Projecting a Close Call Victory for Biden in the 2020 US Presidential Elections*

51% Majority Vote, 274 Electoral College Votes and a bit of Luck

Nayan Saxena

12 April 2022

Abstract

The 2020 US Presidential elections took place not only during a global pandemic but also after several major events like the "Black lives matter" movement. They also occurred after a shocking 2016 US presidential outcome where several polling based models incorrectly predicted the outcome of the election–revealing several issues with statistical approaches to election forecasting. In this paper we attempt to forecast the 2020 US Presidential elections using multilevel-regression with post-stratification using survey data collected prior to the election. Our model predicts an expected 51% +-6% chance of a majority vote and 274 electoral college votes for Joe Biden designating a clear victory for the Democratic presidential nominee across both fronts.

1 Introduction

As the COVID-19 pandemic swept through the United States in early 2020, a challenging contest between the Republicans and Democrats was taking place for the upcoming 2020 US Presidential Election. This was set against the somber backdrop of the various "Black Lives Matter" protests caused due to the untimely death of George Floyd and hate crimes against Asian-Americans which had a polarising effect on the population. Given these tumultuous times, the polling data prior to the election alongside the public perception of each candidate were more volatile than usual, with different events and occurrences altering each candidate's image. The main focus of our paper would be to quantify some of this uncertainty and capture the public sentiment through statistical modeling approaches which have been used to predict electoral outcomes in the past.

When it comes to statistical modeling approaches, the 2020 election was unique in the sense that the previously held 2016 presidential election revealed several inaccuracies and hidden biases within the polling data which failed to capture a large fraction of supporters of Donald Trump (Mercer, Deane, and McGeeney 2016). This led to inaccurate predictions at the time, with most polling data not truly encapsulating the public opinion toward Donald Trump. With Trump's eventual 2016 victory, it has eventually been established that statistical machine learning models still performed better at capturing the election outcome as compared to simpler inferences drawn through polling data (Tien and Lewis-Beck 2016).

Motivated by this superlative performance, we focus our efforts through this paper on creating a parsimonious model that captures the 2020 election sentiment. We use multi-level regression with post-stratification (MRP) as the primary method, with a relatively simple logistic regression model that is trained to predict the likelihood of Joe Biden winning the 2020 election. This method employs two datasets: a non-representative dataset to fit the model, and then a representative dataset to weigh the model prediction outcomes. This model has historically been a popular choice to model voter behavior including election turnout (Ghitza and

*Code and data are available at: https://github.com/the-infiltrator/2020USForecast

Gelman 2013), and forecast elections with particularly non-representative polling data (Wang et al. 2014). Furthermore, MRP models have in the past successfully predicted the 2017 UK general election results (Wong 2019), and might help us more accurately capture the true picture given that trends leading up to the 2020 US election demonstrate a close call between the candidates.

Since MRP uses two datasets we source our model training data or non-representative sample from the Democracy Fund+ UCLA Nationscape survey of US citizens (Tausanovitch and Vavreck 2020). The representative sample or post-stratification dataset is derived from the 2019 American Community Survey (Steven Ruggles and Sobek 2020), which is used to weigh the prediction cells for better predictions. Across both datasets and within our model we consider the variables: state, age group, education level, Hispanic, race, and household income. This study extensively uses the R programming language (R Core Team 2019) and Tidyverse packages (Allaire et al. 2019) to fit the model and engineer the respective features.

Inference based on post-stratification data from our MRP model reveals an expected 51% ± 6% chance of a majority vote with 274 electoral college votes on average in favor of Joe Biden designating a clear victory for Joe Biden across both fronts. It is important to note here that in 2016 Donald Trump became president based on the electoral college win (votes above 270) which plays a major role in the US electoral process. In summary, this paper is organized into 4 major sections with an outline of the datasets used in Section 2; mathematical detail of the modeling framework in Section 3 along with the model coefficients visualized and provided in the form of a table, and a report on our model predictions in Section 4 for each state compared against the survey data.

2 Data

In this paper our main modelling approach will be multi-level regression with postratification (MRP) which requires a non-representative dataset to fit the model and then a representative dataset to weigh the model prediction outcomes. To better assess the public sentiment before the elections, we use the Nationscape survey data collected between October 1-7, 2020 which lies about a month before the United States elections. Furthermore, for post-stratification we extract important variables from the IPUMS America Census Service data.

2.1 Feature Selection

For this study we consider prior research to choose common variables across both datasets for consideration. The main variables throughout this study are: state, age group, education level, hispanic, race and household income. Age has been shown to be a major factor in the political inclination of voters with prior work showing a correlation between age and republican sentiments (Desilver 2014) while other work showing that political ideologies stay relatively consistent over time (J. C. Peterson, Smith, and Hibbing 2020). Furthermore, while every state has their own political inclination we go further to also look at race as well as if a person identifies as hispanic. Race has been shown to be a major driver historically of political idelogies especially during the 2020 election which are ocurring right after the "Black Lives Matter" movement. The consideration of hispanic identity in our study is important because it has been shown that hispanic people are mostly democractic leaning and have mostly voted for democratic candidates in the past (Leon 2020). Education level is considered because of patterns that emerged after the 2016 election, where the majority of voters did not attend university (CAWP 2020).

