Influence of Demographic Factors on Presidential Election Voting Patterns

MGT 6203 Progress Report - Group 18

Ben Fan, Chi Zhang, Diane Kuai, Jessica Zhang, Xiaofei Chen

Introduction

Presidential elections are a frequently discussed topic in the United States. One of the most intriguing questions behind the election is why people cast their ballots the way they did or what the implications of the outcome are. According to the Office of Legislative Research, factors influencing voters' behavior include demographics, candidate influence, and current political concerns [5]. Based on the data published by the United States Census Bureau, voter turnout is higher among certain demographics. For example, people 45 years old or older tend to have a higher voting rate compared to younger generations, women are more likely to vote than men, and white and black Americans vote more frequently compared to other races [6].

The ability of groups of people to impact the result of an election depends on the group's size as well as the group's turnout rate during an election. Minority voters have become increasingly important in recent election cycles as minority population turnout rates rise [3]. Therefore, identifying minority's voting patterns and analyzing the statistical correlation between demographic factors and presidential election voting patterns become crucial. The statistical result could help political strategists to emphasize the significance of personalized messages to target groups as well as creating and distributing customized campaign ads in order to gain an electoral edge [3]. The identified pattern could be especially important in winning the vote in swing states. Not all swing states have sizable non-white populations. In several other 2016 swing states, however, African American, Latino, and Asian American voters compose sizable portions of state populations. Political parties can target the swing states with the most possibility for victory by examining the voting patterns of various demographic groups and concentrating their efforts there.

Literature survey and problem statement

These past couple of years have been marked by a series of life-changing events, from a global pandemic to the violent attack on the capitol, which of course has had a big impact on the political environment of the country. Every election, it is always very hard to predict which candidate will ultimately win the title of president, but the 2020 elections have confirmed that currently, the US is divided more than ever in fundamental views. With the recent change from Republican president Donald Trump to Democratic president Joe Biden, the country has been left with a strongly divided electorate and many people unwilling to compromise even in the midst of a severe health and economic crisis.

In most cases, the major techniques to predict presidential election results are expert opinions and polling. Expert opinion involves asking a group of experts, typically consisting of political scientists and practitioners, the share of the national vote in percentage each party will receive. Polling on the other hand typically utilizes the RealClearPolitics (RCP) poll average as a raw representation of opinion polls and as a benchmark for poll performance. These two techniques are quite dependent on each other since experts tend to follow the polls as basis for their decisions, and unfortunately, polls tend to be subject to various types of error, especially when conducted too early before the election. In a study conducted by Andreas Graefe, he explores the idea of fundamentals-based forecasting to help predict the popular vote in the four U.S. presidential elections from 2004 to 2016 [1]. With this technique, models are based on the theory of retrospective voting and measure performance in different ways. Some examples of fundamentals are economic performance, military fatalities, candidate performance as well as time in office. When combining these forecasts with polls and individual experts, a big majority of the directional error was reduced which demonstrated how useful prior research can be for election forecasting.

Looking more closely at the 2020 election between Biden and Trump, several studies found staggering differences in the demographic patterns and compositions that make up each presidential candidate's electoral coalition as well as county-won populations. According to a Pew Research Center study based on validated 2020 general election voters, compared to Trump voters, Biden voters were younger, more racially and ethnically diverse, more likely to have at least a college degree, and less likely to live in rural areas [4]. Another study conducted by Brookings focused on the total populations residing in the counties that Biden and Trump won and found out that although there are many more Trump-won counties than Biden-won counties (2588 versus 551), 67 million more people lived in counties won by Biden than in those won by Trump (197.9 million versus 130.3 million), signifying the dominance of Biden counties in densely populous metropolitan areas. In addition, similar to the observations in the Pew Research Center study, Biden counties are more likely to be home to a younger, foreign-born, more racially diverse population with higher income and education attainment [2].

Despite the aforementioned studies, **the statistical correlation between demographic factors and presidential election voting patterns has not been analyzed, nor has the predictability of presidential election voting outcomes using demographic factors been explored**. Thus, in our project we are hoping to explore the predictability of presidential election voting outcomes using demographic factors, specifically for the presidential elections from 2012 to 2020. **We hypothesize that the percentage of people with bachelor's degree or higher, total population, and median income are highly correlated with the election voting patterns and most important for predicting the election outcome.** It is important to note that the demographic data provided by the county census includes nonvoters, such as children, noncitizens, and others who did not vote, and therefore certainly did not tell the whole story. Nonetheless, it reflects the communities the voters reside in and is a reasonable proxy for the demographic compositions of the voters.

Methodology

We used multiple linear regression to examine the correlation between demographic factors and presidential election voting patterns in each county. To predict presidential voting outcomes with demographic factors, we used a myriad of regression and classification models including multiple linear regression, random forest regression (regression), logistic regression, and decision trees (classification). The independent variables in these models include total population, median age, median household income, percentage of people with a bachelor's degree or higher, percentage of foreign-born people, and percentage of each race (white, black, Hispanic, Asian). The dependent variable is the share voting for a Democratic candidate (excluding votes for third parties). Whenever needed, we used 5 or 10-fold cross validation to tune the hyperparameters of these models. We used RMSE (root mean squared error) to evaluate regression models and accuracy and AUC (area under the curve) to evaluate classification models.

