Final Written Report: What makes someone more likely to vote?

Approximately Normal: Jenny Huang, Bella Larsen, Jessie Bierschenk, Aidan Gildea

17 November 2020

Introduction and Data

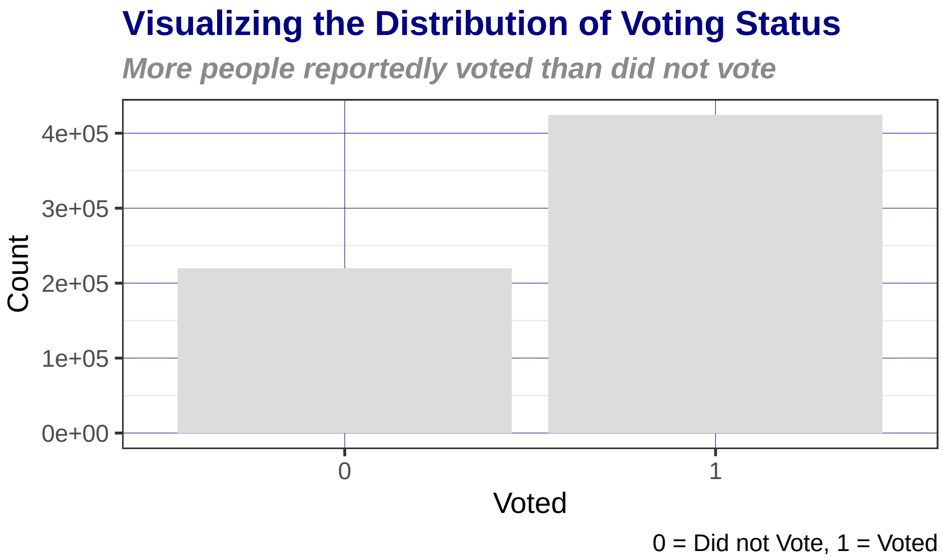
In our regression analysis, we will be assuming the role of a nonprofit political organization aiming to design the most effective GOTV ("get out the vote") effort possible. Our primary focus is to increase voter turnout, however, as an organization funded solely by donations, we need to make sure our outreach efforts are as financially efficient as possible. Therefore, it is critical that we are sending our mailing literature primarily to those we believe may not vote – as this will help turnout voters who may not of voted otherwise. To do so, we will conduct a regression analysis that investigates the different factors and characteristics that appear to be involved or related to voters in the U.S. We are using data on voting behaviors in the U.S. over the past 14 years [1]. These data are sourced from IPUMS (an organization that provides census and survey data) and the American Statistical Association (ASA) [2, 1]. These data include 28 variables on more than 640,000 voters in the U.S. The data collected contains voter characteristics such as age, geographic location, sex, race, marital status, employment, citizenship, ethnicity, education, and voting history and tendencies [1]. This information is particularly relevant in exploring what factors may be related to U.S. - Americans voting or not.

Motivation for our research comes from previous efforts to predict voting behaviors. In an MIT Election data study, researchers describe how understanding voter turnout is important when observing the particular tendencies of certain groups of people as well as factors that motivate individual U.S. citizens to vote [3]. The study highlights general assumptions of voter turnout predictions, noting how higher turnout rates tend to be related to individuals with the following traits: married, white, female, higher education, higher income, older age [3]. The article also addresses how reform may be able to increase voter turnout [3]. In another study done by Harvard graduate student Anthony George Fowler, voter turnout and its implications and repercussions are further examined in the U.S. as well as Australia and Mexico [4]. The study explores the 2008 U.S. election and addresses partisan gaps, voter knowledge (how politically-informed a voter is), and race as main variables of interest in exploring voter tendencies in the U.S. [4]. Both of these studies provide motivation for further and continued investigation into voter data and statistics – especially for our efforts to understand what populations usually do not make it to the voting booth.

In anticipation of our GOTV effort, we are interested in predicting whether a person voted or not based on a list of predictor variables including sex, age, marital status, veteran status, citizenship status (native born or naturalized citizen), whether or not someone is Hispanic or Latinx, employment status and more (described in more detail in Section 2). Our proposed research question is: do voter turnout rates depend on these predictors? Which predictors are more impactful than others? We are also interested in looking at voter turnout over time. We will use the predictor (year) to see if there are changes in voter turnout by demographics over time, or perhaps compare models from different time periods to determine how voter trends have changed over time. Our organization's initial hypothesis is that the significant predictors of voter turnout will include age, level of education, whether they voted in previous elections, and race. Based on historical patterns, people in older age categories tended to vote more than people in younger age categories and people with higher levels of education tended to vote more frequently than people with lower levels of education [5]. Finally, if a person has voted previously, we predict they will be more likely to vote again compared to someone who has not voted previously. Taking these hypotheses into consideration, we will explore which factors are most significant in relation to voting attendance. Our findings will hopefully allow us to identify populations that are statistically less likely to vote, informing who we target with our GOTV literature in the future.

While our GOTV initiative hopes to turnout people who we determine are unlikely to vote, our statistical analysis will use a broad margin to determine who should receive our informational materials. Every vote matters, so we want to make sure no one gets left behind!
We will begin our EDA by visualizing the relationship between the response variable describing whether or not someone voted and several of the other variables of particular interest.

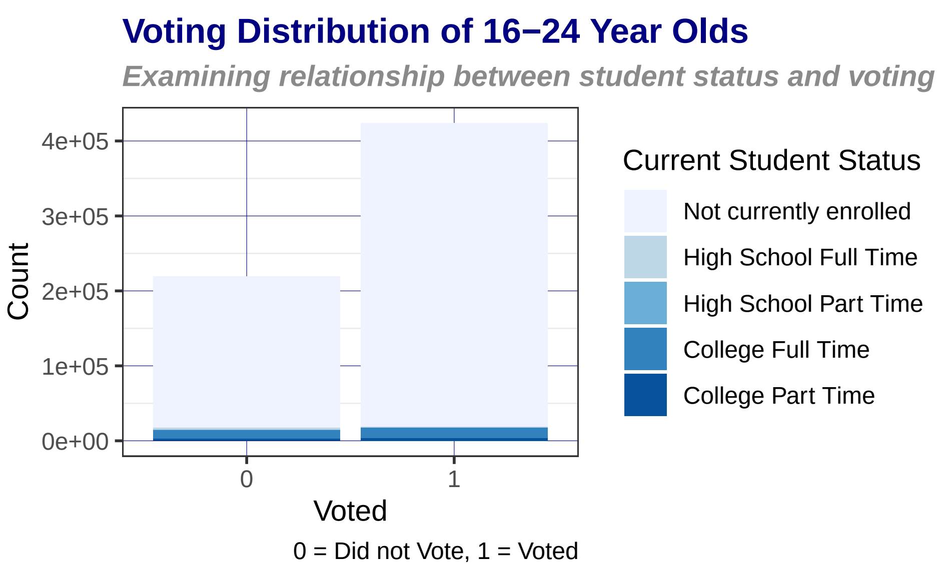
First, we will simply look at the distribution of those who voted throughout the last 8 years of elections.



see theme code inspiration at reference [6] see scale fill code inspiration at reference [7]

From the barplot above, it is clear the more individuals in the data set voted (voted = 1) than did not (voted = 0).