2.2 Survey Data

The primary dataset being used to train the model is the survey data that is sourced from the Democracy Fund+ UCLA Nationscape survey of US citizens (Tausanovitch and Vavreck 2020) collected in the first week of October 2021. The participants were selected from Lucid, a market research platform, based on various

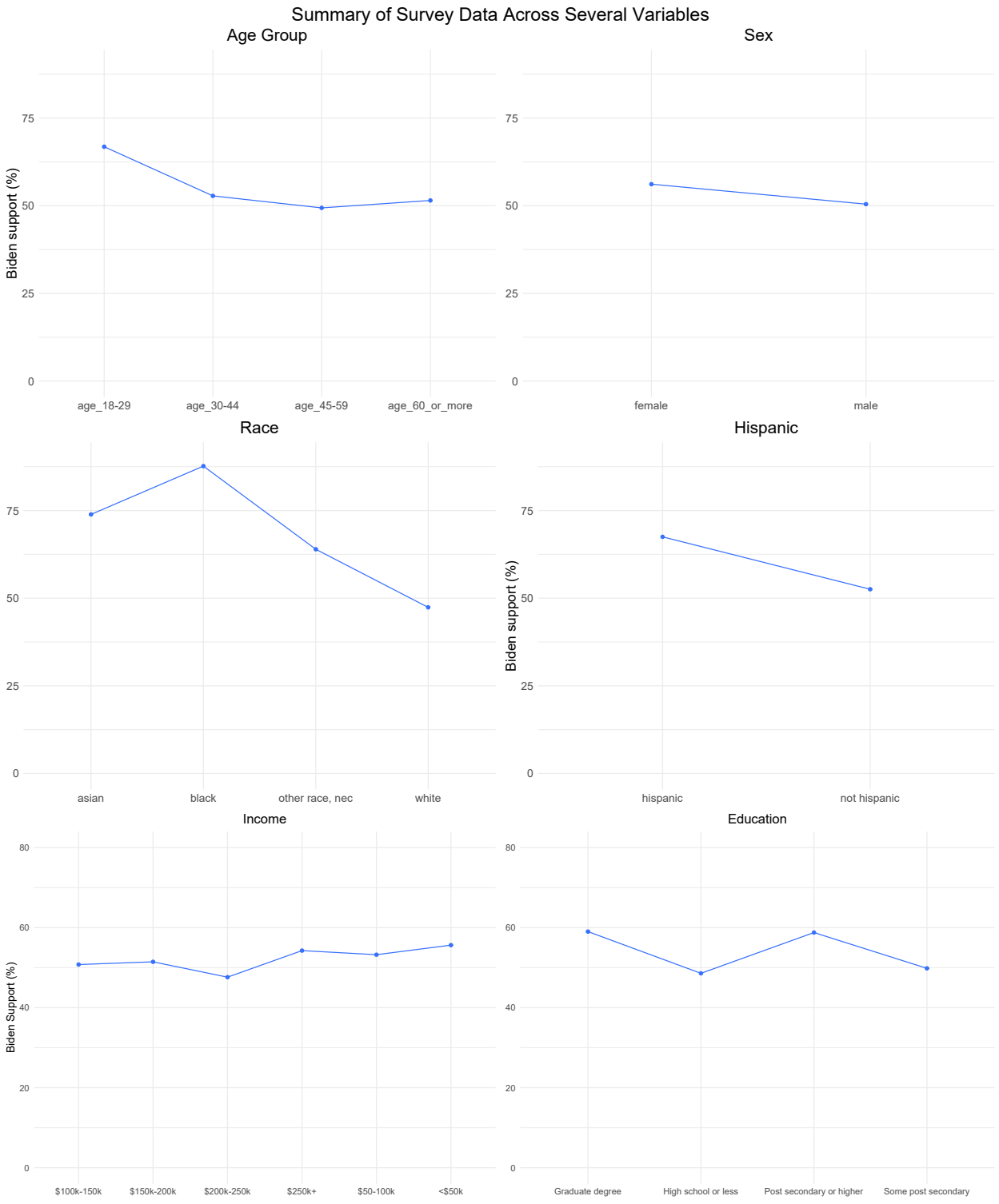

Figure 1: Summary of several important variables present within the survey data

criterias randing from their education, age, gender, income, ethnicity and much more (Tausanovitch and Vavreck 2020).

As observed in Figure 2, there is a small but noticeable difference in partisanship between male and female voters with around 6% more support for Joe Biden based on the survey.

