2020 US Presidential Election Exit Poll Analysis

Kai Jeffreys, Benjamin Jia, Rachael Ren, Kobe Sarausad

2023-06-08

Introduction

The 2020 United States Presidential election marked a moment in our nation's history where political awareness and engagement all across the nation seemingly took an all time high. Given the context of the global Covid-19 pandemic where many people were forced to mail in ballots and the death of George Floyd, which sparked a national cry for outrage against systemic racism, the 2020 presidential election took on even greater significance. Many citizens believed that the future of the United States of America solely depended on the outcome of this election. The 2020 United States Presidential election showcased how important every vote can be and as a group we chose to analyze whether or not an individual's vote can be accurately predicted from the available factors. In the end we had two main research questions:

- Who do we predict will win the 2020 presidential election? (popular vote and electoral college)
- Which variables are significant for predicting who a respondent voted for?

We chose the data set, collected by Cornell, National Election Pool Poll: 2020 National Election Day Exit Poll. This data set contains a telephone survey and an election day exit poll. The survey consisted of demographic data, who each participant voted for, and other opinion questions.

Variable	**Variable Name**	**Data Type**	**Description**
Age	age10	ordinal	Age group participant is a part of
Sex	sex	nominal	Sex of the voter
Education	educ18	ordinal	Education level of the participant.
Income	income20	ordinal	Income bracket of participant.
Racism	racism20	ordinal	Is racism in the US... (the most important problem, one of many important problems, a minor problem, not a problem at all, omit)
Life	life	ordinal	Describes whether or not the participant has a pessimist/optimist view of the world.
Party	party	nominal	Party affiliation of the participant.
State	statenum	nominal	State where the participant is voting from.
City	sizeplac	ordinal	Size of the city where the participant is living in.
Pres	pres	nominal	Who the participant voted for in the 2020 election.
Weight	weight	continuous	How well the participant represents the population.
2016 vote	vote2016	nominal	Who the participant voted for in 2016 presidential election
Voting method	votemeth	nominal	Voting method (election day, by mail, or early in-person)
Abortion	abortion	ordinal	Which comes closest to your position? Abortion should be legal in all cases, legal in most cases, illegal in most cases, or illegal in all cases?
Climate change	climatec	nominal	Do you think climate change, also known as global warming, is a serious problem?

Figure 1: Data Dictionary

Voting Bootstrap

We first must cover a disclaimer over our use of bootstrap on this data. One of the main bootstrap is that each observation of data has equal chance of occurring. However, this data set is weighted, indicating that each observation has a different chance of occurring. To fix this, under the professor's instructions, we sampled the data the way you would using a bootstrap (meaning each observation is the same) and then applied the weights afterward. This allows us to obtain a more valid estimate using the data, while still conducting bootstrap.

Our initial move was just to do a bootstrap just of the popular vote. We wanted to get a good estimate about the variance of outcomes of the election. Below, we have a histogram with simulated percentage of vote going to Biden on the horizontal axis.

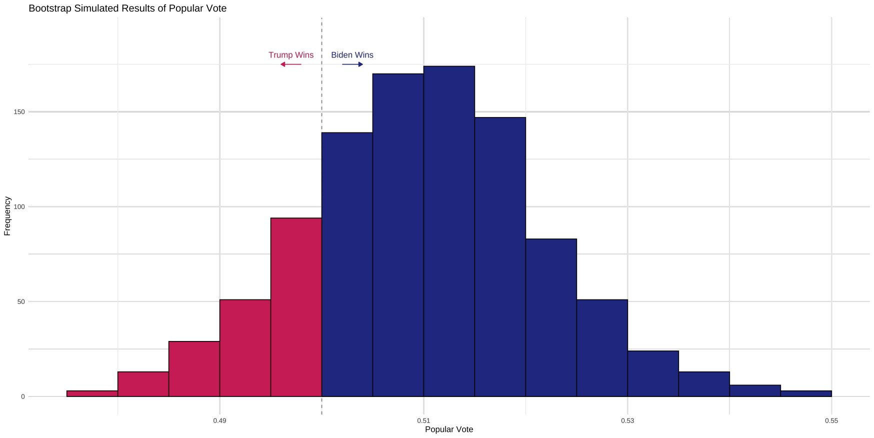


Figure 2: Popular Vote Histogram

As is evident, Biden gets a majority of the votes in most of the bootstrap simulations, around eighty percent of them. However, it is not quite a done deal, as there are still twenty percent of the simulations in which Biden does not get the majority of the votes.