Many political nonprofits engage with college campuses, so we want to analyze whether or not being a student influences the frequency of voting. We will explore this preliminarily by visualizing the distribution of if school aged individuals (18-24) voted or not – categorized by their current student level. This is seen in the bar plot below.



see scale fill brewer code inspiration from reference [6]

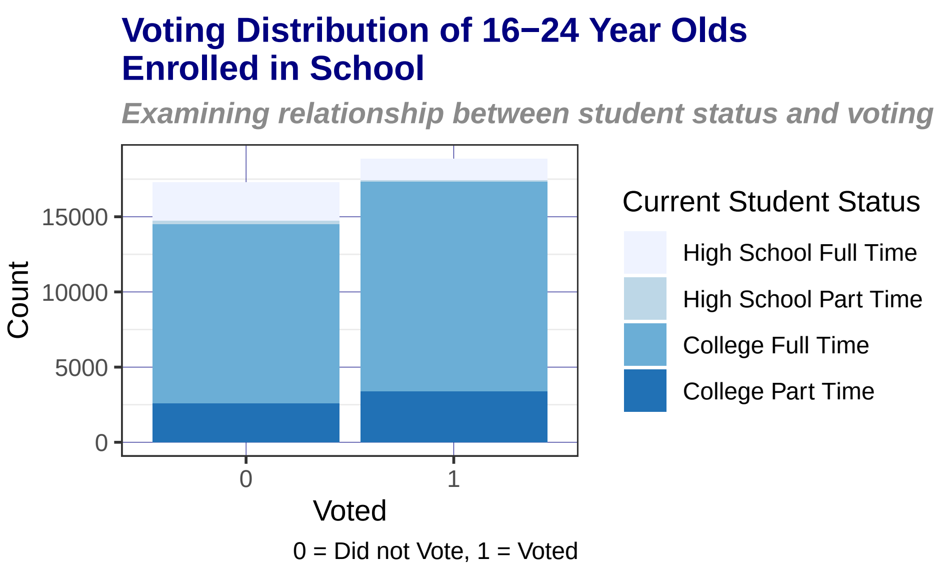
From the bar plot, it is evident that a majority of these individuals were not currently enrolled. This may be a result of a general national trend, but we want to investigate if it is the result of a larger proportion of older individuals within in the range of ages between 16-24. We will investigate this by analyzing those who are not currently enrolled in school within this age range.

Table 1: Proportion of Respondents Not Currently Enrolled In School By Age

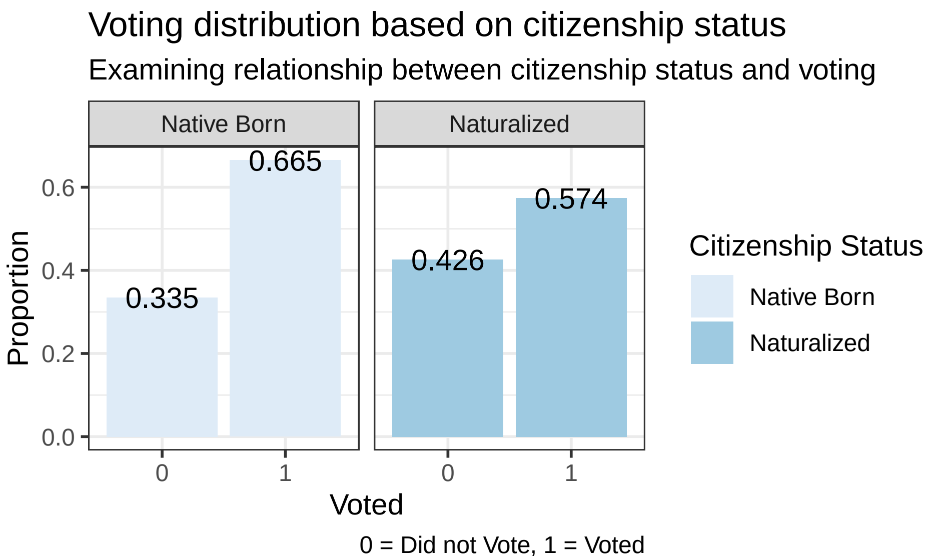
AGE	n	prop
18	2327	0.065
19	3671	0.102
20	4367	0.122
21	4760	0.133
22	5955	0.166
23	6991	0.195
24	7791	0.217

From the table above, it is apparent that more than 40% of those not currently enrolled in school are 23-24 years old. This could be a potential reason for why this age range includes so many who are not currently enrolled as a student.

To more meaningfully analyze the relationship between being a student and if they vote or not, we adjusted our visualization to only include those currently enrolled in some level of education. This is seen in the visualization below.

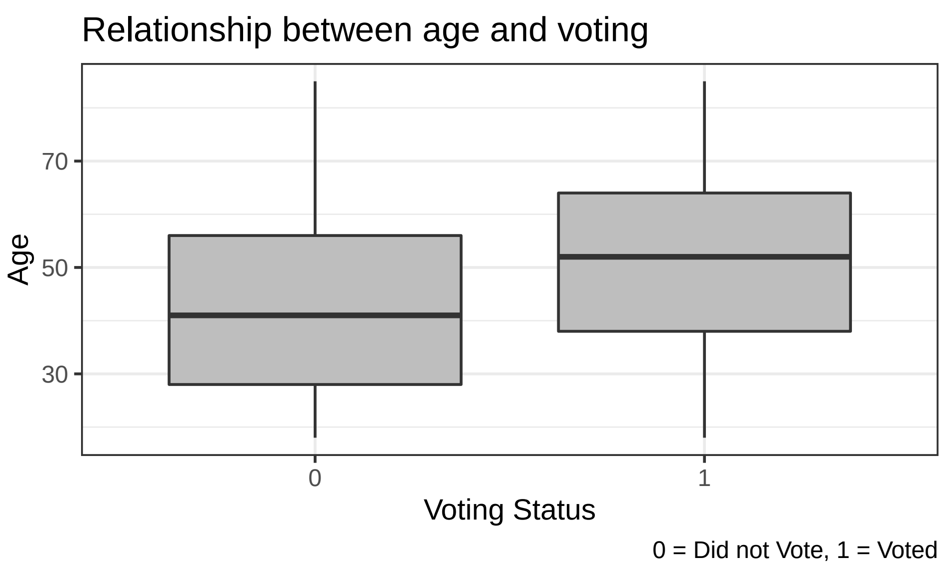


This visualization shows that eligible voters between the age 16-24 enrolled in some education at the time were mainly full time college students. More full time and part time college students that were eligible to vote did vote compared to those that did not vote. The opposite is true for high school students: more full time and part time high school students that were eligible to vote did not vote compared to those that did vote.



