for Trump or Biden. β_1 is a dummy variable for age groups. It divides voters by their ages, containing 7 levels from "less than 20" to "above 70". β_2 is the dummy variable for sex. "Gender" has two categories, "male" and "female". β_3 is the dummy variable for states in the U.S. They contain 52 American states which participate in the election. β_4 is a dummy variable for "Education". They classify voters into diverse education levels from low to high. β_5 is a dummy variable for race, dividing into "White", "Chinese", "other asian or pacific islander", "Black, or African American", "Japanese" and "other race". The rest β_6 is a dummy variable for "Household_income" . They record each voter's family income.

Model Check

The area under the curve (AUC) is = 0.691 for vote_T_model. This illustrates that the ability for this model to demonstrate is about 69%. The AUC that we got for the vote_B_model is 0.587. The demonstration ability is not that good for this model since the value is close to 0,5 instead of 1.

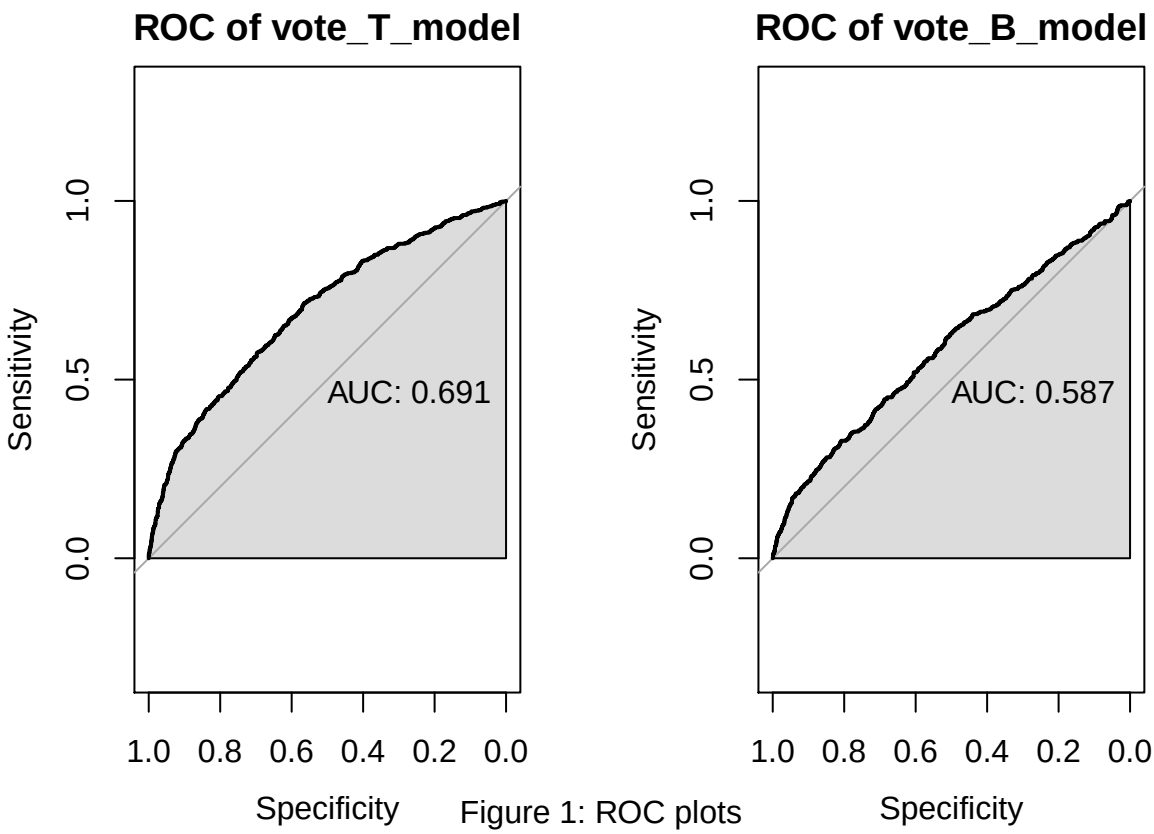

Figure 1: ROC plots

Post-Stratification

The main method of post-stratification is to partition the survey population into cells by different variables. In this survey, we create a total number of 222298 cells from different age, gender, race, education level, household income, and state.Census data is applied to estimate the response variable within each cell to make cell-level estimates upgrade to a population-level estimate by weighting each cell by its relative proportion in the population. Post-stratification is useful for correcting differences between sample and target populations. It can also correct non-sampling errors and lead to less variable estimates.

The variables we used in the estimation are chosen due to their relationship with the characteristics of voters. These factors are gender, age race, education level, household income and the state of the voter. Gender is chosen as a crucial variable in the model. Females and males have different ways of thinking thus need to be discussed separately. In general, males would like to judge a candidate from policies or regulations candidates proposed. More importantly, are candidates' political achievements for this country. In contrast, females, who are sensitive and attentive to details, tend to make evaluations based on candidates' contribution for infrastructures which are related to daily life experiences.

With the rise of feminism, female voters account for more in elections and they will make more independent decisions. Therefore, gender is bound to have an impact on the outcome of the general election. Although the two presidential candidates are males, their views on women would still affect the attitudes of voters of the opposite sex.

Citizens' attitudes towards candidates differ from levels of age groups. Generally, middle-aged crowds are more conservative than young adults. Their expectations towards future lives are obviously distinct from each other. Domestic people who are aged above 18 are considered as qualified voters. Therefore, it is interesting to examine how people in diverse age levels differ in their opinions.

Race issues are controversial in American society. Especially the national awareness raised by the Black Lives Matter movement, and because of how hard voting is being pushed. This would be a problem for both white people and colored people to address.