We also conducted a bootstrap of each of the states. We wanted to be able to see whether we could predict the 2020 election using bootstrap. Below, we have a simulated electoral map where blue indicates states in which Biden had over fifty percent of the vote in the majority of the bootstrap simulations. Meanwhile red indicates states where Biden received less than fifty percent of the vote in most of the simulation, implying that Trump is more likely to win.

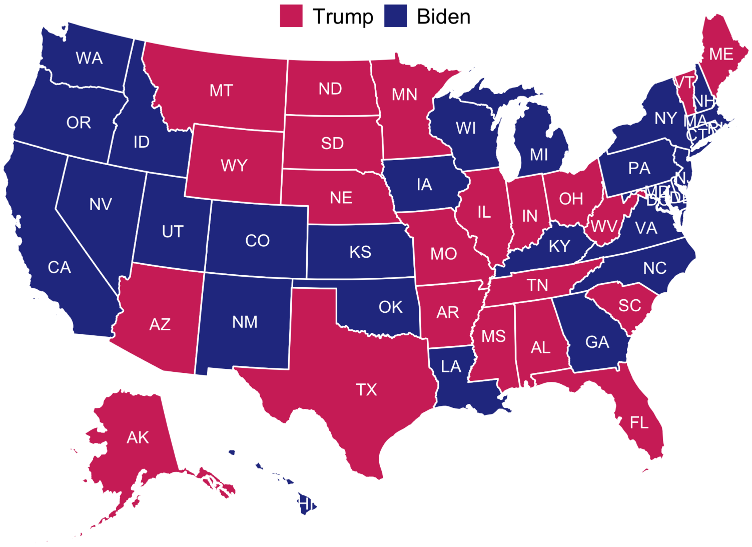


Figure 3: Bootstrap Results By State

This simulation clearly indicates that Biden will win most states. However, this prediction is misguided, as many of the states that Biden wins a majority of the time here he did not in fact win in the actual election. This includes states that are clear Republican strongholds such as Alabama and Idaho, which Biden was not ever going to end up winning in 2020. This is a finding that is evident in real life, as seen in this NYT article, democratic voters are often oversampled in polls which gives a bad representation of the whole country. Thus, we've shown how this is in effect in our analysis.

One final bootstrap we did was a simulation of the popular vote by age. This allows us to get an idea of how each age group will vote, as well as its certainty. Below we have a violin plot, sub grouped by the age bracket of the voter.