see geom_text() code inspiration in reference [8]

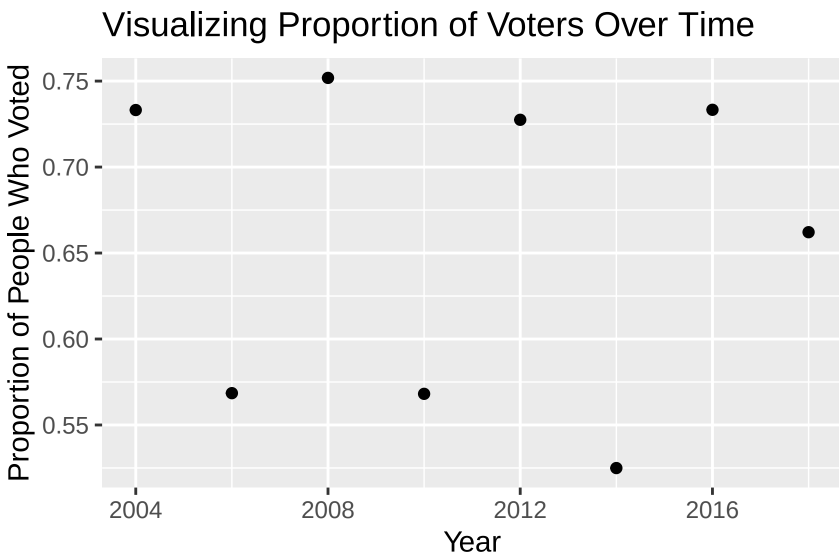
This visualization shows that for both native born and naturalized individuals, more citizens that were eligible to vote did vote compared to those that did not vote; however, the proportion is much greater for native born citizens than for naturalized. For a similar plot comparing respondents by veteran status, please consult Appendix A.



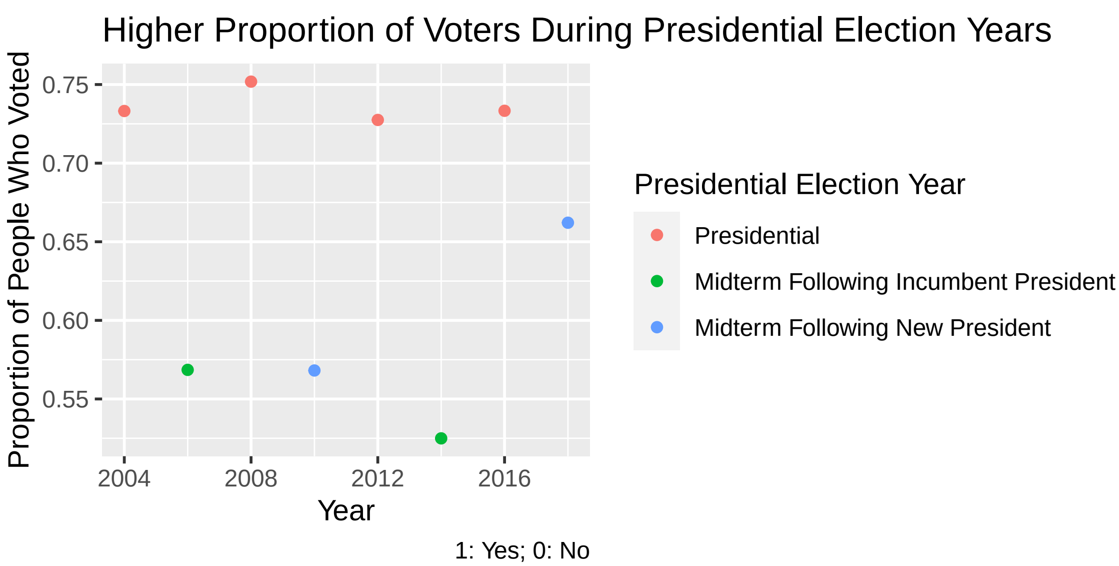
This box plot shows that eligible voters that did vote were generally older than respondents that did not vote.

We are also interested in looking at how voter turnout has changed over the years.

We notice that the proportion of people who voted fluctuates depending on whether the year falls on a presidential election. In the trend of the proportion of voting over time, we see a clear divide between the years when there is a presidential elections versus when there is not. In the future, we may decide to add the variable “Election Year” as an interaction term with year as a divide between years that fall on an election.



We decided to look at the scatterplot of voter turnout over time broken down by the status of the election (Presidential election, midterm after incumbent president, or midterm after new president) to see how it may differ depending on the election.



From the above scatterplot, we confirmed that there is higher voter turnout during presidential elections compared to midterm elections. In addition, there is equal or higher voter turnout for midterm elections following the election of a new president compared to midterm elections following the election of an incumbent president, and an especially high voter turnout in the midterm election after Trump's election in 2016.

Finally, it is important to acknowledge that our any missingness in our data, as it may influence the outcome of our regression analysis. The initial data sourced from the This Is Statistics: Fall Data Challenge used non-uniform values for the different variables, therefore we cleaned the data to be more intelligible for our analysis. While cleaning, we did impute "NA" values by either combining them with other categories and/or removing them to improve our analysis. We believe this will not negatively affect out analysis, yet will move forward taking it into consideration.

After analyzing these preliminary visualizations and observations, we will now begin to build our model.

Methodology

Model Selection

Select a random subset of the data to create model.

In order to make our model, we have decided to take a random sample of 10,000 to be sure that the model selection is accurate and not obscured by too many observations.

Using the random sample of 10,000 observations, we began the model selection process. See Appendix B for the full backward selection model output.

