58% vs 42%, Trump Will Win the 2020 U.S. Presidential Election

With 72% accuracy of multilevel logistics model, Trump will win most of the swing states in the election

Dingyi Yu, Linzi Guan, Liuxuan Wang

03 November 2020

Abstract

The influential US 2020 Election keeps catching people's eyes all over the world. Obtained data from Democracy Fund + UCLA Nation scape "full data-set"(Tausanovitch and Vavreck 2020) and post-stratification data from 2018 1-year American Community Surveys (Team, n.d.), this paper conducts a multilevel logistic regression model with post-stratification to predict the voting result. We expect Donald Trump from the Republican Party to win 58 per cent of the national popular vote and defeat Joe Biden from the Democratic Party in the US 2020 Election. This study appeals for governments and businesses to better prepare future strategies and make decisions, and even countries to consider the political and economical relationship with the US in an international context.

Keywords: forecasting; US 2020 election; Trump; Biden; Multilevel Regression With Post-stratification

Introduction

The upcoming US 2020 Election is watched by the world and the result of the coming election not only has important meaning for the US citizens, but will also has large effects on various international issues, such as diplomatic relations, world economy and international trades. With so much attention, people are curious about the polling situation and a relatively accurate estimation of the US election result is also helpful for governments and related businesses to better prepared for future strategic and make decisions. These needs stimulate the organizations to predict the election result based on various kinds of informations and we are also interested about the US 2020 Eelection results. Hence, in this paper, we used statistical models to make our own prediction in R(R Core Team 2020).

There are two major political parties in the US 2020 Election, the Democratic Party and the Republican Party and the election uses "winner-takes-all" plurality system that if the support rate for one candidate is higher in one state, he takes all the votes in the state. Joe Biden and Donald Trump are candidates for presidents from the two parties separately(Joe Biden from Democratic party and Donald Trump from Republican Party) and our study tends to forecast who would win the US election in 2020. We used survey data from Democracy Fund + UCLA Nation scape "full data-set"(Tausanovitch and Vavreck 2020) and post-stratification data from 2018 1-year American Community Surveys (Team, n.d.).

Based on the datasets, we have applied a statistical method to build two multilevel logistic regression models with post-stratification in this analysis to predict the polling result. We have learned about that most states have a certain preference for one candidate, but some swing states would indeed influence the overall results (2020), so we conducted a multilevel logistic regression model firstly taking states as cells only to see the fixed effects of states on the overall results(whether people would vote for a certain candidate because they are from the same state). However, some states have too few observations in the sample data to analysis, which makes it not convincing to make a valid prediction model from it, we abandoned the model and considered a new one. Noticing that there appears to be significant fixed effects in race and gender, we built a new multilevel regression model taking genders and races as cells. This time we had sufficient data to analyse and we conducted a prediction model using genders and races as cells then. We predicted that Donald Trump would win the election at 58% versus Joe Biden at 42% and we achieved a comparable high accuracy score of 77.6% based on the AUC test.

The paper would be followed by a full disclosure and analysis of the data we built our study on in the Data Section, some detailed discussions on the statistical models that we used for forecasting in Model Section, some discussions and results, as well as some limitations and next steps.

Data

Survey data

The individual-level survey data were retrived from Democracy Fund + UCLA Nationscape “Full dataset”. Conducted by researchers at UCLA, Nationscape is a 16-month election study that completes approximate 6,250 interviews each week starting from July, 2019 to January, 2021. The dataset we used is the summer 2020 release. (Tausanovitch and Vavreck 2020)

Data Variables

The raw survey data consist of following nine variables: vote_2020, vote_intention, age, gender, state, employment, education, household_income and race_ethnicity.We used those nine variables in the survey data after a data cleaning of the raw data in R(R Core Team (2020)) and a data procession with tidyverse package(Wickham et al. (2019)) in R. We did not construct any variables combining various ones on the first stage

vote_2020 is a categorical variable indicates the candidate that the respondent would like to vote for. It has five categories including Donald Trump, Joe Biden, Someone else, I would not vote and I am not sure/don’t vote. Note that we use the data to filter out responses only want to vote for Trump or Biden to make it binary for doing logistics analysis.

vote_intention is a categorical variable indicates whether the respondent want to vote or not. It has four categories including Yes I will vote, No I will not vote but I am eligible, No I am not eligible to vote and Not sure. Note that we only filter people who will vote and not sure for our further analysis, we assume they will go to vote at the election day.

age is a numeric variable that stands for the self reported age of the respondent.

gender is a categorical variable that represents the self-reported gender of the respondent. It has two values: Male, Female

state is a categorical variable that represents the two-character postal abbreviation that the respondent entered. It is expected to have 51 values of the postal abbreviation of each states in U.S.

employment is a categorical variable indicates the current employment status of the respondents. It includes 9 values including Full-time employed, Homemaker, Retired, Unemployed or temporarily on layoff,Part-time employed,Permanently disabled,Student,Self-employed and Other.

education is a categorical variable that represents the highest level of education the respondent have completed. It has 11 values including 3rd Grade or less, Middle School - Grades 4 - 8, Completed some high school, High school graduate, Other post high school vocational training， Completed some college but no degree, Associate Degree,College Degree (such as B.A., B.S.),Completed some graduate but no degree, Masters degree, Doctorate degree.

household_income is a categorical variable that indicates the range of respondents’ annual household income before taxes. It has 24 variables including Less than $14,999;$15,000 to $19,999;$20,000 to $24,999;$25,000 to $29,999;$30,000 to $34,999;$35,000 to $39,999;$40,000 to $44,999;$45,000 to $49,999;$50,000 to $54,999;$55,000 to $59,999;$60,000 to $64,999;$65,000 to $69,999;$70,000 to $74,999;$75,000 to $79,999;$80,000 to $84,999;$85,000 to $89,999;$90,000 to $94,999;$95,000 to $99,999;$100,000 to $124,999;$125,000 to $149,999;$150,000 to $174,999;$175,000 to $199,999;$200,000 to $249,999;$250,000 and above

race_ethnicity is a categorical variable that indicates the race of the respondents.It contains 15 variables: White, Black or African American, American Indian or Alaska Native, Asian (Asian Indian), Asian (Chinese), Asian (Filipino), Asian (Japanese), Asian (Korean), Asian (Vietnamese), Asian (Other), Pacific Islander (Native Hawaiian), Pacific Islander (Guamanian), Pacific Islander (Samoan), Pacific Islander (Other)

Survey methodology

Nationscape conducted the survey through online interviews on Lucid, which is a market research platform running an online exchange for survey respondents with a weekly questionnaire length of 15-minute median time and the interviews are in English. Lucid gathers demographic quotas including age, gender, ethnicity, region, income and education. Before taking the online survey, respondents are asked to complete an attention check.

Population, frame and sample

The target population of Nationscape is all U.S. citizens who are 18 years old or more with voting rights and the sample frame is the respondents of Nationscape survey on Lucid. Nationscape samples were drawn simple randomly from the sample frame without replacement and Respondents on Lucid would be sent directly to survey software operated by Nationscape. The target sample size is 6,479 respondents and the final sample size is 5,395 respondents. To make the survey data representative of the American population, the survey data are

weighted according to the adult population of the 2017 American Community Survey of the U.S. Census Bureau(Team (n.d.)) by a simple raking technique and one set of weights is generated for each week's survey. Factors that influence weights include demographic information and some interactions such as Hispanic ethnicity by language spoken at home, education by gender, gender by race, and etc.

Non-response

Overall, the average non-response rate of the survey is 25%, comprising 12% of those selected to be interviewed immediately rejecting to respond, around 5% interviewed not completing the survey and 8% for speeding or straight-lining through the survey, where speeding is completing the survey less than 6 minutes and straight-line is selecting the same response for every question in the three policy question batteries. To improve the overall accuracy of the survey data, Nationscape removes the 8% for speeding or straight-lining.

Data features and strengths

• Data accuracy: Nationscape removes the 8% for speeding or straight-lining to improve the overall accuracy of the survey data, in case the results of respondents who provided answers too fast or selected same answers for multiple questions would influence the survey results. Therefore, the data accuracy is strongly improved.

• Representativeness: The survey data were modified by weights of different factors according to the ACS, so the weighted survey data is more representative of the dataset than the plain data.

Data weaknesses

• Survey length: The survey takes respondents 15 minutes without counting for screening time. For some people it can take up for 17 minutes. A relatively long survey time with online survey method generates non-respondents with qutting or stop halfway and invalid data with speeding through the whole survey.

• Imperfect coverage: The survey is conducted online on Lucid so the respondents would only be those people who have access to a networked computer or mobile device and at the same time, they are users for Lucid. However, our target population is all American citizens with voting rights, so a proportion of our target population is out of our coverage range in the sampling frame, and this makes our sample data less representative of our target population.

• Non-response: Nationscape sample's non-response rate was generally 75%, and since we know few about the non-respond cases, there would be some biases in the data.