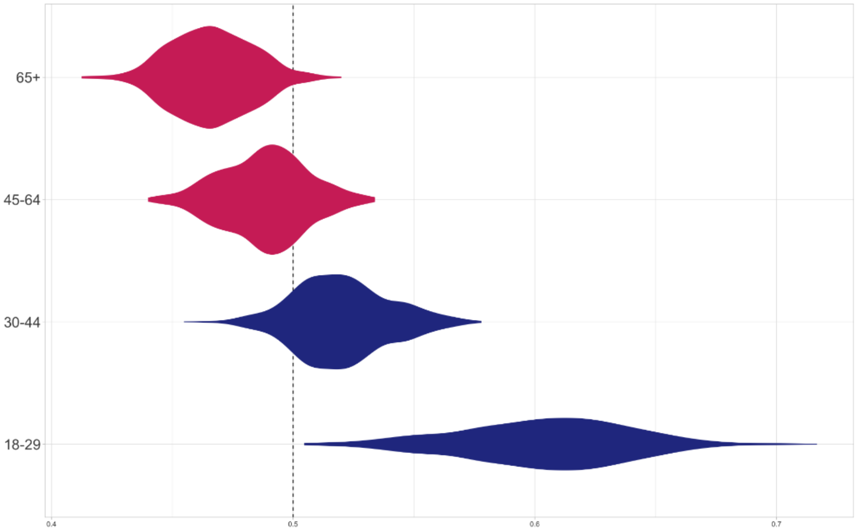


Figure 4: Vote By Age Bracket

We can tell that the older a person gets, the less likely they will vote for Biden. As a result, the vast majority of the simulations show voters over 65 years old giving Biden less than 50 percent of the vote, as well as a majority of voters 45-64 years old doing the same. Meanwhile, Biden gets a majority in most of the 30-44 year old simulations and all of the 18-29 year old ones. However, we can also see that the 18-29 age bracket has by far the most variance. This may be due to how younger voters are less likely to respond to polls, meaning they will be underrepresented and thus are estimates of their voting patterns with be less certain.

Logistic Model

We used logistic regression to predict our binary response: whether a given respondent would vote for Biden or not Biden. We decided to use the following predictors for our initial model: age, sex, education, income, size place, voting method, LGBT, region, racism, abortion, life, climate change, and face mask. The first eight predictors listed are responses to demographic questions and the latter five are responses to opinion questions (see Table 1 for more details).

These predictors were selected because we were interested in whether they would be significant predictors for who a respondent voted for. We excluded obvious variables, such as who they voted for in 2016 and party affiliation, since the main purpose of this model was to see which predictors were significant rather than prediction.

Since all of our predictors were categorical, we used one-hot encoding to relevel the categories. The following levels were set as the reference:

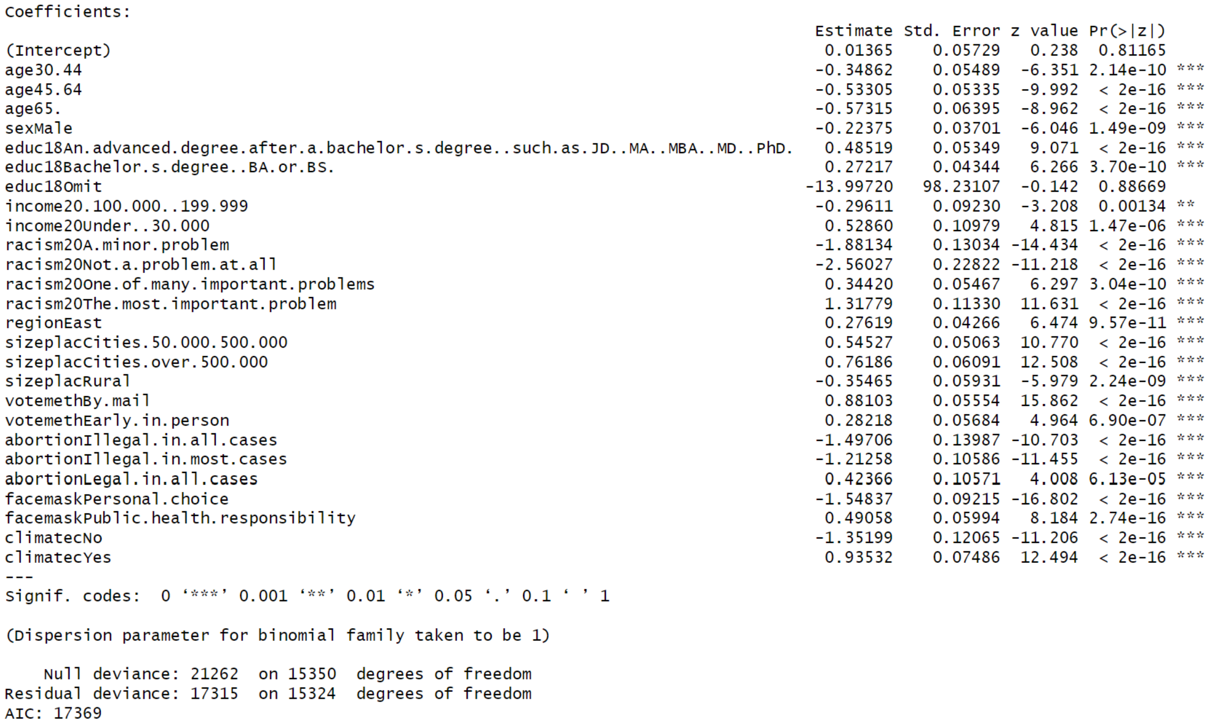
Predictor	Reference level
age	18-29
sex	Female
education	Never attended college
income	Omit
size place	Rural
voting method	Election day
LGBT	Omit
region	South
racism	Omit
abortion	Omit
life	Omit
climate change	Omit