Table 2: Model Resulting From Backward Selection

term	estimate	std.error	statistic	p.value
(Intercept)	0.079	0.130	0.607	0.544
sexMale	-0.076	0.048	-1.576	0.115
marstDivorced/Separated	-0.703	0.071	-9.951	0.000
marstNot Married/Other	-0.475	0.057	-8.367	0.000
citizenNaturalized	-0.684	0.103	-6.644	0.000
employedYes	0.405	0.060	6.754	0.000
highest_educationHigh School Degree/GED	-1.365	0.074	-18.550	0.000
highest_educationSome College	-0.636	0.081	-7.833	0.000
highest_educationSome High School	-2.224	0.099	-22.490	0.000
highest_educationAssociate Degree	-0.532	0.095	-5.587	0.000
highest_educationMasters Degree	0.281	0.125	2.253	0.024
highest_educationProfessional Degree	0.475	0.281	1.689	0.091
highest_educationDoctorate Degree	0.395	0.274	1.440	0.150
highest_educationNone/Unknown	-3.151	0.654	-4.820	0.000
current_studentHigh School Full Time	1.195	0.305	3.921	0.000
current_studentHigh School Part Time	0.284	1.139	0.249	0.803
current_studentCollege Full Time	0.443	0.122	3.617	0.000
current_studentCollege Part Time	0.538	0.247	2.182	0.029
raceBlack	0.606	0.084	7.226	0.000
raceAsian or Pacific Islander	-0.375	0.133	-2.819	0.005
raceNative American	-0.057	0.201	-0.283	0.777
race2 or more races	0.268	0.205	1.305	0.192
AGE	0.040	0.002	22.657	0.000
ElectionMidterm	-0.851	0.048	-17.720	0.000

The final model included predictor variables for sex, whether someone between ages 16 and 24 is currently enrolled in school, citizenship status, employment status, race, marital status, whether it was a presidential election year or a midterm election year, respondent age, and the highest education level of the respondent.

The backward selection based on AIC removed the variables for the year, whether someone is from a metro area, veteran status, and hispanic status (see Appendix B).

Interpretation of coefficients of interest:

Baseline: Female, married, native born, not employed, Bachelors Degree is highest education, not currently enrolled in school if between the age of 16-24, White, and the time of voting is midterm following the election of an incumbent president.

We expect the odds of a survey respondent voting for a divorced/separated eligible voter to be 0.50 (exp(-0.702)) times the odds of individuals that are female, married, native born, not employed, have Bachelors Degree as their highest education, not currently enrolled in school if between the age of 16-24, White, and the election is a midterm election following the election of an incumbent president, after factoring in all other voter characteristics/information.

For each additional year in age, we expect the odds of a survey respondent voting to be 1.04 (exp(0.040)) times the odds of individuals that are female, married, native born, not employed, have Bachelors Degree as their highest education, not currently enrolled in school if between the age of 16-24, White, and the election is a midterm election following the election of an incumbent president, after factoring in all other voter characteristics/information.

We expect the odds of a survey respondent voting when it is a presidential election year to be 2.689 (exp(0.989)) times the odds of individuals that are female, married, native born, not employed, have Bachelors Degree as their highest education, not currently enrolled in school if between the age of 16-24, White, and the election is a midterm election following the election of an incumbent president, after factoring in all other voter characteristics/information.

Checking Model Conditions

Linearity

According to the empirical logit plot (see Appendix C), there is a linear relationship between the empirical logit and the predictor variable age. Hence linearity is satisfied for AGE.

Randomness

It is possible that randomness is not satisfied because our data is from the census survey, which may not be random (ie might select for people who have time to fill it out). However, there is no reason to believe that this will not generalize to the US population as a whole in a significant way, particularly due to the large sample size.

Independence

Independence may be violated because geographic location may influence voting due to factors such as (residuals by state ID). Hence, we will look at misclassification rate by region.

Here, we create a confusion matrix with a decision threshold of 0.5.

We will join the augmented data with region in order to look at misclassification rates by region.

Below is a table of the misclassification rates by region.

Table 3: Missclassification Rates by Region

region	voted	predicted_voted	n	prop
Midwest	0	Will Vote	81790	0.192
Midwest	1	Will Not Vote	32378	0.076
Northeast	0	Will Vote	117077	0.195
Northeast	1	Will Not Vote	45035	0.075
South	0	Will Vote	131989	0.192
South	1	Will Not Vote	51318	0.075
West	0	Will Vote	93919	0.193
West	1	Will Not Vote	37085	0.076
NA	0	Will Vote	8042	0.186
NA	1	Will Not Vote	3424	0.079

Consulted Census data for the region fips number corresponding to region name [9]

We then created a plot of misclassification rate by region to determine if independence is satisfied. See the Appendix C for the boxplot visualizing the misclassification rates by region.

Based on the misclassification rates, we have no reason to believe that the independence condition is not satisfied. The misclassification rates across regions are relatively similar, which suggests that there is not an issue of spatial correlation and that the independence condition is satisfied.

Checking for influential points

Cook's distance

We will also look for influential points using Cook's Distance. See the plot for Cook's distance in Appendix C.

According to the plot of Cook's Distance (see Appendix C), there are no influential points, so all points can be left in the model.

Multicollinearity

All of the VIF values are under the threshold of 10 (see table in Appendix C), indicating that there is no evidence of multicollinearity in our data.

Interaction Terms

We will add in several interaction terms of interest to us and use a drop-in-deviance test to see if they are meaningful predictors of the odds of someone voting.

The reduced model is the same as the above model titled "Model Resulting From Backward Selection," and the full model is the reduced model plus interactions terms for sex and employment, election type and highest education level, sex and election type, and race and election type.

The following hypotheses will be used:

H_0: the coefficients for the interaction between sex and employment, election type and highest education level, sex and election type, and race and election type are all zero

H_0 :

$\beta_{sex*employed} = \beta_{Election*highest_education} = \beta_{sex*Election} = \beta_{race*Election} = 0$

H_a: at least one of these coefficients for the interaction terms $\neq$ zero

$\alpha = 0.05$

Table 4: Drop-In-Deviance Test Results For Interaction Terms

Resid..Df	Resid..Dev	df	Deviance	p.value
9976	10695.20	NA	NA	NA
9963	10645.55	13	49.649	0

The p-value (7.840e-07) is very small (less than the alpha level 0.05), so we can reject the null hypothesis. Thus, we conclude that the data provide sufficient evidence that the coefficients associated with the additional interaction terms are not equal to 0. Therefore, we should add them to the model.