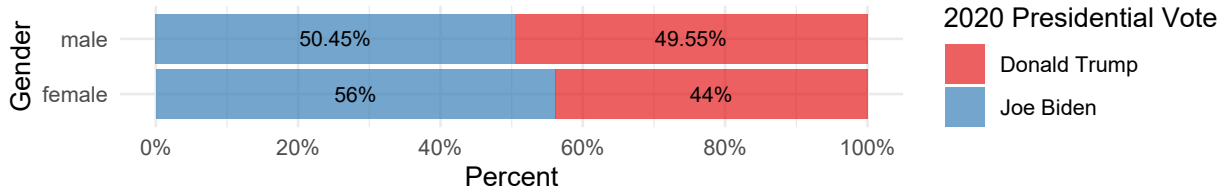

Figure 2: 2020 Percentage of Voters by Sex for Each Candidate

Furthermore, when considering different age groups, as seen in Figure 3 a major difference can be seen between voter sentiment for lower age-groups which are more polarised towards supporting Joe Biden. It should be noted here, that we do not observe a correlation between increasing age and conservative support as noted by some studies. The support for either candidates appears to be almost constant for other age groups except the youngest, where almost approximately 70% of the surveyed people showed support for Joe Biden.

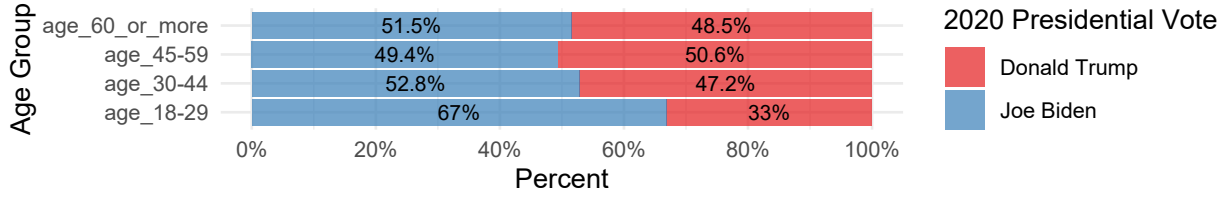

Figure 3: 2020 Percentage of Voters by Age for Each Candidate

When considering race, shown in Figure 4 we notice strong patterns emerging within the survey data with a very large majority of black and asian respondents showing support for Joe Biden. Amongst the black population almost approximately 90% of the respondents indicated support for Joe Biden alonngside around 75% of the asian people. A common trend seems to be emerging here which shows more democratic support in the 2020 elections by people of other races as well as people who are black and asian. This can be explained as a natural consequence of the "Black Lives Matter" movement of 2020 and the various hate-crimes against Asian people during the Trump presidency.

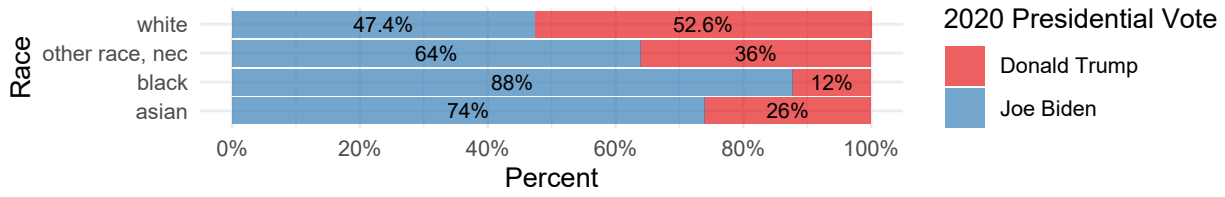

Figure 4: 2020 Percentage of Voters by Race for Each Candidate

Echoing, a similar sentiment as seen previously in Figure 4, people identifying as hispanic also showed more support for Joe Biden as compared to otherwise, shown in Figure 5. The general consensus based on these results is that race and hispanic origin seem to have a significant bearing on the outcome of this election and therefore must be given importance throughout our analysis.

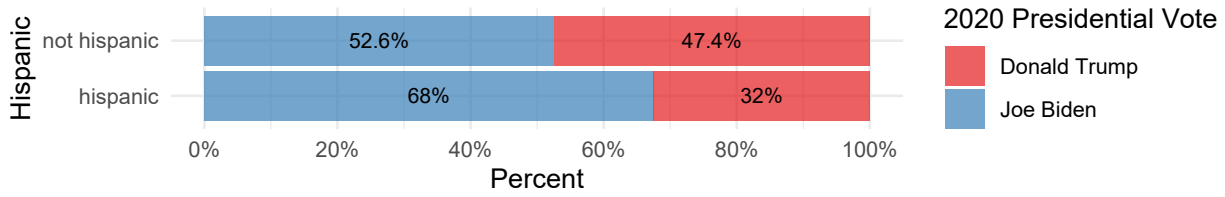

Figure 5: 2020 Percentage of Voters by Hispanic Identity for Each Candidate

Finally, a detailed summary of survey data can also be seen in Figure 1 which outlines the support for Biden for each of these variables. It should be noted that we also observe an interesting constant trend in terms of candidate support for people coming from different household income brackets. This observation is interesting as income has been shown previously to be a major indicator of partisanship (E. Peterson 2016) – something which is not captured within our survey data.

