Forecasting U.S. Presidential election 2020 with non-representative polls

Heran Zhou, Xinyu Zhong, Yaqi Feng, Yuhan Gu

Nov.02 2020

Model

Multilevel logistic regression model is utilized to predict the election results. The survey data was partitioned into 14 cells, with each cell representing an interaction variable of gender and race. This is because gender and race are strong predictors of the election results as determined by our frequency analysis. Furthermore, age is treated as a random intercept to capture the impacts among different age groups and reduce the model variance and standard error.

Model Specifics

The model has an expression as follows:

$$Pr(Y_i \in Trump, Biden) = logit^{-1}(a_m + a_{j[i]}^{labor} + a_{j[i]}^{householdincome} + a_{j[i]}^{state})$$

Where a_m is a random intercept representing an individual's voting intention impact from his age group and m is the individual's corresponding age group. The expressions $a_{j[i]}^{labor}$, $a_{j[i]}^{householdincome}$ and $a_{j[i]}^{state}$ indicate the coefficients for each level 1 categorical variable, and the footnote j[i] represents the cell that the ith respondent belongs. Labor, household income and state are the predictor variables because we assume they have relatively strong correlations with an individual's voting intentions. Such variables exist in both the survey and census data.

The model is run through R studio. To adjust and transform the data on hand into accurate estimates of the 2020 election results, we use the approximate marginal maximum likelihood estimates obtained from the glmer() function in the R package lme4 (Bates, Maechler, Bolker & Walker, 2015). Most of the variables were excluded during the data cleaning process because they simply do not exist in both the survey and census data set.

The Area Under the ROC Curve ("AUC") is employed to measure the model performance since it can most effectively distinguish between two diagnostic groups under a logistic regression model.

Post-Stratification

Post-stratification, in plain language, corrects for any imbalance in the representation of the groups, or strata, in the sample obtained, knowing that the population is composed of distinct groups that differ with respect to the estimating quantity of interest. This technique is useful because it not only allows for more accurate estimates of population quantities to be obtained in the context of survey sampling, but also corrects non-sampling errors as well. (Reilly, Gelman, & Katz, 2001)

We calculate the post-stratified proportions of voters voting for Biden and hence the democratic party by the expression as follows:

$$\widehat{y}^{PS} = \frac{\sum N_j \widehat{y_j}}{\sum N_j}$$

$$\widehat{y}^{PS} = \frac{\sum\limits_i [if(\frac{\sum\limits_j \widehat{y_j} * N_{ji}}{\sum\limits_j N_{ji}} > 0.5, 1, 0) * V_i]}{\sum\limits_i V_i}$$

Where N is the weight by person, $\widehat{y_j}$ is the voting proportion estimator for either Biden or Trump, V is the electoral college votes, j represents the j^{th} cell of race combined with gender and i represents the U.S. states. The core idea is to calculate a weighted average of the proportion of Biden or Trump voters across the U.S. states.

The process comprises two steps. First, we estimate whether an election candidate has over half of the polling support in a specific state by calculating whether a state's weighted average number of Biden or Trump voters is greater than 0.5. In accordance with the "winner takes all" system, if either presidential candidate has over half of the polling support in a certain state, he is taking all the electoral college votes. Therefore, his weight for the state's electoral votes will be 1 (100%). In contrast, if his support is under 0.5, he will take 0 (0%) of the electoral votes. The cell hence has two levels. The first level is a combination of gender and race which is used to differentiate between individuals, and the second level is U.S. states to differentiate between states' electoral votes.