Another non negligible factor is Education-level. Trump's campaign policy includes providing school choices for every child in the USA and pushing the idea of American exceptionalism. These policies have very different levels of impact on people from different education levels.

What is more important behind the factor of Household income is the huge impact caused by income inequality. The voters could be sensitive to their own economic situation, and furthermore, to the economic policies promised by presidential candidates. Therefore, people with different incomes will treat elections differently.

Due to the special electoral system in the United States, presidential candidates tend to focus on certain states. Obviously, the state of the voters would be a very important factor.

Result

Based on the post-stratification analysis, our estimation of the proportion of voters in favor of voting for Donald Trump is 0.3906846 and that of Joe Biden is 0.4376753 accounted for age, gender, race, household income, state and education modelled by 2 separate logistic models. From the result of our estimations, we can predict that Joe Biden is more likely to win the popular vote in the 2020 American federal election. Focusing on the factor "race", White people and American Indian or Alaska Natives tended to vote for Trump instead of Biden, about 49% and 44% of these voters respectively in favor of Republican Party. On the other hand, most Asians, Black or African Americans are more willing to vote for Biden. In addition,the vote_T_model has Area Under Curve(AUC) equals 0.691, which means this model has 69.1% probability to make the true prediction. Generally, AUC displays the overall diagnostic accuracy of the test, the value here is calculated as 0.691 for the model of Trump, the conclusion can be drawn that this model has a relatively good discrimination ability.

Table 1: Proportion of vote based on Race for Trump

race	alp_predict_t
American Indian or Alaska Native	0.4426272
Black, or African American	0.1179371
Chinese	0.2127651
Japanese	0.2561780
other asian or pacific islander	0.3740386
Other race	0.3262108
White	0.4915463

Table 2: Proportion of vote based on Race for Biden

race	alp_predict_t
American Indian or Alaska Native	0.4426272
Black, or African American	0.1179371
Chinese	0.2127651
Japanese	0.2561780
other asian or pacific islander	0.3740386
Other race	0.3262108
White	0.4915463

Discussion

Survey data in 2020 is downloaded from IPUMS website and census data is downloaded in 2018 Voter Study Group website. Logistic regression model is applied to predict the proportion of American people who will vote for Donald Trump. Variables used for building the logistic regression models are age, gender, race, education, household income and state.Then we use the post-stratification technique, the census data is partitioned to 222298 cells to estimate the national vote result in 2020. The predicted percentage of voters are obtained from the population by applying cell estimation on census data. Based on the estimated proportion of voters in favour of voting for Republican Party being 39.1% while the proportion of voters in favour of voting for Democratic Party being 43.8%. By this estimation, Joe Biden is predicted to win the 2020 American Federal election. The cell estimation is grouped by states and races separately as well, to find out the tendency of voters with these certain characteristics. The group result shows that Trump and Biden both have advantages in nine states. This supports our results, that is, Trump and Biden's expected votes are fairly close, although Biden leads with only a slight advantage of 5%. More white people and American Natives are supposed to vote for Trump instead of Biden This tendency is reasonable, since one of Donald Trump's political ideologies is to protect domestic citizens, but not propose social acceptance towards foreigners. Therefore, domestic people are more willing to choose him. In contrast, most people, who are not locals, tend to vote for Biden. According to Biden's political concept, he protects immigrants' and their generations' rights, as well as admitting their values. Interaction with Asian countries and investing in smarter technologies are necessary for the United States.

Weakness

Non-response rate and missing data rate are still problems that cannot be ignored. In survey data, only 66% data is kept and in census data, 97% data kept. This means the model we build from survey data cannot represent the majority.

In this case, each observation satisfies the assumption, which is, variables should be independent of each other. These assumptions would simplify the model while being troublesome if leading to multicollinearity. The above assumptions might lead to a procession error, that is, a non-sampling error.

In addition, the survey which the survey data comes from was performed five months ago. Due to the special impact of this year's COVID-19 situation, this data is possibly out of time.

The census dataset we used in the analysis is the 2018 data, it might not match the real situation in 2020.

In the survey data collecting process, it is assumed that the winner of the state's vote receives all of that state's electors. However, things are different in Maine and Nebraska.

Next Step

We can increase the sample size to make survey data more representative. At the same time, putting more gender options other than man and woman in the survey would also be a good choice.

Right now, we are using SRS. In the future, we could use more advanced sampling methods like Multi-Stage Sampling. In our future analysis, we can try to use multilevel regression models.

The2020 census dataset should be used if it was available.

Reference

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Men and Women in the U.s. Continue To Differ in Voter Turnout Rate, Party Identification Ruth Igielnik - https://www.pewresearch.org/fact-tank/2020/08/18/men-and-women-in-the-u-s-continue-to-differ-in-voter-turnout-rate-party-identification/

Trump, Biden Offer Dire Warnings in Saturday Campaign Stops As Early Voting Continues To Set Records John Colby Itkowitz - https://www.washingtonpost.com/elections/2020/10/31/trump-biden-live-updates/

Receiver Operating Characteristic Curve in Diagnostic Test Assessment Jayawant Mandrekar - https://www.sciencedirect.com/science/article/pii/S1556086415306043

Wang., et al., Forecasting elections with non-representative polls. International Jorunal of Forecasting (2014), http://dx.doi.org/10.1016/j.ijforecast.2014.06.001

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Paul-Christian Bürkner (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software, 80(1), 1-28. doi:10.18637/jss.v080.i01

Paul-Christian Bürkner (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal, 10(1), 395-411. doi:10.32614/RJ-2018-017

Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77 http://www.biomedcentral.com/1471-2105/12/77/

Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1. https://CRAN.R-project.org/package=kableExtra

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

Appendix

Code and data supporting this anaylysis is available at: https://github.com/Wang-Lucy107/Prediction-on-U.S.-Election-2020