2.3 Post-stratification Data

Table 1: Comparison of Nationscape (Survey Data) and ACS Demographics (Postratification Data)

Variable	n	Nationscape	ACS
Age Group			
age_18-29	811	17%	17%
age_30-44	1212	26%	22%
age_45-59	1213	26%	24%
age_60_or_more	1400	30%	37%
Sex			
female	2862	62%	51%
male	1774	38%	49%
Race			
asian	226	5%	7%
black	503	11%	10%
other race, nec	258	6%	3%
white	3649	79%	80%
Hispanic			
hispanic	440	9%	12%
not hispanic	4196	91%	88%
Education			
Graduate degree	383	8%	12%
High school or less	1165	25%	39%
Post secondary or higher	1916	41%	28%
Income			
Some post secondary	1172	25%	22%
$100,000 to $149,999	579	12%	18%
$150,000 to $199,999	206	4%	9%
$200,000 to $249,999	84	2%	4%
$250,000 and above	94	2%	12%
$50,000 to $99,999	1345	29%	30%
Less than $50,000	2328	50%	28%

Another dataset used in our analysis is the post-stratification data which is derived from the 2019 American Community Survey (Steven Ruggles and Sobek 2020), which is a large random sample which will help us weigh the cells during prediction to better approximate the sentiments of the American people. A comparison of the differences between our survey data (non-representative) and post-stratification data are shown in Table 1 where it can be observed that there are certainly differences between the two samples. Specifically, there is roughly a 10% difference between both male and female respondents across both surveys with approximately 10% difference also emerging in the respondents with high-school education or lower or post-secondary education or higher. Finally, another major difference across both surveys is revealed in respondents who have a household income below $50,000.

3 Model

To model the election outcome we employ multilevel regression with post-stratification (MRP) which has been shown to be a powerful tool to examine state and individual differences across populations. As explained in Kennedy and Gelman (n.d.), MRP can come in handy when we have two datasets with the goal of investigating common variables across both. In our particular case, we use the aforementioned Nationscape survey data to train our model, and then post-stratify our model using the IPUMS Census data. The model used is a logistic regression model that outputs 0 or 1 to predict if the vote is for Donald Trump or Joe Biden respectively. We use R Core Team (2019) programming language to fit our model which takes the following mathematical form,

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \beta_0 + \beta_1 x_{\text{sex}} + \beta_2 x_{\text{agegroup}} + \beta_3 x_{\text{race}} + \beta_4 x_{\text{state}} + \beta_5 x_{\text{education}} + \beta_6 x_{\text{hispanic}} \tag{1}$$

$$+ \beta_7 x_{\text{household income}} + \beta_8 x_{\text{race}} * x_{\text{hispanic}} \tag{1}$$

A central use of our model will be to predict future likelihood of voting for Biden for different voters which is denoted by p in Equation 1. Furthermore, each β_i indicates a coefficient estimate for each respective variable. We also incorporate an interaction term based on the high influence of hispanic and racial identities observed in our survey data as seen in Figures 4 and 5. This term can be simply thought of as a variable that accounts for additional black-hispanic, asian-hispanic and otherrace-hispanic identities. We should also note that this model is limited due to it simply being binary and also not considering individual level differences in race and gender which can be accounted by a more Bayesian approach. Furthermore, a more sophisticated but simple mixed model can also come in handy to incorporate state-level or individual-level differences. The model is fitted using R programming language (R Core Team 2019) and the model coefficient estimates are outlined in Table 2 with each estimate visualized alongwith the 95% confidence interval in Figure 6.

4 Results

Inference based on post-stratification data from our MRP model reveals an expected 51% $\pm$ 6% chance of a majority vote with 274 electoral college votes on average in favor of Joe Biden designating a clear victory for Joe Biden across both fronts. It is important to note here that in 2016 Donald Trump became president based on the electoral college win (votes above 270) which plays a major role in the US electoral process.

As seen in Figure 7 we observe that the proportion of people voting for Joe Biden and the predictions from our MRP model align closely, with most values overlapping or lying within the 95% confidence intervals as denoted by the error bars. It should be noted that there are a few incorrect predictions by our model that can certainly effect the overall outcome of the election and some that do not overlap or resonate with our training data.

Finally we show using Figure 8 a more clearer picture of the voting outcomes for each states with some states clearly being in the democratic majority and some in the conservative.