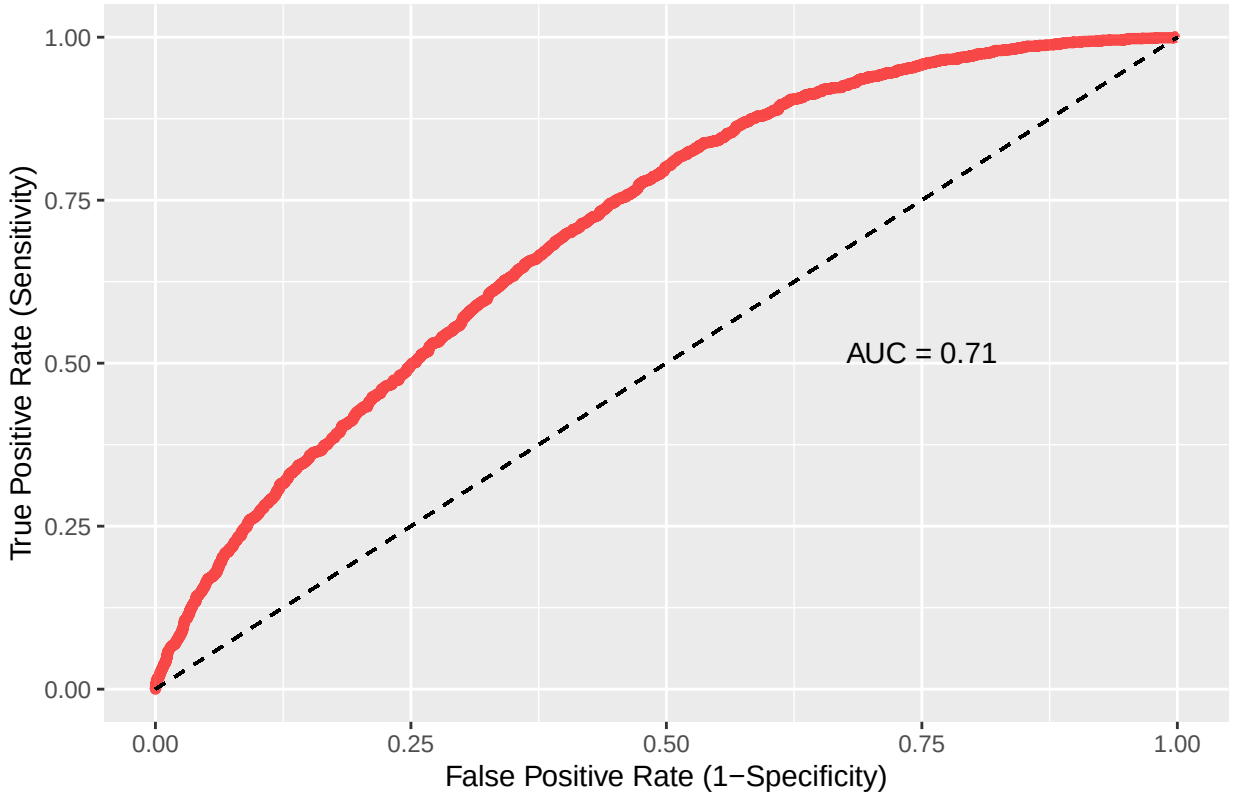
Aditional information

The post-stratification weighting used herein was conducted based on electoral college votes rather than specific state populations. This is because the polling for the U.S. presidential candidates is ultimately calculated in electoral votes rather than population polling. It would be more precise to use the electoral votes as the weights for state-level proportions of voters for each party.

Results

The proportion of voters in favour of voting for Biden and thus the democratic party was calculated to be 0.578 approximately, winning the republican party by 84 votes. This calculation is based on the post-stratification analysis of the number of voters in favour of the candidates modelled by a multilevel logistic regression model. Such a model accounts for an individual's combination of age and race, labor status, household income and state. The Area Under the ROC Curve ("AUC") is 0.71, which implies that the model is relatively effective in distinguishing between the proportion of voters in favour of the democratic party and the proportion of those in favour of the republican party.

Figure 1: ROC Curve



.

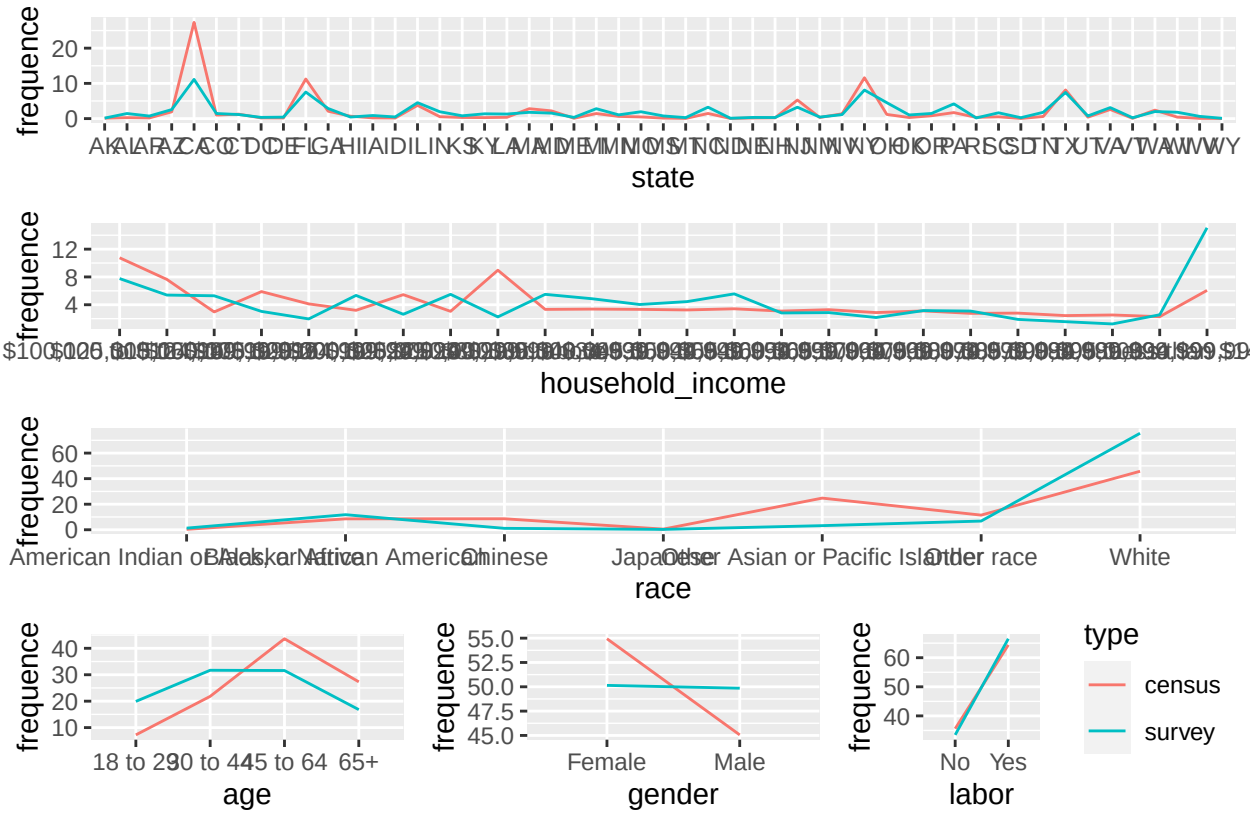
The Area Under the ROC Curve ("AUC") is 0.71, which implies that the model is relatively effective in