Figure 5: One-hot Encoding Reference Levels

Final Model

After building our initial model, we ran backwards stepwise BIC, which resulted in our final model:

$$\hat{\text{president}} = \beta_0 + \beta_1(\text{age}) + \beta_2(\text{sex}) + \beta_3(\text{education}) + \beta_4(\text{income}) + \beta_5(\text{racism}) +$$
$$\beta_6(\text{size place}) + \beta_7(\text{voting method}) + \beta_8(\text{abortion}) + \beta_9(\text{climate change})$$



```
Coefficients:
                                                                                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)                                                                         0.01365    0.05729   0.238  0.81165    
age30.44                                                                           -0.34862    0.05489  -6.351 2.14e-10 ***
age45.64                                                                           -0.53305    0.05335  -9.992  < 2e-16 ***
age65.                                                                             -0.57315    0.06395  -8.962  < 2e-16 ***
sexMale                                                                            -0.22375    0.03701  -6.046 1.49e-09 ***
educ18An.advanced.degree.after.a.bachelor.s.degree..such.as.JD..MA..MBA..MD..PhD.   0.48519    0.05349   9.071  < 2e-16 ***
educ18Bachelor.s.degree..BA.or.BS.                                                  0.27217    0.04344   6.266 3.70e-10 ***
educ18Omit                                                                        -13.99720   98.23107  -0.142  0.88669    
income20.100.000..199.999                                                          -0.29611    0.09230  -3.208  0.00134 ** 
income20Under..30.000                                                               0.52860    0.10979   4.815 1.47e-06 ***
racism20A.minor.problem                                                            -1.88134    0.13034 -14.434  < 2e-16 ***
racism20Not.a.problem.at.all                                                       -2.56027    0.22822 -11.218  < 2e-16 ***
racism20One.of.many.important.problems                                              0.34420    0.05467   6.297 3.04e-10 ***
racism20The.most.important.problem                                                  1.31779    0.11330  11.631  < 2e-16 ***
regionEast                                                                          0.27619    0.04266   6.474 9.57e-11 ***
sizeplacCities.50.000.500.000                                                       0.54527    0.05063  10.770  < 2e-16 ***
sizeplacCities.over.500.000                                                         0.76186    0.06091  12.508  < 2e-16 ***
sizeplacRural                                                                      -0.35465    0.05931  -5.979 2.24e-09 ***
votemethBy.mail                                                                     0.88103    0.05554  15.862  < 2e-16 ***
votemethEarly.in.person                                                             0.28218    0.05684   4.964 6.90e-07 ***
abortionIllegal.in.all.cases                                                       -1.49706    0.13987 -10.703  < 2e-16 ***
abortionIllegal.in.most.cases                                                      -1.21258    0.10586 -11.455  < 2e-16 ***
abortionLegal.in.all.cases                                                          0.42366    0.10571   4.008 6.13e-05 ***
facemaskPersonal.choice                                                            -1.54837    0.09215 -16.802  < 2e-16 ***
facemaskPublic.health.responsibility                                                0.49058    0.05994   8.184 2.74e-16 ***
climatecNo                                                                         -1.35199    0.12065 -11.206  < 2e-16 ***
climatecYes                                                                         0.93532    0.07486  12.494  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 21262  on 15350  degrees of freedom
Residual deviance: 17315  on 15324  degrees of freedom
AIC: 17369
```

Figure 6: Final Model Summary

Each coefficient estimate indicates the change in log odds for voting Biden in comparison with the reference level. As we can see from the model output, at least one level for each predictor was significant at the

$\alpha = 0.05$ level. The only predictors omitted from our final model were whether a respondent identified as LGBT, their opinion on face masks, and which region of the US they were from. Note that due to the nature of our predictors, there is likely collinearity.