• Sampling error: Sampling error is unavoidable. The results will vary from sample to sample. It willsurely be different from the results of the census.

Key facts about the survey data:

✓ Overall, more respondents intended to vote for Joe Biden.

✓ 42 voting states out of the total 50 have clear proclivity for one certain candidate, while respondents in 8 of them show swinging preferences for both candidates.

✓ Male voters like Donald Trump more while female voters prefer Joe Biden.

✓ Preference for Donald Trump is not consistent among all age groups in male voters, while all age groups of female support Joe Biden more than Donald Trump.

✓ More white respondents in the US support Donald Trump.

✓ Employment Status does not make distinguished difference.

✓ There are no specific pattern of voting intentions for different income groups.

✓ Respondents with comparable high degrees are more likely to vote for Donald Trump, while respondents with lower degrees tend to support Joe Biden.

✓ There is no significant difference for voting intentions of different employment status.

In our study, two multilevel logistic regressions were conducted to predict the voting result. For the multilevel logistic regression, two datasets are required(Alexander 2020): one is the individual-level survey data retrieved from Democracy Fund + UCLA Nationscape "Full dataset" (Tausanovitch and Vavreck 2020), which is the sample of our interest, and another is the Post stratification data got from the American Community Surveys(ACS) (Team, n.d.) that we used to adjust for some sample bias. In the following part, a detailed introduction, explanation and interpretation of the data would be discussed.

42 voting states out of the total 51(50 states and Washington D.C.) have clear proclivity for one certain candidate, while respondents in 8 of them show similar preferences for both candidates(Figure 1).The swing in those 8 states creates uncertainty that might have influenced the overall prediction. In the election, most states would steadily vote for one candidate, so what really matters in the result of the election is the voting results of those swing states(2020). We could see from Figure 2 that in the eight swing states, Biden wins Michigan, North Carolina, Ohio and Wisconsin, while Trump only wins Arizona and Pennsylvania. As of Florida and iowa, there is no significant difference shown.

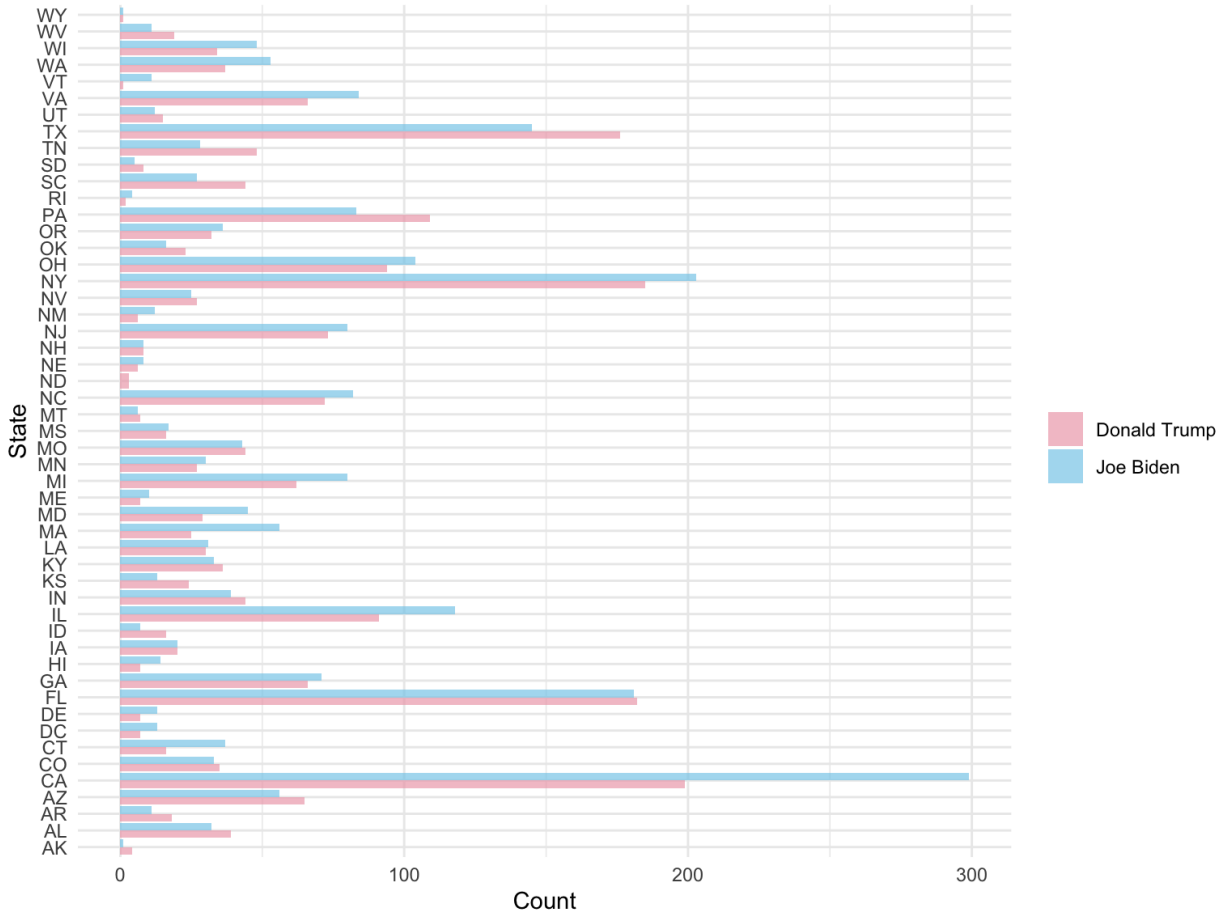


Figure 1: National States Poll.The color represent different voting intentions: blue for Joe Biden and pink for Donald Trump. A longer bar represents more people of this race group will vote for certain candidate.

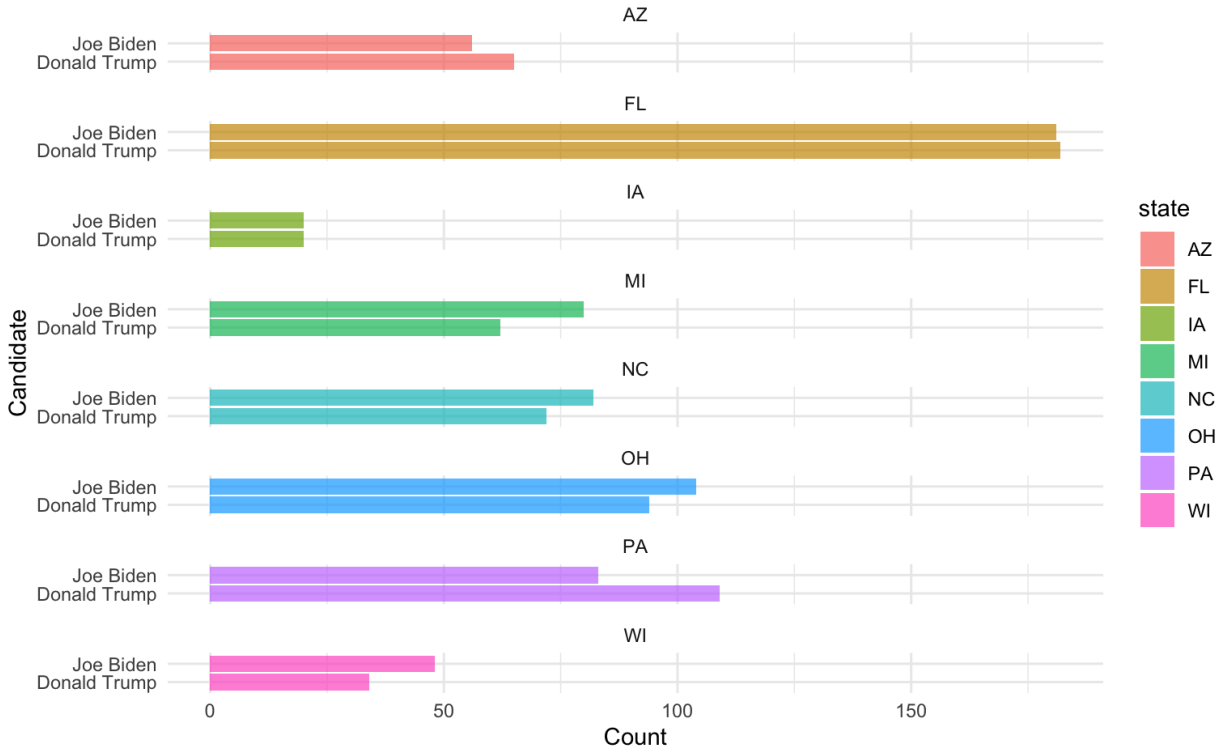


Figure 2: Election Poll in Swing States. Of the 51 states(50 states and Washington D.C.) in the U.S., 8 did not show clear proclivity of their preference for voting with similar number of respondents supporting both candidates. These 8 states include Wisconsin, Pennsylvania, Ohio, North Carolina, Michigan, Iowa, Florida, and Arizona. They still swayed around which candidate to vote for.