Table 2: Coefficients from the Model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.52	0.14	3.79	0.00	0.25	0.78
gendermale	-0.07	0.01	-4.74	0.00	-0.10	-0.04
age__groupage__30-44	-0.07	0.02	-3.32	0.00	-0.12	-0.03
age__groupage__45-59	-0.10	0.02	-4.27	0.00	-0.14	-0.05
age__groupage__60__or__more	-0.06	0.02	-2.87	0.00	-0.11	-0.02
stateicpalaska	0.00	0.16	0.03	0.98	-0.30	0.31
stateicparizona	0.19	0.07	2.78	0.01	0.06	0.33
stateicparkansas	0.17	0.08	2.12	0.03	0.01	0.33
stateicpcalifornia	0.31	0.05	5.90	0.00	0.21	0.41
stateicpcolorado	0.35	0.08	4.41	0.00	0.19	0.50
stateicpconnecticut	0.28	0.08	3.68	0.00	0.13	0.43
stateicpdelaware	0.24	0.10	2.35	0.02	0.04	0.44
stateicpdistrict of columbia	0.30	0.24	1.24	0.22	-0.17	0.76
stateicpflorida	0.23	0.05	4.38	0.00	0.13	0.34
stateicpgeorgia	0.09	0.06	1.48	0.14	-0.03	0.21
stateicphawaii	0.22	0.12	1.76	0.08	-0.03	0.46
stateicpidaho	0.29	0.12	2.43	0.02	0.06	0.53
stateicpillinois	0.23	0.06	4.02	0.00	0.12	0.34
stateicpindiana	0.11	0.06	1.73	0.08	-0.01	0.24
stateicpiowa	0.26	0.08	3.32	0.00	0.11	0.42
stateicpkansas	0.20	0.08	2.57	0.01	0.05	0.35
stateicpkentucky	0.12	0.07	1.68	0.09	-0.02	0.26
stateicplouisiana	0.01	0.07	0.20	0.85	-0.13	0.16
stateicpmaine	0.25	0.12	2.14	0.03	0.02	0.48
stateicpmaryland	0.25	0.07	3.38	0.00	0.10	0.39
stateicpmassachusetts	0.36	0.07	5.17	0.00	0.22	0.49
stateicpmichigan	0.26	0.06	4.34	0.00	0.14	0.38
stateicpminnesota	0.19	0.06	3.04	0.00	0.07	0.32
stateicpmississippi	-0.09	0.09	-0.99	0.32	-0.27	0.09
stateicpmissouri	0.13	0.06	2.04	0.04	0.00	0.26
stateicpmontana	0.26	0.13	2.02	0.04	0.01	0.50
stateicpnebraska	0.23	0.12	1.89	0.06	-0.01	0.47
stateicpnevada	0.12	0.08	1.44	0.15	-0.04	0.28
stateicpnew hampshire	0.20	0.13	1.57	0.12	-0.05	0.46
stateicpnew jersey	0.23	0.06	3.76	0.00	0.11	0.35
stateicpnew mexico	0.32	0.12	2.69	0.01	0.09	0.56
stateicpnew york	0.28	0.05	5.12	0.00	0.17	0.38
stateicpnorth carolina	0.18	0.06	2.79	0.01	0.05	0.30
stateicpnorth dakota	0.37	0.16	2.27	0.02	0.05	0.69
stateicpohio	0.17	0.06	3.05	0.00	0.06	0.29
stateicpoklahoma	0.07	0.08	0.82	0.41	-0.09	0.23
stateicporegon	0.28	0.08	3.64	0.00	0.13	0.43
stateicppennsylvania	0.20	0.06	3.50	0.00	0.09	0.32
stateicprhode island	0.13	0.13	1.00	0.32	-0.13	0.40
stateicpsouth carolina	0.08	0.07	1.12	0.26	-0.06	0.23
stateicpsouth dakota	0.34	0.17	2.00	0.05	0.01	0.68
stateicptennessee	0.15	0.07	2.25	0.02	0.02	0.29
stateicptexas	0.17	0.05	3.07	0.00	0.06	0.27
stateicputah	0.23	0.09	2.48	0.01	0.05	0.42
stateicpvermont	0.38	0.27	1.39	0.17	-0.16	0.92
stateicpvirginia	0.19	0.06	3.07	0.00	0.07	0.32
stateicpwashington	0.36	0.07	5.40	0.00	0.23	0.49
stateicpwest virginia	0.10	0.10	0.96	0.33	-0.10	0.29
stateicpwisconsin	0.27	0.07	3.99	0.00	0.14	0.40
stateicpwyoming	-0.28	0.18	-1.51	0.13	-0.63	0.08
education__levelHigh school or less	-0.16	0.03	-5.47	0.00	-0.22	-0.10
education__levelPost secondary or higher	-0.01	0.03	-0.40	0.69	-0.06	0.04
education__levelSome post secondary	-0.12	0.03	-4.00	0.00	-0.17	-0.06
raceblack	0.17	0.14	1.17	0.24	-0.11	0.44
raceother race, nec	0.11	0.13	0.83	0.41	-0.15	0.37
racewhite	0.02	0.13	0.12	0.90	-0.24	0.27
hispanicnot hispanic	0.09	0.13	0.73	0.47	-0.16	0.35
household__income$150,000 to $199,999	-0.04	0.04	-0.93	0.35	-0.11	0.04
household__income$200,000 to $249,999	-0.09	0.06	-1.57	0.12	-0.19	0.02
household__income$250,000 and above	-0.02	0.05	-0.47	0.64	-0.13	0.08
household__income$50,000 to $99,999	0.02	0.02	1.01	0.31	-0.02	0.07
household__incomeLess than $50,000	0.04	0.02	1.87	0.06	0.00	0.09
raceblack:hispanicnot hispanic	0.07	0.15	0.46	0.64	-0.22	0.36
raceother race, nec:hispanicnot hispanic	-0.28	0.14	-1.94	0.05	-0.56	0.00
racewhite:hispanicnot hispanic	-0.23	0.13	-1.70	0.09	-0.49	0.03