Results

Some notable trends our model predicted were:

1. As education level increased, the more likely they were to vote for Biden.
2. The more a respondent considered racism to be a problem, the more likely they were to vote for Biden.
3. If a respondent voted by mail or early in-person, they were more likely to vote for Biden. The positive coefficient associated with voting by mail may have been due to COVID-19 concerns in 2020. The positive coefficients for both levels accurately describe voting behavior in 2020. Early polls tended to overstate Biden's lead in the election (Keeter et al., 2021).
4. Men were less likely to vote for Biden and people who omitted sex, which likely included non-binary individuals, were more likely to vote for Biden.
5. The more a respondent favored the right to abortion, the more likely they were to vote for Biden.
6. If a respondent did not believe in climate change, they were much less likely to vote for Biden.

Trends for other predictors were less clear, but still significant.

Model Validation

We validated our final model by creating a confusion matrix for our test data.

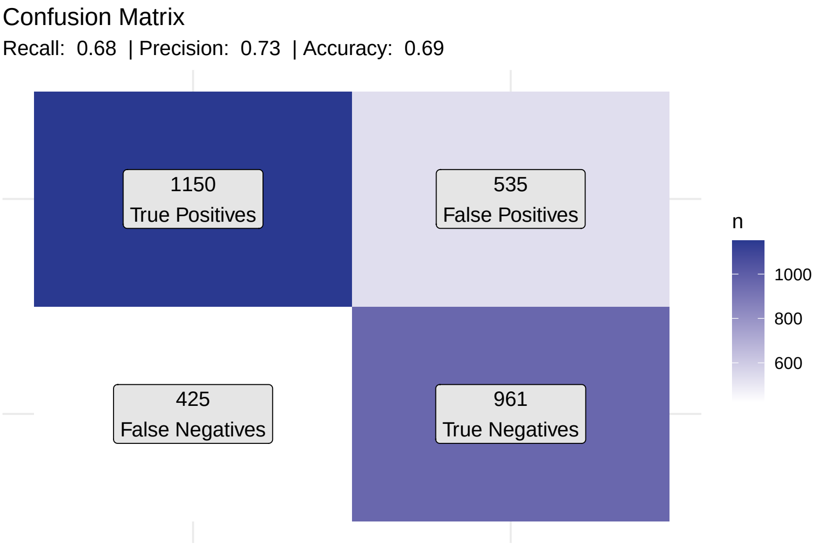


Figure 7: Confusion Matrix

As we can see from the confusion matrix, our model performed relatively well. The recall, precision, and accuracy of our model were 0.68, 0.73, and 0.69, respectively. However, recall that the primary purpose of our model was to determine which predictors were significant, rather than obtaining the highest accuracy.

Bootstrap of Coefficients

After creating our model through the aforementioned model selection process, we wanted to obtain an estimate of the variance of the coefficients. Since bootstrap is a way for us to estimate how much coefficients vary in a model, we decided to conduct a bootstrap on our model. The following chart are violin plots of every coefficient in our model, categorized by question. The colors below indicate whether the median of the coefficient is above or below zero. A blue coefficient means that the median coefficient is above zero, saying that in the majority of simulations responding a question this way means that person is more likely to vote for Biden. Meanwhile, a red coefficient is the reverse.