Significant interaction terms: The effect of the election being a presidential election for an individual with the highest level of education as a High School Degree/GED is significant with a p-value of 0.011 assuming an alpha level of 0.05. The coefficient is negative, indicating that this effect would mean that during a presidential election for an individual with the highest level of education as a High School Degree/GED, they are less likely to vote than individuals that are female, married, native born, not employed, have Bachelors Degree as their highest education, not currently enrolled in school if between the age of 16-24, White, and the election is a midterm election following the election of an incumbent president.

The effect of the election being a presidential election for an individual that is Black is significant with a p-value of 0.013 assuming an alpha level of 0.05. The coefficient is positive, indicating that this effect would mean that during a presidential election for an individual that is Black, they are more likely to vote than individuals that are female, married, native born, not employed, have Bachelors Degree as their highest education, not currently enrolled in school if between the age of 16-24, White, and the election is a midterm election following the election of an incumbent president

We will use a drop-in-deviance test to determine whether or not sex is a meaningful predictor of the odds of someone voting.

The following hypotheses will be used:

H_0: the coefficients for the main effect for sex, the interaction between sex and employment, and sex and election type are all zero. All of the coefficients associated with sex are equal to zero.

$H_0 : \beta_{sex} = \beta_{sex*employed} = \beta_{sex*Presidential-Election-Status} = 0$ H_a: at least one of these coefficients for the coefficients associated with sex $\neq$ zero

$\alpha = 0.05$

Table 5: Drop-In-Deviance Test Results For Sex

Resid..Df	Resid..Dev	df	Deviance	p.value
9965	10649.50	NA	NA	NA
9963	10645.55	2	3.954	0.138

From the drop-in-deviance test to include variable sex, the p-value (0.138) is greater than an alpha-level of 0.10, so we will exclude the main effect for sex and the interaction terms including sex from the model. This contradicts the model output from the backward selection, which is most likely due their different criteria for statistical significance. Backward selection makes decisions based on the AIC, while the drop-in-deviance test uses the p-value. A potential discrepancy between the two values may have resulted in the different determinations of the significance of sex. We have chosen to use the results of the drop-in-deviance test, and therefore, will be excluding sex from the final model.

Results

Table 6: Final Model Output

term	estimate	std.error	statistic	p.value
(Intercept)	0.241	0.150	1.605	0.109
current_studentHigh School Full Time	1.155	0.302	3.819	0.000
current_studentHigh School Part Time	0.354	1.137	0.311	0.756
current_studentCollege Full Time	0.455	0.123	3.709	0.000
current_studentCollege Part Time	0.569	0.246	2.314	0.021
citizenNaturalized	-0.681	0.103	-6.599	0.000
employedYes	0.395	0.060	6.629	0.000
raceBlack	0.897	0.134	6.717	0.000
raceAsian or Pacific Islander	-0.743	0.191	-3.882	0.000
raceNative American	-0.312	0.287	-1.090	0.276
race2 or more races	-0.133	0.279	-0.477	0.633
marstDivorced/Separated	-0.710	0.071	-10.031	0.000
marstNot Married/Other	-0.477	0.057	-8.396	0.000
ElectionMidterm	-1.152	0.128	-8.965	0.000
AGE	0.040	0.002	22.696	0.000
highest_educationHigh School Degree/GED	-1.647	0.118	-13.981	0.000
highest_educationSome College	-0.834	0.130	-6.432	0.000

term	estimate	std.error	statistic	p.value
highest_educationSome High School	-2.552	0.144	-17.759	0.000
highest_educationAssociate Degree	-0.621	0.154	-4.029	0.000
highest_educationMasters Degree	0.619	0.237	2.608	0.009
highest_educationProfessional Degree	1.204	0.615	1.959	0.050
highest_educationDoctorate Degree	0.050	0.460	0.110	0.913
highest_educationNone/Unknown	-3.452	0.811	-4.259	0.000
ElectionMidterm:highest_educationHigh School Degree/GED	0.475	0.150	3.163	0.002
ElectionMidterm:highest_educationSome College	0.314	0.164	1.908	0.056
ElectionMidterm:highest_educationSome High School	0.590	0.191	3.082	0.002
ElectionMidterm:highest_educationAssociate Degree	0.139	0.197	0.708	0.479
ElectionMidterm:highest_educationMasters Degree	-0.486	0.281	-1.730	0.084
ElectionMidterm:highest_educationProfessional Degree	-1.002	0.697	-1.437	0.151
ElectionMidterm:highest_educationDoctorate Degree	0.538	0.572	0.941	0.347
ElectionMidterm:highest_educationNone/Unknown	0.824	1.360	0.606	0.544
raceBlack:ElectionMidterm	-0.494	0.171	-2.893	0.004
raceAsian or Pacific Islander:ElectionMidterm	0.630	0.247	2.551	0.011
raceNative American:ElectionMidterm	0.489	0.399	1.225	0.221
race2 or more races:ElectionMidterm	0.836	0.406	2.056	0.040

Checking Model Conditions for Final Model

Checking for influential points

Cook's distance

We will also look for influential points in our final model using Cook's Distance.

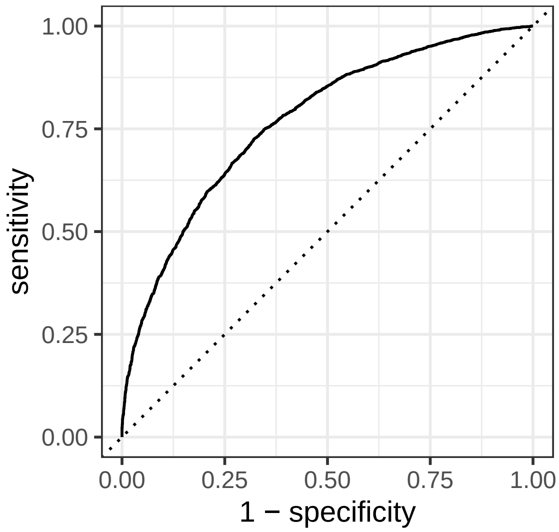
According to the Cook's Distance plot for the final model (see Appendix D), there are no influential points, so all points can be left in the final model.

Multicollinearity

All of the VIF values are under the threshold of 10 (see table in Appendix D), indicating that there is no evidence of multicollinearity in our data.

Creating Classifier

We'll fit an ROC curve to help us determine a decision-making threshold.



Below we look at values within a range of thresholds in order to choose the threshold that yields high sensitivity and low values of the false positive rate (1 - specificity).