What surprising is that male voters like Donald Trump more while female voters prefer Joe Biden! Figure 3 shows a large gap in preferences for Donald Trump and Joe Biden between males and females. Of the 2366 male respondents, a significant larger proportion of 9.64% intended to vote for Donald Trump. On contrast, of the 2241 female respondents, an even larger difference 18.34% existed in preference for Joe Biden, which is nearly twice the voting difference in male. This shows that there does exist a fixed effect of gender in voting intention.

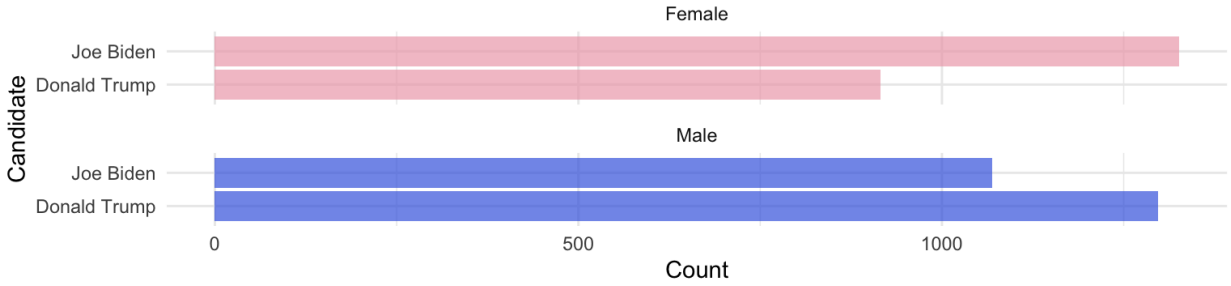


Figure 3: Male and Female voters voting intention count. The colours represent different voting intentions: blue for Joe Biden and red for Donald Trump. A longer bar represents a higher proportion of voting choice.

It is interesting to see that the preference for Donald Trump is not consistent among all age groups in male voters, where younger age groups tended to like Joe Biden more, while all age groups of female support Joe Biden more than Donald Trump. Figure 4 presented the voting intention for male and females separately in different age groups. Male respondents that are younger than 35 years old were more supportative to Joe Biden while older male respondents liked Donald Trump more. For female respondents, all age groups have shown a strong preference for Joe Biden.

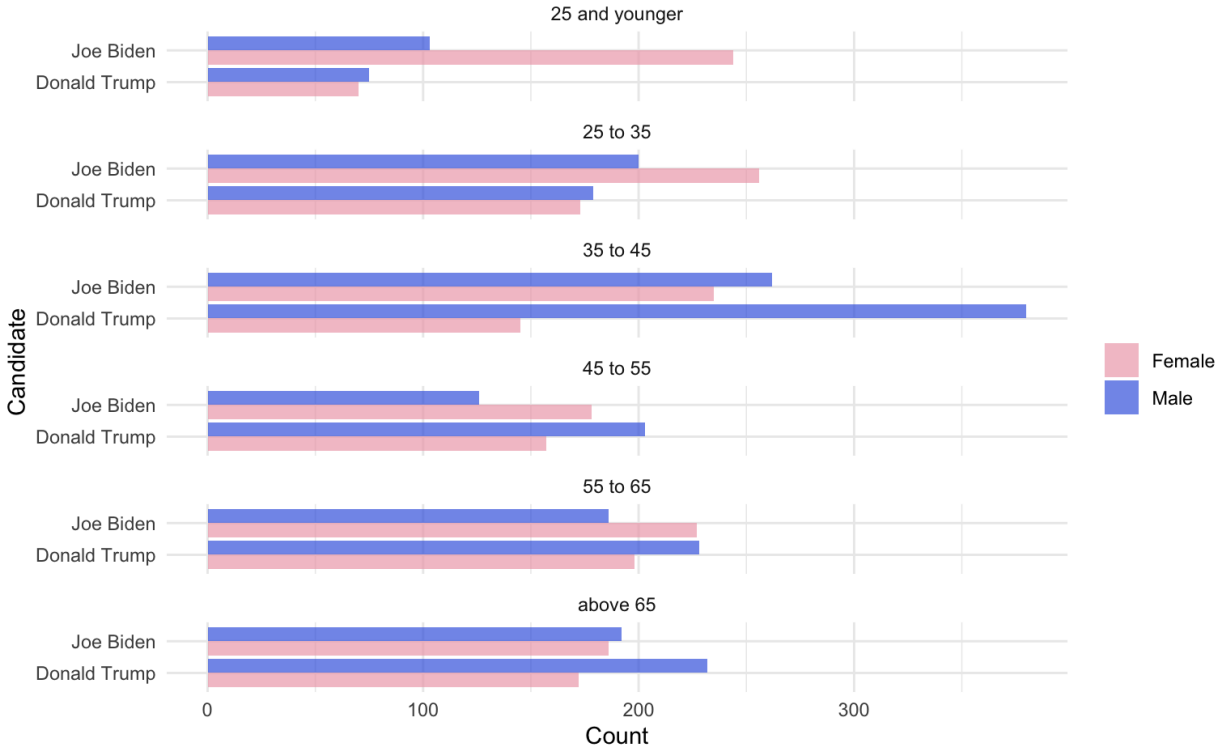


Figure 4: Voting intention in different age group for male voters(blue) and female voters(pink). The bar plot shows voting choices of different age groups separately for male and female. A longer bar represents more people of this race group will vote for certain candidate.

For different race groups, the survey data reflects that more white people in the US support Donald Trump. However, for other group of races, except for American Indian or Alaska Native who have no apparent distinction of support between two candidates, minority races such as Asian, pacific islander, Japanese and Chinese, support Joe Biden more than Donald Trump. Another issue reflected from the survey is that white people is the absolute majority of US citizens, which makes up over 70% of the survey observations. This indicates that white people's intentions are much more effective.(Figure 5)

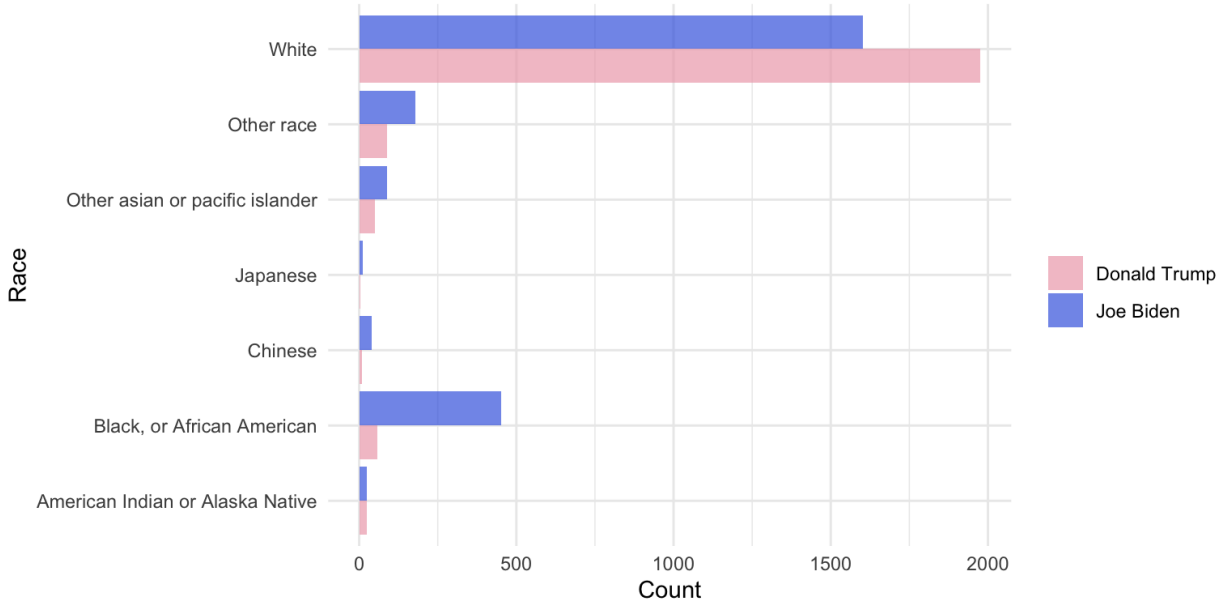


Figure 5: The vote intention of different groups of race. The bar plot shows voting intentions for different race groups. The colours represent different voting intentions: blue for Joe Biden and pink for Donald Trump. A longer bar represents more people of this race group will vote for certain candidate.