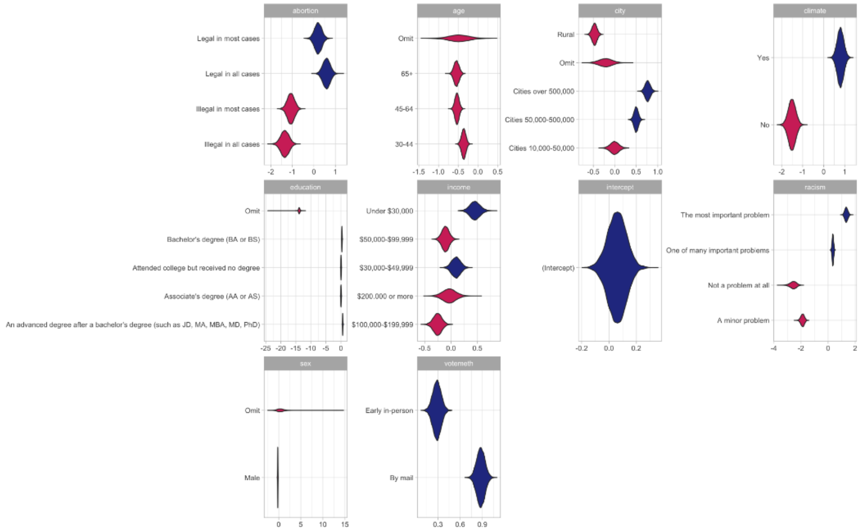


Figure 8: Distribution of Coefficients

A few observations can be made from looking at this chart. One is that the intercept seems to be around zero, with slightly more of the simulations being above, which indicates that our reference levels lead to a voter close to the middle of the political spectrum. However, we can see that some of these coefficients are heavily dependent on the set reference level for that question, such as age and vote method, with all of their coefficients being on only one side of zero.

Other insights from this chart include that the omit response seems to have plenty of variance, especially compared to other responses in the same question. This means that a person omitting answering a question does not tell us much in regards to who they will vote for. We also can see that the climate question has a clear effect, with Yes being completely above zero, and No being completely below zero. Meanwhile, we also can tell that the racism question is similar, with two of the response being clearly above zero and two clearly below zero. Abortion is similar in this respect, but not as extreme, since some of the coefficients appear to cross zero.

Conclusion

In conclusion, this analysis utilized bootstrap techniques to gain insights into the 2020 election, examining the popular vote, state-level predictions, voting patterns by age, and coefficients in a logistic model. The results shed light on important aspects such as oversampling, variable significance, and the predictive capabilities of the models.

When delving into the results, we discovered that certain variables held more significance than others, as indicated by lower p-values. Particularly noteworthy were the variables related to opinions on "Racism" and

“Climate.” These variables consistently exhibited strong significance across all categories, suggesting their importance in determining whether an individual voted for Biden. This finding adds validity to our model, highlighting the significance of these factors in real-world voting behavior.

It is worth mentioning that our logistic regression model was primarily developed for interpretation purposes rather than accurate prediction. However, recognizing the practical importance of accurate predictions, we are interested in exploring the application of undersampling techniques. By rectifying the issue of democratic overrepresentation, we aim to improve the reliability of our model and assess how it impacts the results.

Additionally, we plan to test alternative tree-based models in comparison to our regression model. This exploration will provide insights into their predictive performance and allow us to evaluate their effectiveness in capturing the complexities of voter behavior. By broadening our analysis and incorporating these modeling approaches, we strive to develop a more robust predictive model for future elections.

In summary, this analysis not only uncovers key findings about oversampling, variable significance, and the impact of certain factors on voting behavior but also emphasizes our commitment to enhancing the predictive capabilities of our model. Through undersampling techniques and the exploration of alternative models, we aim to refine our understanding of the data and create a more accurate and reliable model for predicting election outcomes.

References

Keeter, Scott, et al. "What 2020's Election Poll Errors Tell Us About the Accuracy of Issue Polling." Pew Research Center, 1 June 2023, https://www.pewresearch.org/methods/2021/03/02/what-2020s-election-poll-errors-tell-us-about-the-accuracy-of-issue-polling/.

Leonhardt, David. "The Hidden Lessons of the 2016 Election." The New York Times, 19 May 2023, https://www.nytimes.com/2018/03/29/opinion/2016-exit-polls-election.html.

Roper Center for Public Opinion Research. "The New York Times/CBS News Monthly Poll # 1981-10: The GOP and Election '82." Roper Center for Public Opinion Research, 7 May 2023, https://ropercenter.cornell.edu/ipoll/study/31119913.