Table 7: ROC Curve Threshold Table

.threshold	specificity	sensitivity	pred_prob	false_pos_rate
0.715	0.786	0.602	0.672	0.214
0.716	0.786	0.602	0.672	0.214
0.716	0.787	0.602	0.672	0.213
0.716	0.787	0.602	0.672	0.213
0.716	0.787	0.602	0.672	0.213
0.716	0.787	0.602	0.672	0.213
0.716	0.787	0.602	0.672	0.213

According to our ROC curve and modeling objectives, we will choose a threshold of 0.716 because we want to prioritize having a lower false positive rate (and avoid making a type I error) than having higher true positive rate, or sensitivity, as it does not hurt to mail a few extra mailers to people who may already be planning to vote. Because our decision will be to not send a mailer if the predicted probability is greater than the threshold, we want to be sure that we are not inaccurately predicting someone to vote when they truly will not, which means that we want to minimize our false positive rate.

At this threshold, we have a sensitivity of about 0.602, which means that our model prediction will correctly identify about 60% of people who will vote. In other words, using this threshold, of the people who will actually vote, we will capture about 60.2% of them. Our false positive rate is 0.213, which means that with our current threshold, we are predicting "yes" for about 21.3% of people who will actually not vote.

Any data point with a probability over 0.716 will be predicted to be in the "voted" category and will therefore not be sent a mailer.

Discussion

While completing this project provided excellent insight into the voting patterns in the United States, there are several limitations of this project that must be acknowledged. First, the original dataset hard-coded several variables with "unknown" and "missing" combined into one category. This could have obscured the effects of missingness in our data and contributed to decreasing the predictive accuracy of our model and results. For this reason, if given the opportunity to expand this project, we would like to explore missingness in our dataset and its implications with more depth to improve our model. Because several categorical variables already had levels that combined missing and unknown, we decided to code the other variables in the same way, effectively imputing a new category for missing that was used in the model. While this was the best way to deal with missingness within the scope of this project, we would be interested in expanding this in the future and teasing out missing from unknown to create a more accurate model and improve our results.

References

[1]“Fall Data Challenge | This Is Statistics.” n.d. Accessed November 16, 2020. https://thisisstatistics.org/fal ldatachallenge/.

[2] University of Minnesota. “What Is IPUMS?” Text. IPUMS. February 7, 2019. https://ipums.org/what-is-ipums.

[3] MIT Election Data + Science Lab. n.d. “Voter Turnout.” Accessed November 16, 2020. https://plotly.com/~cwimpy/69/.

[4] Fowler, Anthony George. 2013. “Five Studies on the Causes and Consequences of Voter Turnout,” October. https://dash.harvard.edu/handle/1/11156810.

[5] McDonald, Michael. n.d. “Voter Turnout Demographics - United States Elections Project.” Accessed November 16, 2020. http://www.electproject.org/home/voter-turnout/demographics.

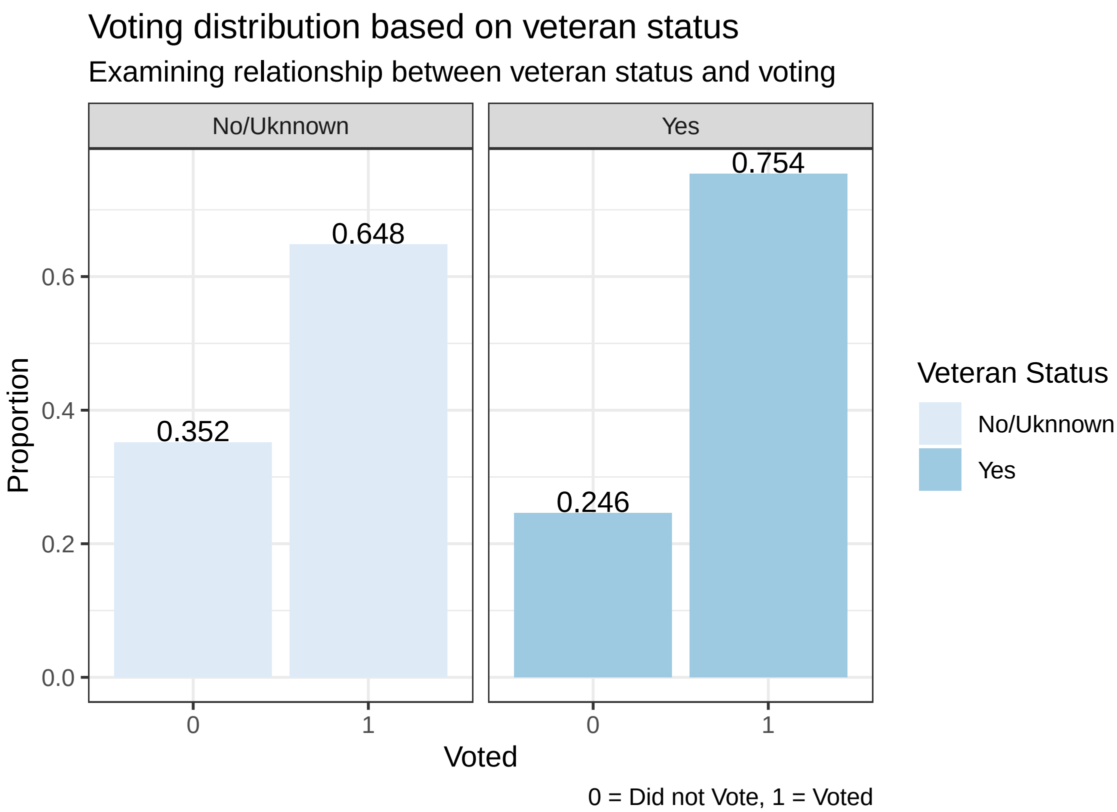
[6] “GGPlot Cheat Sheet for Great Customization - Articles - STHDA.” http://www.sthda.com/english/arti cles/32-r-graphics-essentials/125-ggplot-cheat-sheet-for-great-customization/ (November 15, 2020).

[7] “The A - Z Of Rcolorbrewer Palette You Must Know.” 2018. Datanovia. https://www.datanovia.com/en /blog/the-a-z-of-rcolorbrewer-palette/ (November 15, 2020).

[8] “How to Put Labels over Geom_bar for Each Bar in R with Ggplot2.” Stack Overflow. https://stackoverf low.com/questions/12018499/how-to-put-labels-over-geom-bar-for-each-bar-in-r-with-ggplot2 (November 15, 2020).

[9] Prepared by the Geography Division of the U.S. Census Bureau. “Census Regions and Divisions of the United States.” https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf.