Generally, for people with different income levels, there are no specific pattern of voting intentions for high-income groups and low-income groups. Generally speaking, low-income and middle income groups with under $75,000 anual household income may support Joe Biden more and high-income groups with over $100,000 annual household income may support Donald Trump more. For middle-income groups with household annual income between $75,000 and $100,000, the voting intentions vary among groups. More households with income level of "$75,000 to $79,999" and "$90,000 to $94,999" intent to vote for Donald Trump, while the other three income level groups between $75,000 and $100,000 will vote more for Joe Biden, according to the survey data.(Figure 6)

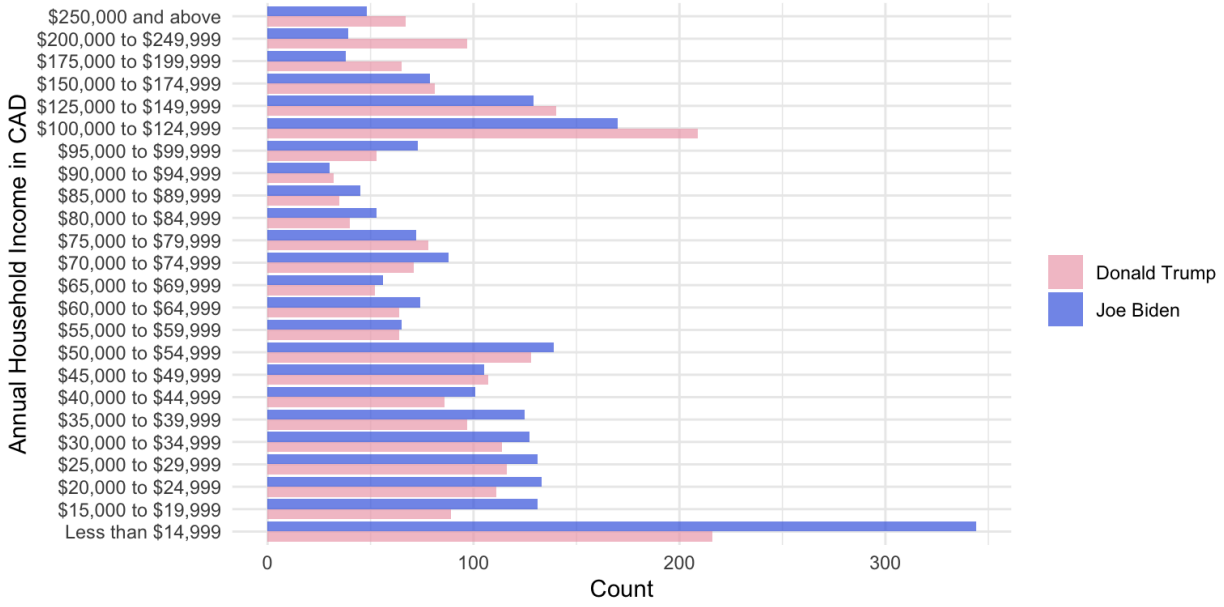


Figure 6: Support of two candidates under different level of income. The colours represent different voting intentions: blue for Joe Biden and pink for Donald Trump. A longer bar represents more people of this race group will vote for certain candidate.

It is noticeable that respondents with comparable high degrees who have doctorate and masters degrees are more likely to vote for Donald Trump, while respondents with lower degrees including college degree, Associate degree, completed some college without degree, and completed some high school without degree are more likely to support Joe Biden, and exceptions are high school graduates who are more supportative for Donald Trump and no obvious difference for middle school or under. (Figure 7)

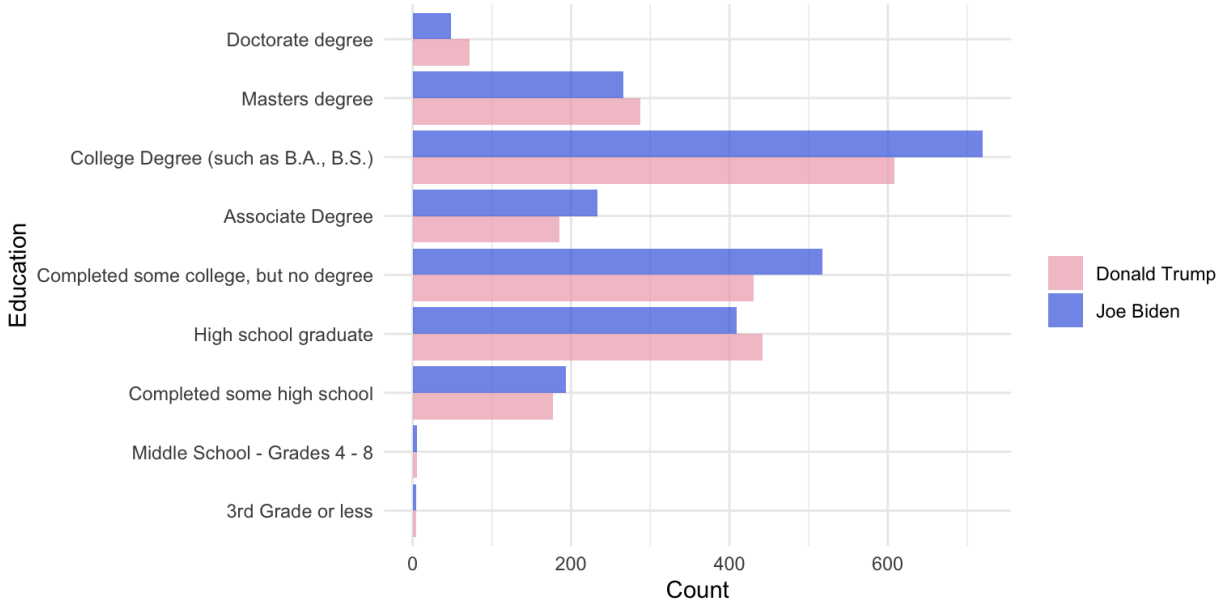


Figure 7: Support of two candidates under different education background. The colours represent different voting intentions: blue for Joe Biden and pink for Donald Trump. A longer bar represents more people of this race group will vote for certain candidate.

There does not exist significant difference for voting intentions of different employment status. For all respondents from all employment status, more respondents support Joe Biden compared to Donald Trump. (Figure 8)

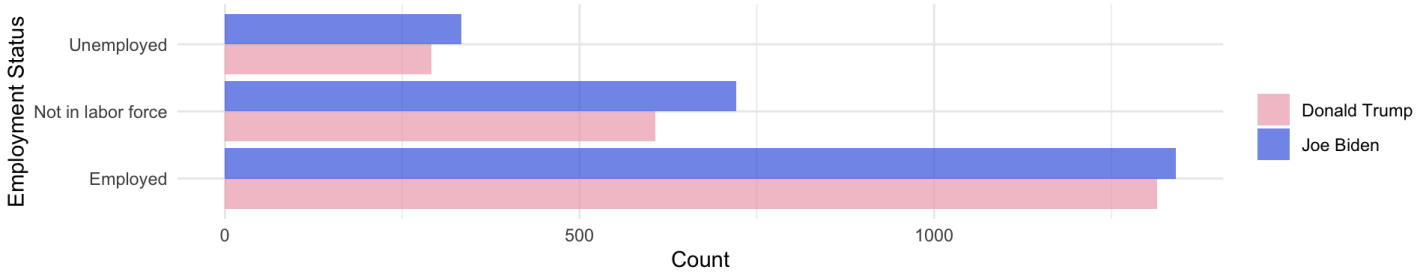


Figure 8: Support of two candidates under different education background. The colours represent different voting intentions: blue for Joe Biden and pink for Donald Trump. A longer bar represents more people of this race group will vote for certain candidate.

Census Data

The post-stratification data source is from American Community Surveys. The American Community Survey (ACS) is conducted by the U.S. Census Bureau to provide continuously updated information on the numbers and characteristics of the nation's people and housing (2007).The sample survey methodology is online survey.

Data Variables

The raw census data consist of following variables: perwl, citizen, age, sex, stateicp, empstat, educd, hhincome and race. The detailed explanation of the variables are shown in the following part. To compute the post stratification, we constructed states-groups out of states, and constructed gender*race groups out of gender and races. The detailed process of post-stratification is also discussed below.

"perwt" is a numeric variable called personal weight. It should be a positive whole number ranging from 1 to 2313. More specifically, for example, the person with 10 perwt vote for Trump is equivalent to ten votes gained by Trump.

"citizen" is a categorical variable reflecting citizen status, The categories include N/A, Born abroad of American parents, Naturalized citizen, Not a citizen, Not a citizen but has received first papers, and Foreign born citizenship status not reported. In details, Citizen status reports the citizenship status of respondents, distinguishing between naturalized citizens and non-citizens. For 1900-1940, respondents who were not yet citizens but who had begun the naturalization process ("received first papers") are identified. Note that we only select naturalized citizen and born abroad of American parents to filter out people who are not eligible to vote

"age" is a numeric variable with positive whole numbers indicating the respondents' age. The youngest is 1 and the oldest is 97. Pay attention that age reports the person's age in years as of the last birthday.