distinguishing between the proportion of voters in favour of the democratic party and the proportion of those in favour of the republican party.

Discussion

We have utilized the data from Democracy Fund + UCLA Nationscape survey data (2020) and the IPUMS USA census data (2018) to predict the upcoming 2020 general election result. Note that the Nationscape survey is voluntary and therefore is subject to volunteer bias. Furthermore, the 2020 election results are estimated based on a 2018 survey. Respondents' voting intentions may change over the two-year period and therefore the sample data is not very representative.

Figure 2: Survey data vs. Census data



In previous sections, a multilevel logistic regression model with post-stratification was performed to estimate the number of voters for each party. The survey data was partitioned into 14 cells, with each cell representing a combination of gender and race. The sample model was then used to calculate the proportion of each party's voters within each cell, and subsequently the cell-level results were proportionally aggregated to a national-level estimate by post-stratification methods.

Conclusion

Based on the estimation, the proportion of the voters in favour of the Democratic party is 0.578, or 311 out of 538 votes. Therefore, we predict that the Democratic Party will win this election. The model has an AUC of 0.71, which indicates a high distinguishableness between the proportion of voters voting for the Democratic Party and The Republican Party. Therefore, the result is relatively liable. Gender and whether the individual is African American are the most significant variables in determining an individual's probability of voting for each party.

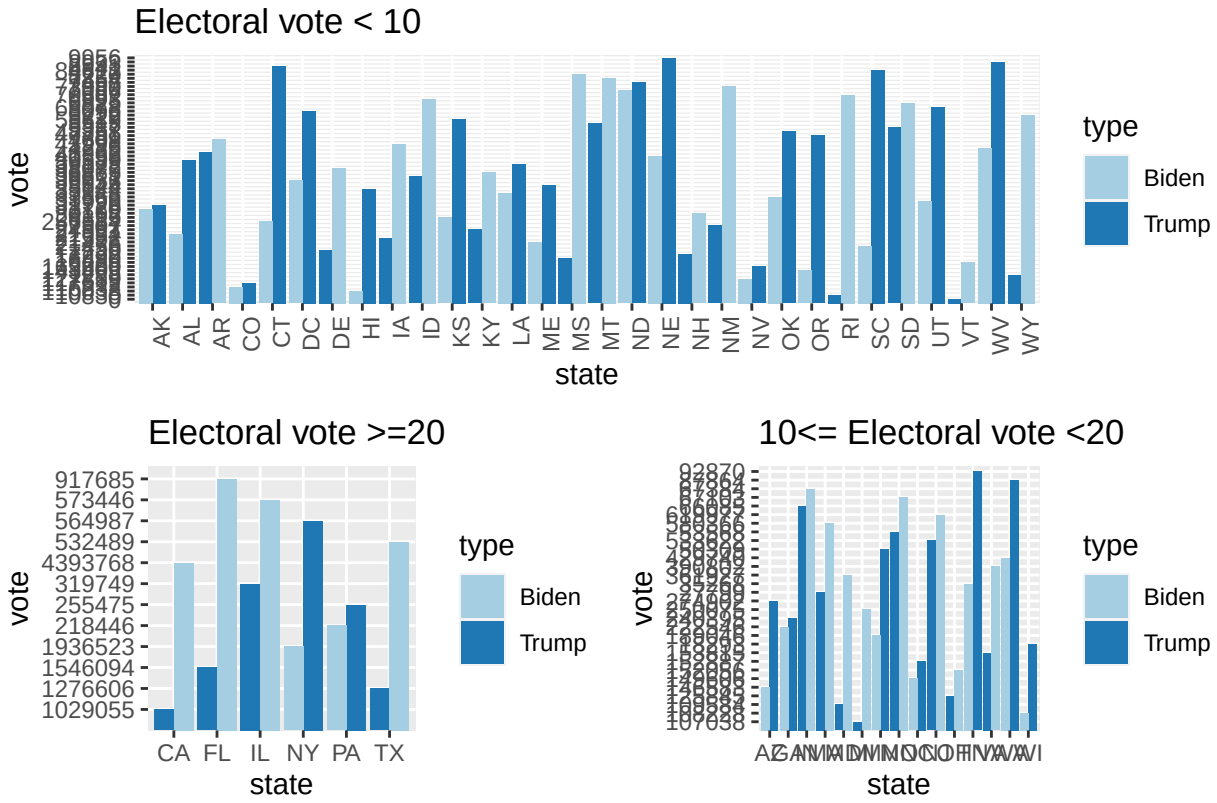
Table 1: Table 1:surpport for Trump and Biden in each cell