Appendix A



see geom_text() code inspiration in reference [4]

Appendix B

Backward Selection Output

```
## Start:  AIC=10747.79
## voted ~ metro + sex + marst + veteran + citizen + hispanic_status +
##     employed + highest_education + current_student + race + AGE +
##     Election
##
##                     Df Deviance   AIC
## - metro              1    10694 10746
## - veteran            1    10694 10746
## - hispanic_status    1    10695 10747
## <none>                    10694 10748
## - sex                1    10696 10748
## - current_student    4    10722 10768
## - citizen            1    10731 10783
## - employed           1    10740 10792
## - race               4    10759 10805
## - marst              2    10824 10874
## - Election           1    11018 11070
## - AGE                1    11218 11270
## - highest_education  8    11577 11615
##
```

```
## Step:  AIC=10745.79
## voted ~ sex + marst + veteran + citizen + hispanic_status + employed +
##     highest_education + current_student + race + AGE + Election
##
##                     Df Deviance   AIC
## - veteran            1    10694 10744
## - hispanic_status    1    10695 10745
## <none>                    10694 10746
## - sex                1    10696 10746
## - current_student    4    10722 10766
## - citizen            1    10731 10781
## - employed           1    10740 10790
## - race               4    10760 10804
## - marst              2    10825 10873
## - Election           1    11018 11068
## - AGE                1    11218 11268
## - highest_education  8    11584 11620
##
## Step:  AIC=10743.83
## voted ~ sex + marst + citizen + hispanic_status + employed +
##     highest_education + current_student + race + AGE + Election
##
##                     Df Deviance   AIC
## - hispanic_status    1    10695 10743
## <none>                    10694 10744
## - sex                1    10696 10744
## - current_student    4    10722 10764
## - citizen            1    10731 10779
## - employed           1    10740 10788
## - race               4    10760 10802
## - marst              2    10826 10872
## - Election           1    11018 11066
## - AGE                1    11238 11286
## - highest_education  8    11585 11619
##
## Step:  AIC=10743.2
## voted ~ sex + marst + citizen + employed + highest_education +
##     current_student + race + AGE + Election
##
##                     Df Deviance   AIC
## <none>                    10695 10743
## - sex                1    10698 10744
## - current_student    4    10723 10763
## - citizen            1    10739 10785
## - employed           1    10741 10787
## - race               4    10762 10802
## - marst              2    10828 10872
## - Election           1    11020 11066
## - AGE                1    11253 11299
## - highest_education  8    11609 11641
```

Appendix C

Checking Model Conditions for the resulting model from the selection process

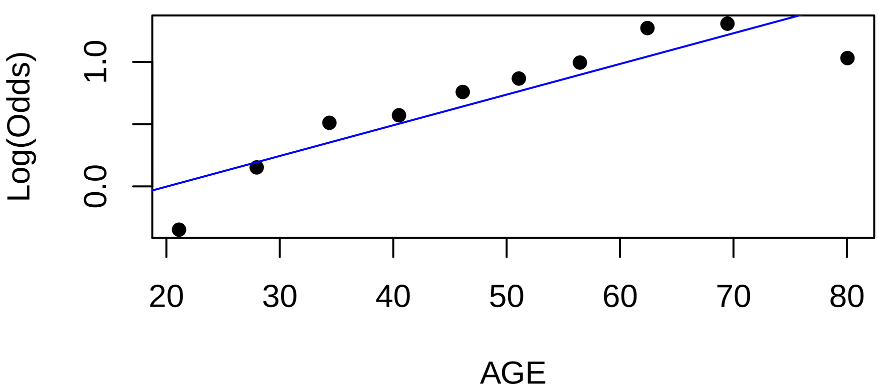
Linearity

Table 8: Empirical Logit Results For Sex

sex	voted	n	prop	emp_logit
Female	1	3582	0.67	0.710
Male	1	3029	0.65	0.621

Table 9: Empirical Logit Results For Marital Status

marst	voted	n	prop	emp_logit
Married	1	4145	0.735	1.018
Divorced/Separated	1	783	0.588	0.355
Not Married/Other	1	1683	0.556	0.226



Independence

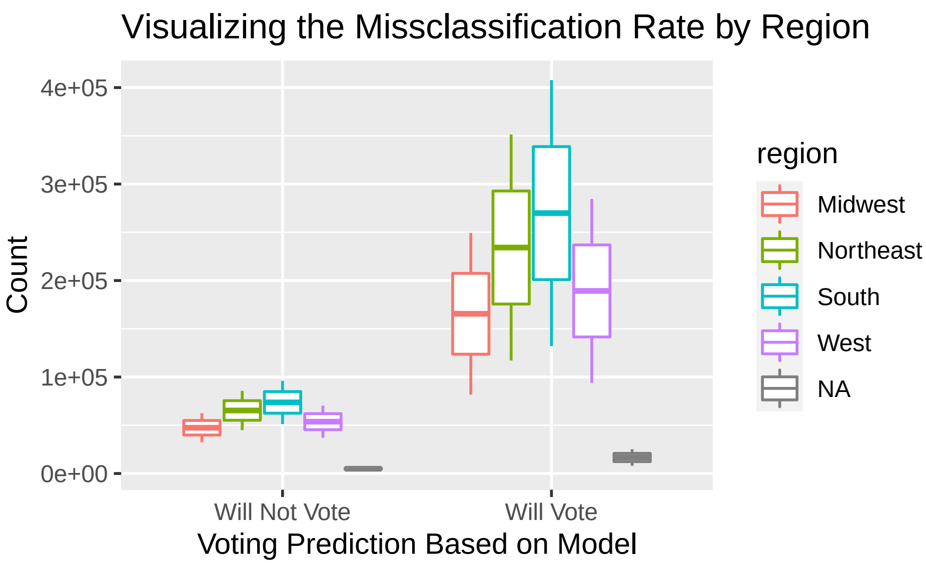
Misclassification rate by region

Table 10: Missclassification Rates by Region

region	voted	predicted_voted	n	prop
Midwest	0	Will Vote	81790	0.192
Midwest	1	Will Not Vote	32378	0.076
Northeast	0	Will Vote	117077	0.195
Northeast	1	Will Not Vote	45035	0.075
South	0	Will Vote	131989	0.192
South	1	Will Not Vote	51318	0.075
West	0	Will Vote	93919	0.193
West	1	Will Not Vote	37085	0.076
NA	0	Will Vote	8042	0.186
NA	1	Will Not Vote	3424	0.079