"sex" is a categorical variable reports the gender of the respondents, that is whether the person was male or female. The categories are Male and Female.

“stateicp” is a categorical variable identifies the state in which the housing unit was located, using the coding scheme developed by the Inter-University Consortium for Political and Social Research (ICPSR). The ICPSR scheme orders states first by geographic division and then alphabetically within each division. The categories contain Arizona, West Virginia, Tennessee, Oklahoma, Maryland, Kentucky, Texas, South Carolina, North Carolina, Mississippi, Louisiana, Florida, Georgia, Alabama, Arkansas, Virginia, South Dakota, Nebraska, North Dakota, Minnesota, Missouri, Iowa, Kansas, Michigan, Ohio, Wisconsin, Illinois, Indiana, Pennsylvania, New York, New Jersey, Vermont, Delaware, Rhode Island, New Hampshire, Massachusetts, Maine, Connecticut, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming, California, Oregon, Washington, Alaska, Hawaii, Puerto Rico, State groupings (1980 Urban/rural sample), Military/Mil.Reservations, District of Columbia, and State not identified.

“empstat” is a categorical variable indicates whether the respondent was a part of the labor force – working or seeking work – and, if so, whether the person was currently unemployed. The categories contain N/A, Employed, Unemployed, and Not in labor force.

“educd” is a categorical variable indicates respondents’ educational attainment, as measured by the highest year of school or degree completed. Specifically, the categories include N/A, No schooling completed, Nursery school to grade 4, Nursery school preschool, Kindergarten, Grade 1, 2, 3, or 4, Grade 1, Grade 2, Grade 3, Grade 4, Grade 5, 6, 7, or 8, Grade 5 or 6, Grade 5, Grade 6, Grade 7 or 8, Grade 7, Grade 8, Grade 9, Grade 10, Grade 11, Grade 12, 12th grade, no diploma, High school graduate or GED, Regular high school diploma, GED or alternative credential, Some college, but less than 1 year, 1 year of college, 1 or more years of college credit, no degree,2 years of college, Associate’s degree, type not specified, Associate’s degree, occupational program, Associate’s degree, academic program, 3 years of college, 4 years of college, Bachelor’s degree, 5+ years of college, 6 years of college (6+ in 1960-1970), 7 years of college, 8+ years of college, Master’s degree, Professional degree beyond a bachelor’s degree, Doctoral degree, and Missing.

“hhincome” is a numerical variable ranging from -16400 to 9999999(means no income reported). It reports the total money income of all household members age 15+ during the previous year. The amount should equal the sum of all household members’ individual incomes, as recorded in the person-record variable INCTOT. The persons included were those present in the household at the time of the census or survey. People who lived in the household during the previous year but who were no longer present at census time are not included, and members who did not live in the household during the previous year but who had joined the household by the time of the census or survey, are included.

“race” is a categorical variable reflecting self-reported sociopolitical constructed race. The categories include Other race nec, Two major races, Other Asian or Pacific Islander, Japanese, Chinese, American Indian or Alaska Native, Black/African American/Negro, White, and Three or more major races.

Data features and strengths

• Timeliness: Instead of 2 years, ACS data are released 8 to 10 months compared to 2 years. • Frequency: ACS data are updated every year compared to every 10 years • Higher quality of data(completeness of response): more complete responses are achieved by ACS with computer-assisted telephone and personal interviewing of households that do not responded by mail. The ACS interviewers are more experienced and highly trained.

Data weaknesses

• ACS estimates have significantly larger margins of error, which is because of the much smaller sample size of the ACS and the greater variation in the sample weights due to the subsampling for field interviewing of households not responding by mail or telephone • Unmeasured estimation error makes postcensal population and housing estimates used as survey controls less effective.

Data Visulization

Figure 9 shows the distribution of gender for each state of the census data. Overall, there are more female than male in each state. California has the most population among all states in the US. Interestingly, we can observe that for states with small population, the male to female ratio is close to 50%. While for states with large population, the male to female ratio becomes larger.

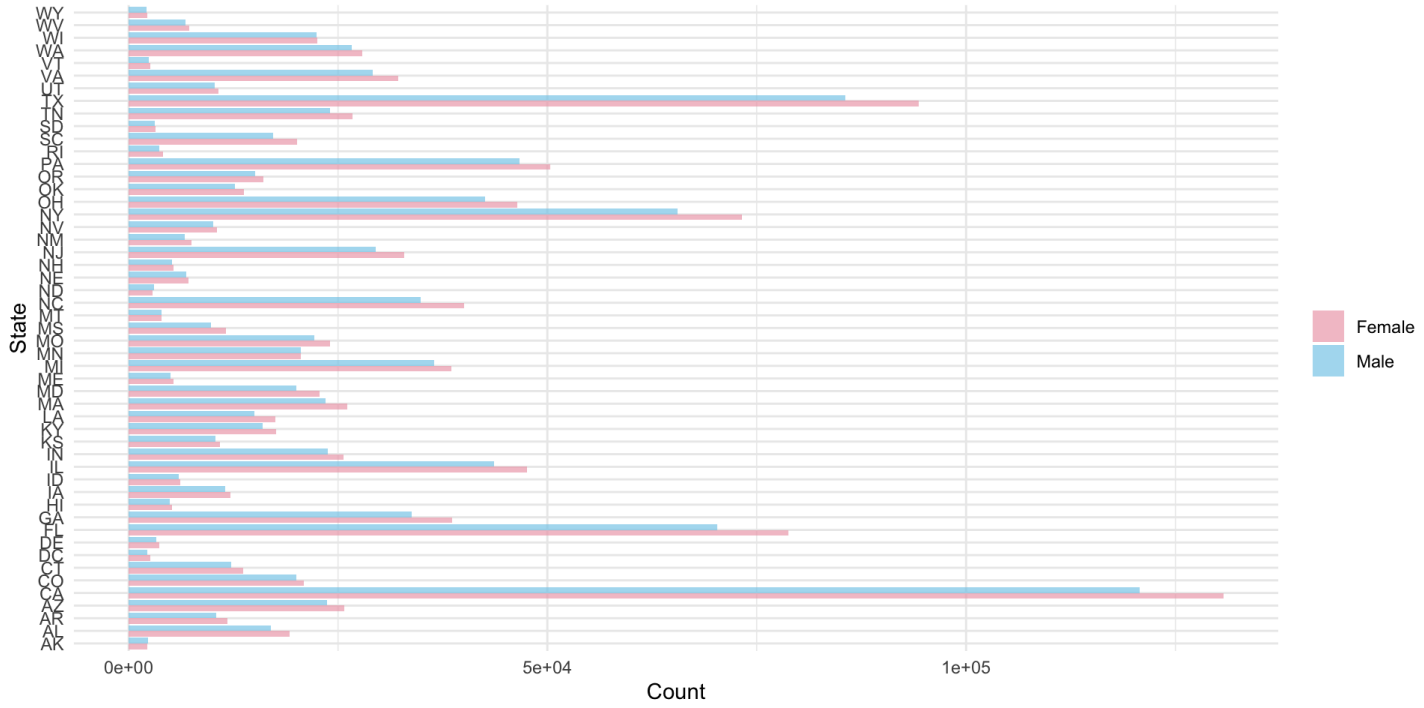


figure 9: gender counts for each state of census data. Blue bar represents male and pink bar represents female. The longer the bar is, the larger number of people it represents.

Post-Stratification

The size of survey data, as what was mentioned in previous part, only contains less than five thousands observations. Therefore, it would be overly biased to simply use their respond to represent the vote result of all U.S. citizens that are eligible to vote, which can be a population with a size as large as over one hundred million. In order to reduce bias and generalize the result as possible, we imply the technique of Post-Stratification.

Post-Stratification is a statistical technique to correct estimates when there are known differences between target population and study population. According to (Wang et al. 2015), we could have a big picture that how post-stratification could eliminate an obvious bias in study population (X box players in the example).

The post-stratification technique could be completed by diving the population into smaller group (which is called cells here) based on selected cell variable. Note that the accuracy of the model is highly dependent on the size of the cell, as the more detailed the cell is, the more accurate the model is expected to be. Then we use sample to estimate the response variable inside each cell, and aggregate the value of each cell by assigning a weight based on each cell's proportion in population. Mathematically, it could be expressed as:

$\hat{Y} = \sum_i n_i \hat{y}_i / \sum_i n_i$ where $\hat{y}_i$ is the estimated value based on the model in cell i and n_i is the number of population in cell i

Following this way, we could reduce the bias of the model by assigning less weight to groups that occur rarely, and add weight for them who can represent the majority well.

In order to do post-stratification using census data, we need to convert certain variables to make them match the name and format. For example, we have converted the numeric annual household income in census data to categorical annual household income range for better estimation.