cell	Trump	Biden
Female American Indian or Alaska Native	0.5040519	0.4959481
Female Black, or African American	0.0000000	1.0000000
Female Chinese	0.0000000	1.0000000
Female Japanese	0.0049157	0.9950843
Female Other Asian or Pacific Islander	0.0622688	0.9377312
Female Other race	0.0167590	0.9832410
Female White	0.5283042	0.4716958
Male American Indian or Alaska Native	0.8408687	0.1591313
Male Black, or African American	0.0000000	1.0000000
Male Chinese	0.0209943	0.9790057
Male Japanese	0.0128478	0.9871522
Male Other Asian or Pacific Islander	0.4434238	0.5565762
Male Other race	0.2233517	0.7766483
Male White	0.8378089	0.1621911

Table 2: Table 2: Election Result

win	total_votes
Donald Trump	227
Joe Biden	311

Figure 3: Biden vs. Trump each state



Weaknesses

In the post-stratification process, we assumed the same election system across the states. In particular, the number of voters in favour of voting for either party was calculated based on the "winner-takes-all" assumption, that the winning candidate is receiving all electoral college votes of the respective state. However, Maine and Nebraska adopt a different system such that one electoral vote is awarded to each candidate, and the rest go to the winning candidate. Such differences could lead to an overestimated winning chance for the Democratic party.

Moreover, after a series of cleaning processes performed on the survey and census data, the size of the datasets was significantly reduced. More specifically, the filtered survey and census data contain 4,793 and 198,370 entries, respectively, in contrast with 6,479 and over 20 million entries before cleaning. The reduced sample size may have an adverse impact on the accuracy of the model. As well, survey respondents who haven't decided on their voting intentions or currently have no voting intentions (i.e. "Don't Know" and "N/A" for vote_2020) were excluded from our model. Their voting behaviors can cause deviations from actual results.

Next Steps

Post-hoc analyses and follow-up surveys will be conducted after the report. Comparisons between the actual and estimated election results will be made to evaluate the model efficiency and the significance of each predictor variable. Follow up surveys will cover questions for critical variables appearing in the census data but not included in the previous Nationscape Survey, such as religion and health insurance coverage. Gathering such data will grant us more predictor variables to work with, thereby improving the model accuracy.

References

1. Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from https://www.voterstudygroup.org/downloads?key=9e6f71ed-8c3b-4238-be7b-9d332bf90590
2. 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
3. Reilly, C., Gelman, A., & Katz, J. (2001). Poststratification Without Population Level Information on the Postratifying Variable, With Application to Political Polling. Journal of the American Statistical Association.
4. 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.
5. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
6. Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.
7. 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
8. Kay M (2020). *tidybayes: Tidy Data and Geoms for Bayesian Models*. doi: 10.5281/zenodo.1308151 (URL: https://doi.org/10.5281/zenodo.1308151), R package version 2.1.1, <URL: http://mjskay.github.io/tidybayes/>.
9. Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86. https://github.com/topepo/caret/
10. 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/

11. Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. http://haozhu233.github.io/kableExtra/, https://github.com/haozhu233/kableExtra.

12. Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://rpkgs.datanovia.com/ggpubr/