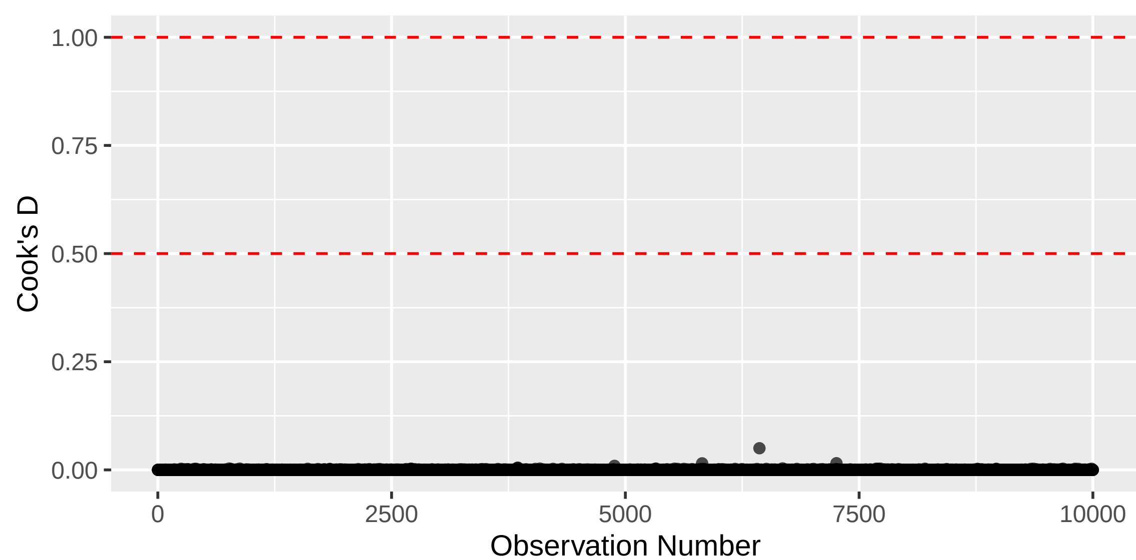
Consulted Census data for the region fips number corresponding to region name [3]

Boxplot of misclassification rate by region



Checking for influential points

Cook's distance plot



Multicollinearity: VIF Table

names	x
current_studentHigh School Part Time	1.005
raceNative American	1.008
race2 or more races	1.009
highest_educationNone/Unknown	1.015
sexMale	1.022
ElectionMidterm	1.026
current_studentCollege Part Time	1.036

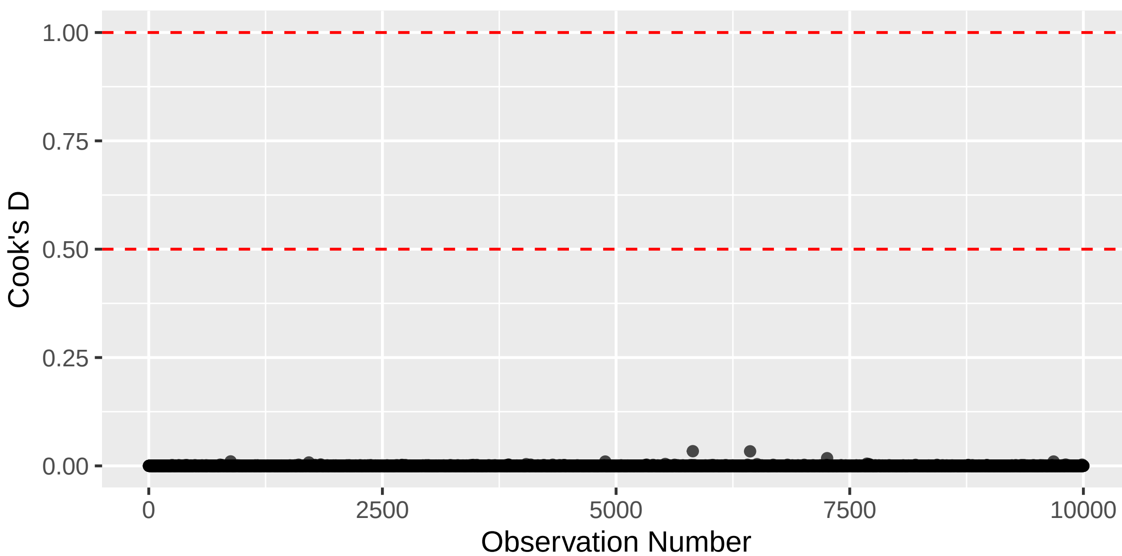
names	x
highest_educationProfessional Degree	1.043
raceBlack	1.048
highest_educationDoctorate Degree	1.049
current_studentHigh School Full Time	1.086
marstDivorced/Separated	1.116
raceAsian or Pacific Islander	1.204
citizenNaturalized	1.213
highest_educationMasters Degree	1.246
marstNot Married/Other	1.286
current_studentCollege Full Time	1.301
employedYes	1.432
highest_educationAssociate Degree	1.509
AGE	1.705
highest_educationSome High School	1.741
highest_educationSome College	1.951
highest_educationHigh School Degree/GED	2.218

Appendix D

Checking Model Conditions for the Final Model

Checking for influential points

Cook's distance plot for final model



Multicollinearity: VIF Table for final model

names	x
current_studentHigh School Part Time	1.005
current_studentCollege Part Time	1.039
current_studentHigh School Full Time	1.090
marstDivorced/Separated	1.116
citizenNaturalized	1.212
marstNot Married/Other	1.283

names	x
current_studentCollege Full Time	1.304
employedYes	1.420
ElectionMidterm:highest_educationNone/Unknown	1.556
highest_educationNone/Unknown	1.572
AGE	1.705
race2 or more races	1.910
race2 or more races:ElectionMidterm	1.915
raceNative American	2.077
raceNative American:ElectionMidterm	2.088
raceAsian or Pacific Islander:ElectionMidterm	2.336
raceAsian or Pacific Islander	2.501
raceBlack	2.607
raceBlack:ElectionMidterm	2.690
ElectionMidterm:highest_educationDoctorate Degree	2.951
highest_educationDoctorate Degree	2.952
ElectionMidterm:highest_educationSome High School	3.147
highest_educationSome High School	3.708
highest_educationAssociate Degree	3.915
ElectionMidterm:highest_educationAssociate Degree	3.921
highest_educationMasters Degree	4.321
ElectionMidterm:highest_educationMasters Degree	4.472
highest_educationProfessional Degree	4.715
ElectionMidterm:highest_educationProfessional Degree	4.744
ElectionMidterm:highest_educationSome College	4.923
highest_educationSome College	4.966
highest_educationHigh School Degree/GED	5.711
ElectionMidterm:highest_educationHigh School Degree/GED	5.933
ElectionMidterm	7.295