Note that since we set variable $State$ as the cell of the first model, there should be 51 values of the cell in total, which match with the 51 states. However, the model could be potentially problematic due to the lack of "diversity" of the cell element, which means the number of variables that consist of the cell is not enough. We have tried to combine more variables such as $AgeGroup$ or $Race$ to the $State$ to complicate our cell in order to get a more accurate model. However the problem is that for certain states, such as Wyoming(only two observations), the data is very limited and we cannot train our model with the combined cell due to lack of training data. Similarly that is also the reason of creating age group variable instead of using age. Even though we want the cell to be as accurate as possible, we cannot make sure we have the combination of all cell variables for each single age.

Model

Multi-level Logistics Regression Model

Given the fact that the target population in the United States election includes all U.S. citizens, the logistic regression model contains information on state, age, gender, race, education, household income and employment status of U.S. citizens respondents and hopefully draws a general portrait of voters for supporters of each candidate. The variable selection is based on the common demographic information, as we could apply them to almost every U.S. citizen. The mode selection is based on the property of data and what we would like to predict. Logistic regression is an approach for predicting binary results. Since the variables all have the nature in the factor form rather than numeric form, for instance male and female for gender, and we would like to predict a binary outcome of whether vote Trump or not (note that one does not vote Trump must vote Biden, as we have filtered all other options in cleaning process), we believe that logistic regression model is a fitted one to make the prediction. Our first attempt was to build a multilevel logistic regression model with R (R Core Team 2020) that set states as our "cell" which means dividing survey data into groups based on states. We then set household income as the coefficient changing variables since the lowest income standard and overall income level in different states is known as different. The model could be expressed as:

$$Pr(Y_{ij}) \in \{Trump, Biden\} =$$

$$logit^{-1}(\alpha_0 + a_j + \beta_{1[i]}(gender_i) + \beta_{2[i]}(education_i) + \beta_{3[i]}(AgeGroup_i) + \beta_{4[i]}(EmploymentStatus) + \beta_{5[i]}(Race) + \beta_{j[i]}^{Householdincome} + \epsilon_{ij})$$

$$\epsilon_{ij} \sim N(0, \sigma_e^2)$$
$$a_j \sim N(0, \sigma_a^2)$$
$$\beta_{j[i]}^{householdincome} \sim N(0, \sigma_{householdincome}^2)$$

where α_0 is the fixed baseline intercept and a_j corresponds to the varying intercept associated different cell variable, which is the variable State in the case here. a_j is the varying intercept that change as the variable $State$ changes(Wang et al. 2015). For example, fixing other variables to be the same, males in California might vote differently from males in Michigan , by having different intercept $\alpha_0 + a_{CA}$ and $\alpha_0 + a_{MI}$.

Further, The notation β_{1i} to β_{5i} are the fixed slope for the variable inside the bracket respectively. Note that the subscript i here corresponds to the *i*th variable. For example $\beta_{1[male]}(Gender_{male})$ and $\beta_{1[female]}(Gender_{female})$represents the voting difference between male and female by fixing other variable.

Then we see the notation $\beta_{j[i]}^{householdincome}$, which corresponds to the varying slope of household income level affected by the change of cell variable $State_j$. For example, $\beta_{CA[\$50{,}000 to \$54{,}999]}^{householdincome}$ will not be the same as $\beta_{MI[\$50{,}000 to \$54{,}999]}^{householdincome}$, which means for voters with household income in the range of $50,000 to $54,999, fixing other variable to be the same,voters in California behave differently from voters in Michigan.

The goal we want to achieve, which is the value of $Pr(Y_{ij})$, represents the probability of voting for Biden (set Trump as reference) in $State_j$ with $Householdincome_i$ and other listed variable with corresponding fixed slope.

By observing the summary output of the model, we could see that the p value for $Agegroup$, $Race$ and $Gender$ are relatively small, which shows that these three variable tend to have significant effect to the candidate preference. That leads us to build a alternative model by combining $Race$ and $Gender$ together to create a $Cell$ variable, and add $Agegroup$ as the layer. Notice the reason why not combining $Agegroup$ into the cell variable is for the convenience of latter post-stratification, since We don't have enough training data to cover all the cases. The alternated model could thus be expressed as:

$$Pr(Y_{ij}) \in \{Trump, Biden\} =$$

$$logit^{-1}(\alpha_0 + a_j + \beta_{1[i]}(Householdincome_i) + \beta_{2[i]}(Education_i) + \beta_{3[i]}(State_i) + \beta_{4[i]}(EmploymentStatus_i) + \beta_{j[i]}^{Agegroup} + \epsilon_{ij})$$

$$\epsilon_{ij} \sim N(0, \sigma_e^2)$$
$$a_j \sim N(0, \sigma_a^2)$$
$$\beta_{j[i]}^{Agegroup} \sim N(0, \sigma_{Agegroup}^2)$$

The notation interpretation of the cell model is similar, but one needs notice that since $Cell$ is consist of $Gender$ and $Race$, there are 2 * 7 = 14 different cells values (2 $Gender$ categories and 7 $Race$ categories).Some examples of the $Cell$ value could be: White Male, Chinese Female… Also, the $Agegroup$ variable becomes the layer, which means people in different age group might tend to vote different candidate.

We then apply to Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to compare the two models (Brownlee 2020).

$AIC = 2k - 2ln(L)$ where k is the number of parameters and L stands for the likelihood function.

Generally, once the complexity of the model increases, which usually increases the value of k, the likelihood function will increase as well, then lower the AIC value is supposed to be. However, if the model is over-fitted, the AIC value will increase consequently.The purpose of AIC function is to reduce the complexity of the model in order to prevent over-fitting, so we tend to pick the model with smaller AIC value. By observing the model summary, the first model have an AIC value of 6282.90 while the second model have a comparatively smaller AIC value of 5856.40.

$BIC = kln(n) - 2ln(L)$ where k is the number of parameters, L stands for the likelihood function and n is the sample size.

Similarly, BIC is also used to prevent the over-complexity of the model so we tend to pick a model with smaller BIC value.By observing the model summary, the first model have an AIC value of 8361.50 while the second model have a comparatively smaller AIC value of 6532.10.

From AIC and BIC values, the second model is obviously a better choice to avoid the complexity. We then want to use Receiver Operating Characteristic (ROC) Curve to see the accuracy of the model by using package (Robin et al. 2011). Note that the area under the curve, represented by the gray square, have a sum of 1. The blue area stands for the accuracy that our predicted voting preference matches the real response. In the curve, the accuracy is measured to be 72% (Figure 10) (Narkhede 2019).