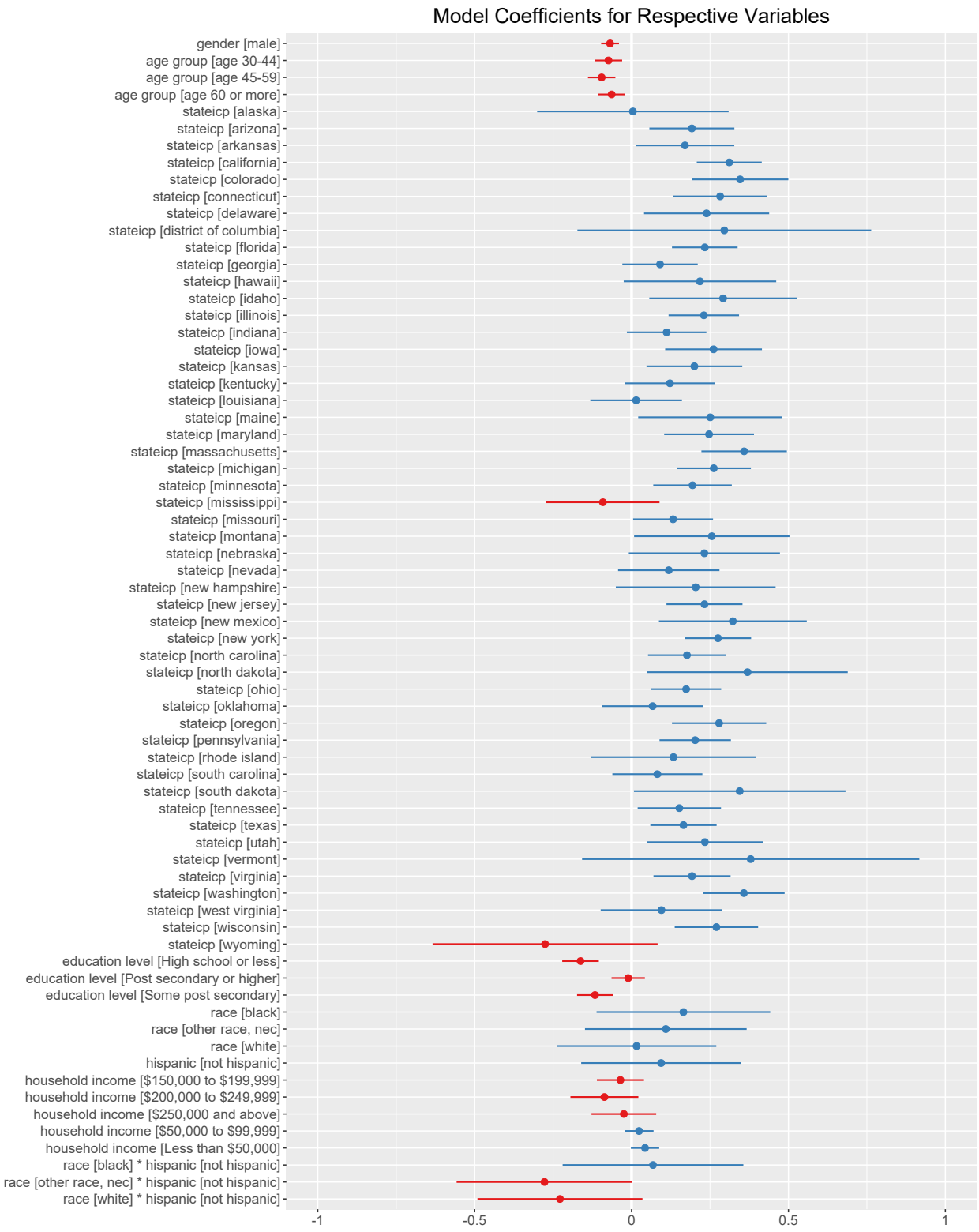

Figure 6: Graph of model coefficients with their respective confidence intervals

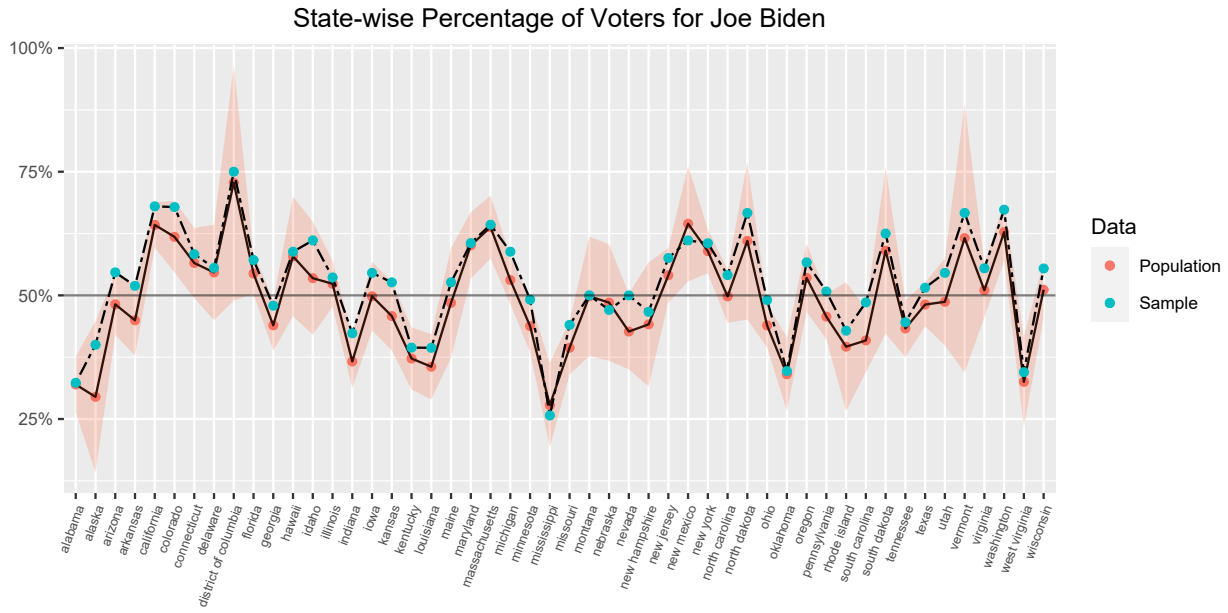

Figure 7: State-wise Percentage of Voters for Joe Biden

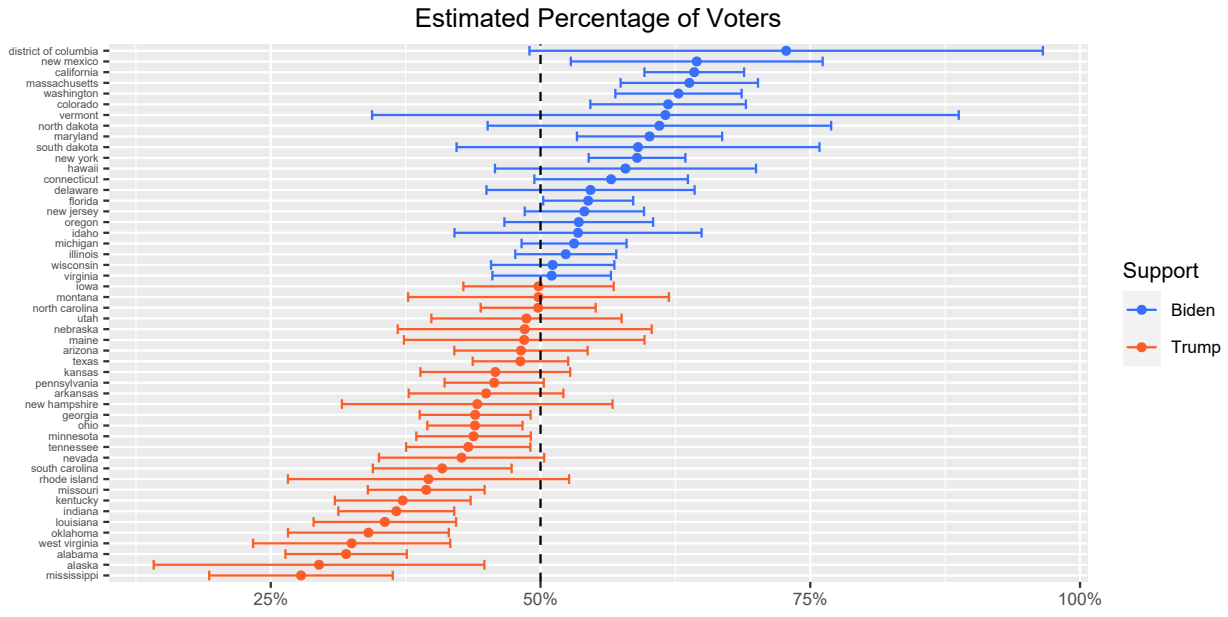

Figure 8: 2020 Percentage of Voters by State

Table 3: Estimated Percentage of Biden Voters

	Lower Estimate	Mean Estimate	Upper Estimate
Number of Colleges	163	274	400
Proportion of Vote (%)	45	51	57

5 Discussion

The results from our paper demonstrate that our model predicts a victory for Democratic presidential candidate Joe Biden, with a close call between the votes. Based on Table 3 it can be observed that our model predicts an expected 51% +-6% chance of a majority vote and 274 electoral college votes for Joe Biden designating a clear victory for the Democratic presidential nominee across both fronts. Furthermore, based on the estimated proportions for every state, it seems like the swing states that can completely alter this close election outcome are Virginia, Wisconsin, Iowa, and Montana which lie quite close to the 50% reference threshold. The model seems to also accurately capture the effect of different age groups and races based on the coefficient estimates seen in Table 2 and observed in Figure 6. This is consistent with our preliminary findings based on survey data in Section 2.2.