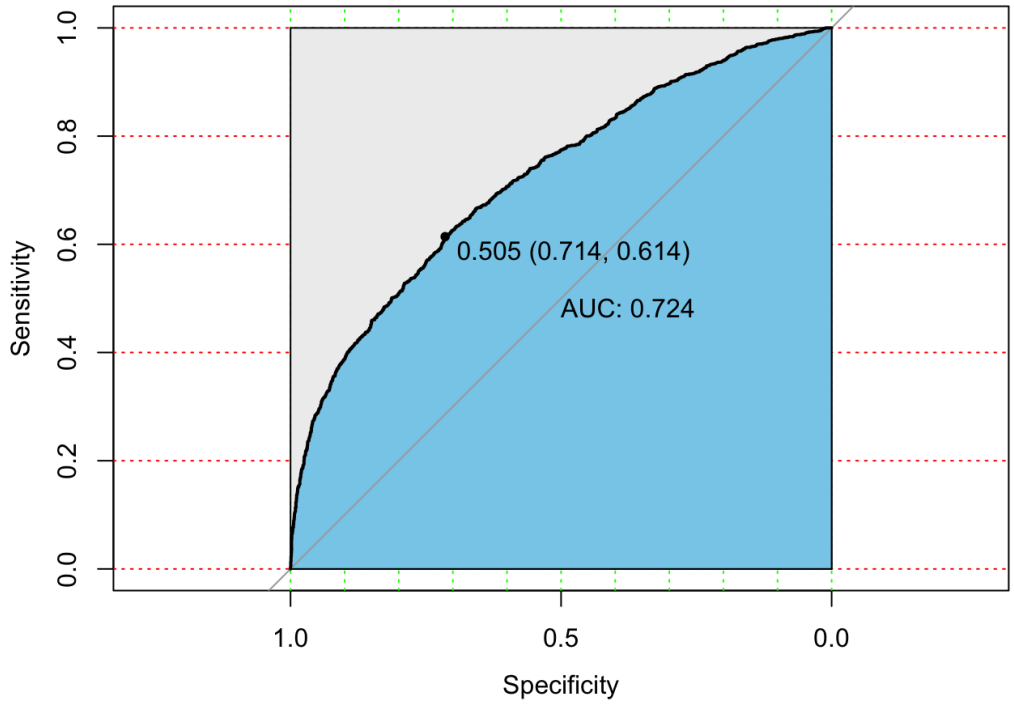
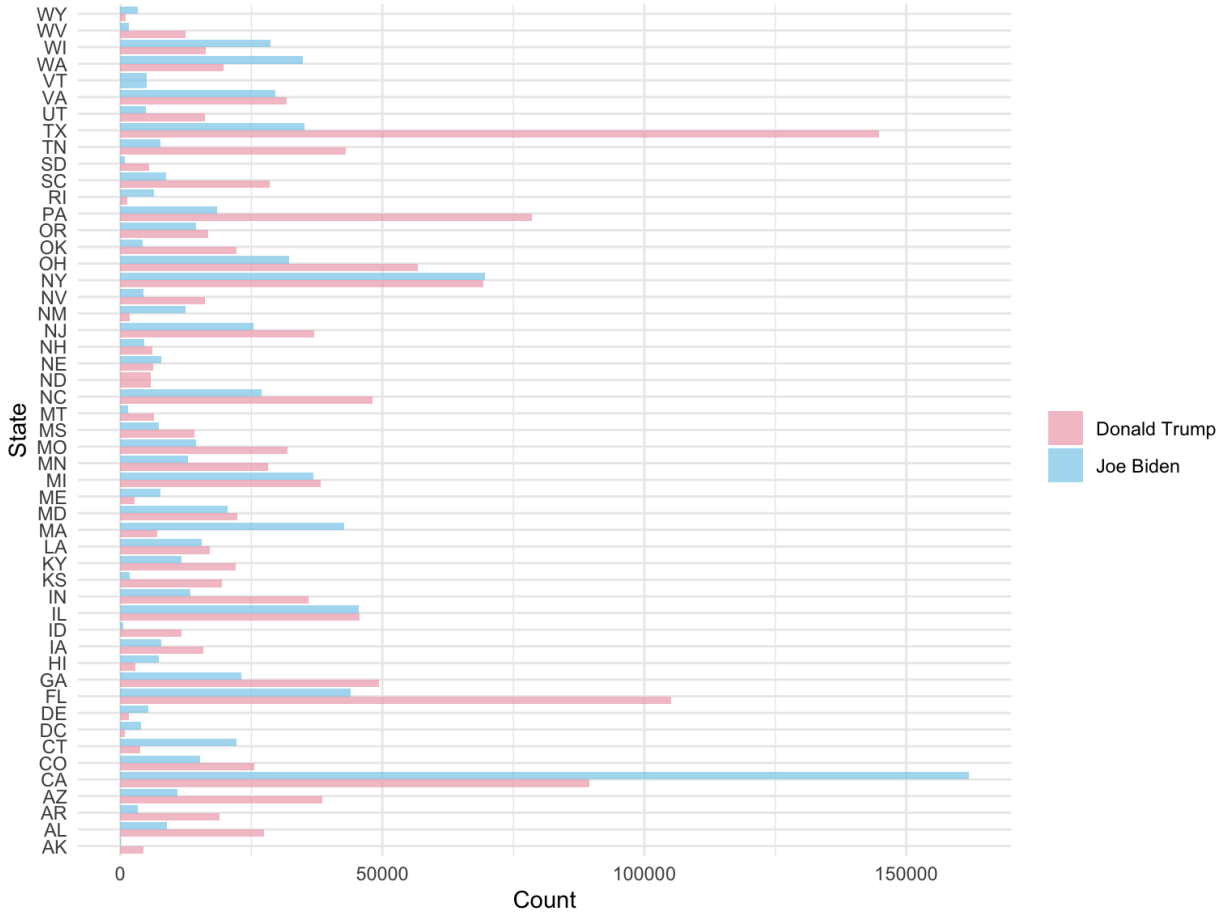


Figure 10: ROC Curve Diagnostics of Multilevel Logistics Regression Model, with the accuracy of 72.4%

Results

Applying the model in census data, which had over two millions of observations with certain personal weight added to each, we hope to have a big picture of voting behaviors of all Americans and hopefully forecast the election result.

By observing Figure 10, which represent for the voting preferences of the weighted population in each state, we could see that Trump leads most of the large states that cover more electorates. For example, besides Texas (which is a traditional Republican State), winning Pennsylvania and Florida contributes a lot on the sucess of Trump. The only large states remained for Biden is California, which is also a traditional Democratic State.



We then want to focus on the trend of the swing state, which have been proved as the key aspect of winning the election. We again filter out the eight swing states same as what we do in Figure 2 and surprisingly find that they are no more “swinging” under our model. Actually, most of the states voters favors Trump, and Trump is definitely ahead a lot in Florida and Pennsylvania.

Figure 11. Predicted vote result by each state: longer blue bar for Joe Biden taken the state and longer pink bar for Donald Trump taken the state.

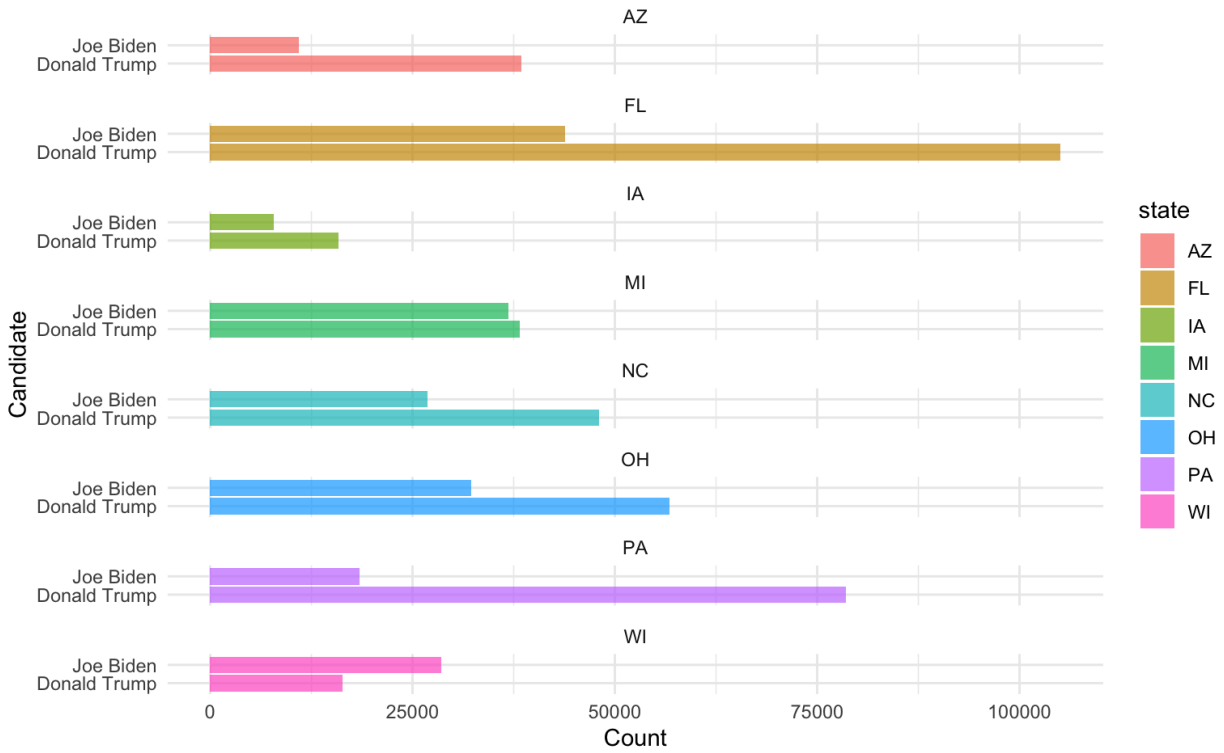


Figure 12. Predicted vote result in the eight swing state:A longer bar represents more people of this race group will vote for certain candidate.

Discussion

The overall outcome

Based on the results from data and models, we are going to make some real life discussion in this section. To start with, we need to have a brief understanding on the U.S. election system. The process is generally called “Electoral College”. To put it simply, presidents of the US are not directly selected by citizens but by electors from each state. The electors’ choices are determined by the citizens represented by them. Two major political parties in the US 2020 Election are the Democratic Party with candidate Biden and the Republican Party with candidate Trump. For Nebraska and Maine, partial electors will choose the candidate with more supports in the whole state, the other electors nevertheless, will determine according to the situation of supports for each electoral district within the state (e.g. Maine is divided in to two electoral districts. There are four electors from Maine. If the whole state has over half of people voting for Biden, and the two district has one voting more for Biden and one for Trump, then the final voting from Maine will be three for Biden and one for Trump.). For the other states, all the electors from that state will vote for the one who wins more voting in that state.

In our forecasting, we predict that Trump would win the election at 58% versus Joe Biden at 42%. Looking this result more in detail, the Democratic and the Republican both have fixed supporters, such as Massachusetts, New York and the other Blue States for Democratic, along with Texas, Utah and the other Red States for Republican. The final result is largely dependent on the swing states. It seems that number of Blue states is smaller than that of Red states. However, since the number of electors varies across states and the “winner-take-all” rule for each state, the polls from these states with fixed choice for two parties are almost the same. Under this circumstance, the swing states’ votes are extremely decisive.(2020)

Candidate	Total Electoral Vote Gained By The Candidate	Total Electoral Vote Percentage Gained By The Candidate
Donald Trump	313	58.18%
Joe Biden	225	41.82%

Table 1. The predicted total electoral vote result for Trump and Biden

Understanding Result

As the result predicted by the model has shown, Trump would win the 2020 U.S. presidential election. However, his win is yet carved in stone. From Figure 11, unlike the result presented by Figure 2, we could see that the Trump wins most of the swing states except for Wisconsin. One remarkable state is Florida, in which Trump is over twice ahead of Biden. However, the voting result in certain states such as MI (16 electoral votes), IL(20 electoral votes) and VA(13 electoral votes) are all closed games. If Biden could take these three states back at the election date, he could get 274 total electoral votes overturn the battle. Therefore, Biden should pay more attention on those states and it is very likely to get it back at the election date. As of Trump, he should consolidate what he already had, and make sure he can keep the lead in certain swing states.

However, it should come to notice that in Indiana, all the people vote for Trump, which is not going to happen in practice. That might be caused by the lack of training dataset while training the model, which caused monotony behavior under certain conditions.

In addition, the voting behavior in traditional “mining zones” for both the Democratic Party and the Republican Party is predicted to vote as expected, which also increase the confidence of the model. Take California and Texas as example, we could see increadiable lead for Trump in Texas and similar for Biden in California.

Small World vs Large World (McElreath 2020)

For the small world of our self-contained model, the accuracy score of our forecasting model is 77.6%, which is high enough to proof the fitness of the model to the small world. From the result of survey data, we also found that gender and races have obvious pattern of effect on voting intention. Therefore for our multilevel regression model’s small world, it is reasonable to select gender and race as cells to build the model and generate results.

For the large world of broader context, there are factors that have not been considered and imagined in the small world. For example, the small world only focus on fitting the survey result, however, the situation of the respondents that are not covered in the small world while exist in the large world, has never been imagined. Also, in the large world with more realistic context, the effect of other demographic factors such as different social status, working industry and life style are not in the small world’s consideration, as well. Overall, in a broad-context large world, it is not as fitted as it was in the small world. And the issues noticeable here are just the things that we could improve in the future.

Weaknesses and next steps

Such outcome of the predicted model is indeed surprising, as it contradicts to most of the election poll results. Many of the election poll shows that Biden is ahead of Trump in support rate, especially seeing Trump's incapability of handling COVID-19 crisis. Take the recent poll result of CNN, we can see that Biden is still 10% ahead of Trump ("View Latest 2020 Presidential Polling," n.d.). That leads to the reflection of the potential weaknesses and problems of our model.

As what we have mentioned above, the result of our model is highly dependent on the accuracy of the survey responsed data. However, as the survey took place in four months ahead of the election, the survey result cannot be translated to the exact voting result in the election date, with the question in the survey phrased as "If the election for president were going to be held now…".Referring back to the 2016 election held between Hilary Clinton and Donald Trump, the poll demonstrated that Clinton was ahead of Trump in certain swing states where she failed in the election day. Especially under the COVID-19 pandemic, as the second-wave of the plague comes back in winter, many people might not consider to vote due to safety concern.

Moreover, since the opinion polls were conducted in English only, which is expected to weaken the representativeness of our race variable in the model. By observing Figure 5, we could see that there are not enough minority race groups sample compared with the white group. It is reasonable to infer that the mono-linguistic design of the survey kept those who are not expert in English from express their opinion. Second, the survey response might not be representative in terms of insufficient sample size. Note that the original dataset contains of 6479 responses, which was further reduced to 4607 after doing necessary cleaning process. That limits the behavior of our model to some extent, as we could not set adequate variables as our cell to increase the accuracy of the model. The cell of our model, as mentioned in the model part, only contains 14 values, which is far less than the expected standard to make accurate prediction. However, due to the restriction of computational capacity and lack of data, we could not build a more complex model so far. For doing further analysis, more training data is needed for constructing a more accurate cell variable by combining more variables together.

Additionally, in the cleaning process, observations with NA values were all simply removed rather than being approximately handled. That somewhat cripples our model again as the sample size is further reduced. While filtering data, we pick only people who indicated that they intended to vote, or they were not sure if they would vote. The model assumes those people who were not sure whether to vote but indicate their preferred candidate would go to vote at the election day.

One more thing needs to be mentioned is the time-effectiveness of model, as the survey data was collected at the end of June and could possibly be outdated. Especially under the COVID-19 pandemic, the crisis caused many deaths and many people lost their jobs. All these circumstances could harm Trump's reputation of leadership.

The last weakness that must be addressed is the way we calculate the final result, as we simply assume Nebraska and Maine also follow the same electoral vote methods as other states. However, candidates may separate the electoral vote in these two states depending on the vote result within each of the congressional district. For instance, in 2016 election between Trump and Hilary, even though Hillary won the overall vote in Maine, Trump still got one electoral vote from Maine by winning one congressional district. However, that weakness does not necessarily affect the final outcome. In fact, the model predict that Biden will win both Nebraska and Maine, but still lose the overall election.

Next Steps

• Survey with closer date to election for the future predictions: Since the election will take place just in a few days, the improvements should be more focus on the prediction for the next time. The survey data will be closer to the citizens final choice as the date of survey data be closer to the election. Getting a survey data closer to election will make the model and results more reflective.

• Larger sample size: The respondents for our survey data this time has only 6479 responses without cleaning. The lack of sample for certain demographic groups of people affect the accuracy of the following logistic model with post-stratification. With a larger sample size, the data for all factors, especially for minority demographic groups, will be more adequate to find out more accurate results.

• Mitigate the effect of inconsistency between survey and post-stratification data: In this analysis, our survey data analysis moved out the effect of inactive voters while the model assumed that they will all go to work. In the future forecasting, this can be eliminate by selecting or gathering a more informative post-stratification data in order to better fit the survey data.

• Consider Nebraska and Maine's special cases: We simplify the model here by considering Nebraska and Maine the same electoral system as the other states. However, since this does not fit to the reality, for future forecasting, it could be improved by adding this affect of different electoral system into the model for these two states.

References

Agreement Among the States to Elect the President by National Popular Vote. (2020, March 08). Retrieved November 02, 2020, from https://www.nationalpopularvote.com/written-explanation (https://www.nationalpopularvote.com/written-explanation)

Alexander, Rohan. 2020. “Rohan Alexander: A Review of ’Forecasting Elections with Non-Representative Polls’.” https://rohanalexander.com/posts/2020-02-11-a-review-of-forecasting-elections-with-non-representative-polls/ (https://rohanalexander.com/posts/2020-02-11-a-review-of-forecasting-elections-with-non-representative-polls/).

Brownlee, Jason. 2020. “Probabilistic Model Selection with Aic, Bic, and Mdl.” *Machine Learning Mastery*. https://machinelearningmastery.com/probabilistic-model-selection-measures/ (https://machinelearningmastery.com/probabilistic-model-selection-measures/).

McElreath, Richard. 2020. “Small Worlds and Large Worlds.” In *Statistical Rethinking a Bayesian Course with Examples in R and Stan*. CRC Press, Taylor & Francis Group.

Narkhede, Sarang. 2019. “Understanding Auc - Roc Curve.” *Medium*. Towards Data Science. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 (https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5).

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/ (https://www.R-project.org/).

Robin, Xavier, 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: 77.

Tausanovitch, Chris, and Lynn Vavreck. 2020. “New: Second Nationscape Data Set Release.” *Democracy Fund Voter Study Group + UCLA Nationascape*. https://www.voterstudygroup.org/publication/nationscape-data-set (https://www.voterstudygroup.org/publication/nationscape-data-set).

Team, MPC UX/UI. n.d. “U.S. CENSUS Data for Social, Economic, and Health Research.” *IPUMS USA*. https://doi.org/10.18128/D010.V10.0 (https://doi.org/10.18128/D010.V10.0).

“View Latest 2020 Presidential Polling.” n.d. *CNN*. Cable News Network. https://www.cnn.com/election/2020/presidential-polls (https://www.cnn.com/election/2020/presidential-polls).

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. “Forecasting Elections with Non-Representative Polls.” *International Journal of Forecasting* 31 (3): 980–91. https://doi.org/10.1016/j.ijforecast.2014.06.001 (https://doi.org/10.1016/j.ijforecast.2014.06.001).

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686 (https://doi.org/10.21105/joss.01686).

2007. https://doi.org/10.17226/11901 (https://doi.org/10.17226/11901).

2020. https://www.nationalpopularvote.com/written-explanation (https://www.nationalpopularvote.com/written-explanation).