Another consistent and common trend observed through our model coefficients is that men are more likely to vote for Donald Trump as compared to women as shown in Table 2. Furthermore, we notice an almost consistent voting pattern for different age groups. There are a few states which do not contribute significantly to the model predictions and therefore there is some uncertainty surrounding those state predictions. These states have a high p value in Table 2, some of which are Kentucky, Louisiana, and Oklahoma. Our findings also show that throughout both survey data and our model estimates, different income brackets largely share the same political ideologies. In terms of our model, only the lowest income households have minor statistical significance in contributing to the predictions.

Through this analysis, a major finding is that race and Hispanic identity are significant drivers of polling behavior. Our model captures this intricate behavior through an added interaction effect between the two variables towards the end, to help capture identities like black-Hispanic and Asian-Hispanic for prediction. It should further be noted that while our model is simple, it can be further improved by further hierarchical multilevel modeling approaches that account for these individual-level differences in a more Bayesian fashion by placing a prior over their assumed values. Furthermore, we need to note that there are several other extraneous variables at play that can certainly significantly impact the election output.

During the months leading up to the election, USA was hit by the COVID-19 pandemic which has further polarised public opinion and is not captured by the surveys used for model training which is another limitation of our study. Several metrics like the social distancing policies and COVID-19 restrictions in different states have shown to be correlated with political ideology and can be major drivers of election behavior during the 2020 season. There is also the important consideration we need to keep in mind of our model over-fitting to the data we have, making it hard to use the model to draw repeated or regular inferences at regular intervals in time to make predictions. This would make the model limited in its usefulness as ideally, such models should be robust enough to be deployed in systems for weekly, monthly, or bi-weekly predictions on election outcomes and voter sentiment.

Such issues can be addressed by using better training techniques and also model comparison techniques. A likelihood ratio test or Analysis of Variance tests (ANOVA) should certainly be used to compare candidate models in such a scenario to determine which variables are useful and which are not. A more granular understanding of data can also be obtained if we use a multiple logistic regression model (multi-class classification) to account for other candidates within the election cycle. Future work can perhaps incorporate COVID-19 data into this analysis to examine the effect of the pandemic on partisanship. There can also be studies carried out focused on examining the data before and after the election, particulary after the January 6 seige of the Capitol building by Trump supporters to see the changes in public perception.

References

Allaire, JJ, Jeffrey Horner, Yihui Xie, Vicent Marti, and Natacha Porte. 2019. *Markdown: Render Markdown with the c Library 'Sundown'*. https://CRAN.R-project.org/package=markdown.

CAWP. 2020. "2020 Presidential Gender Gap Poll Tracker." *CAWP*. https://cawp.rutgers.edu/presidential-poll-tracking-2020.

Desilver, Drew. 2014. "The Politics of American Generations: How Age Affects Attitudes and Voting Behavior."

Ghitza, Yair, and Andrew Gelman. 2013. "Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups." *American Journal of Political Science* 57 (3): 762–76.

Kennedy, Lauren, and Andrew Gelman. n.d. "Know Your Population and Know Your Model: Using Model-Based Regression and Post-Stratification to Generalize Findings Beyond the Observed Sample." https://arxiv.org/pdf/1906.11323.pdf.

Leon, Luis de. 2020. "2020 Election: The Impact of the Latino Vote in Texas." *Kvue.com*. https://www.kvue.com/article/news/politics/vote-texas/election-2020-texas-latino-vote-impact/269-e27ced91-8f67-4bcf-bcc9-f0f56c5bc67f.

Mercer, Andrew, Claudia Deane, and Kyley McGeeney. 2016. "Why 2016 Election Polls Missed Their Mark."

Peterson, Erik. 2016. "The Rich Are Different: The Effect of Wealth on Partisanship." *Political Behavior* 38 (1): 33–54.

Peterson, Johnathan C, Kevin B Smith, and John R Hibbing. 2020. "Do People Really Become More Conservative as They Age?" *The Journal of Politics* 82 (2): 600–611.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Steven Ruggles, Ronald Goeken, Sarah Flood, and Matthew Sobek. 2020. "IPUMS USA: Version 10.0 [Dataset]." Minneapolis, MN: IPUMS. https://doi.org/https://doi.org/10.18128/D010.V10.0.

Tausanovitch, Chris, and Lynn Vavreck. 2020. "June 25-July 1, 2020 (Version 20200814)." Democracy Fund + UCLA Nationscape. https://www.voterstudygroup.org/downloads?key=a19c1a98-554a-474a-b513-70b2afb78ed2.

Tien, Charles, and Michael S Lewis-Beck. 2016. "In Forecasting the 2016 Election Result, Modelers Had a Good Year. Pollsters Did Not." *USApp–American Politics and Policy Blog*.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2014. "Forecasting Election with Non-Representative Polls." *International Journal of Forecasting* 31 (September). https://doi.org/10.1016/j.ijforecast.2014.06.001.

Wong, Sam. 2019. "What Is MRP and Can It Predict the Result of the UK General Election?" *New Scientist*. New Scientist. https://www.newscientist.com/article/2224783-what-is-mrp-and-can-it-predict-the-result-of-the-uk-general-election/.