To improve the performance of the regression models, we explored three different approaches. First, we used L1-regularization (Lasso) and L2-regularization (Ridge) to avoid overfitting. Second, we used natural log transformation on the severely right-skewed independent variables total_population and median_income. Last but not least, we used the difference between 2016 and 2012 data to predict the difference between 2020 and 2016 data.

Overview of Datasets

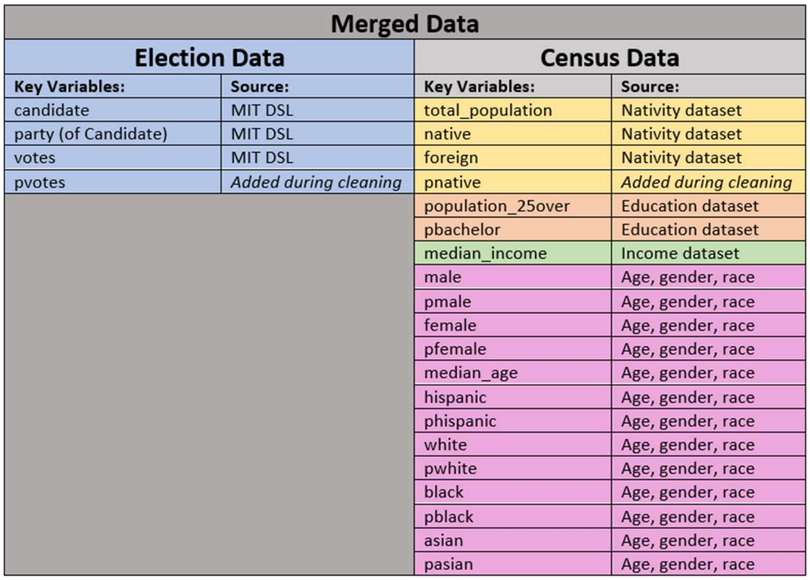
For this project, we are pulling data from the U.S. Census Bureau and the MIT Election Data Science lab. In particular, from the Census Bureau, we used data related to income, education attainment, nativity status, age, gender, and race. From the Data Science lab, we pulled data on the presidential election returns. All datasets from the Census are organized by county, and correspond to the election years of 2012, 2016, and 2020. The election data is also by county; however, the original dataset spans the years 2000-2020.

Data Cleaning Process

Election data: From the original dataset, we extracted the columns of interest – *year*, *state*, *state_po* (two-letter abbreviation of state name), *county_name*, *county_fips* (code that uniquely identifies counties within the U.S.), *candidate*, *party*, and *candidatevotes*. With the exception of *candidatevotes* (class int), all other variables were character data types. To simplify the analysis, we focused on only the Democratic and Republican parties by filtering out the Green party and other parties. After pulling data for election years 2012, 2016, and 2020, we added a new variable *pvotes* for each election year. This variable represents the percent of votes received for each candidate by county.

Census data: The general process of cleaning the datasets from the Census Bureau was similar for each of the four datasets. For each dataset, we first took the original file and pulled out the columns of interest into a separate spreadsheet. After importing this filtered data into R, we renamed the columns so that they were more easily identifiable. The contents of the fips column were then modified so that only the fips for each county, and not any extraneous characters, were included. For the nativity status dataset, we added a new variable, *pnative*, to represent the percent of native-born individuals within the total population of each county.

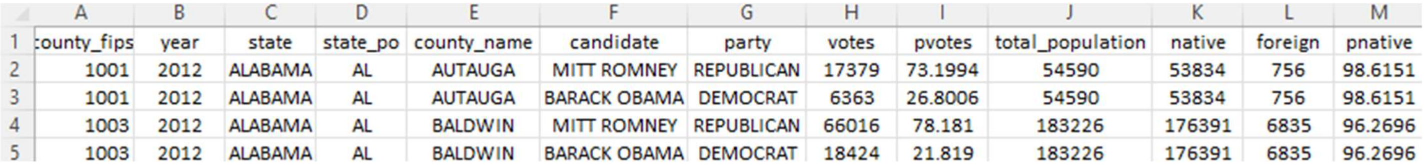
Merging data: For each election year (2012, 2016, and 2020), we merged the election data for that year with the corresponding Census data. As an example, **Table 2** is a screenshot of the first five rows of the cleaned, merged dataset for the 2012 election year. After merging the datasets, any rows with “Na” values were removed. A detailed explanation of each variable can be found in the Appendix, and a general overview of the key variables and their sources can be viewed in **Table 1**.



Merged Data			
Election Data		Census Data	
Key Variables:	Source:	Key Variables:	Source:
candidate	MIT DSL	total_population	Nativity dataset
party (of Candidate)	MIT DSL	native	Nativity dataset
votes	MIT DSL	foreign	Nativity dataset
pvotes	*Added during cleaning*	pnative	*Added during cleaning*
		population_25over	Education dataset
		pbachelor	Education dataset
		median_income	Income dataset
		male	Age, gender, race
		pmale	Age, gender, race
		female	Age, gender, race
		pfemale	Age, gender, race
		median_age	Age, gender, race
		hispanic	Age, gender, race
		phispanic	Age, gender, race
		white	Age, gender, race
		pwhite	Age, gender, race
		black	Age, gender, race
		pblack	Age, gender, race
		asian	Age, gender, race
		pasian	Age, gender, race

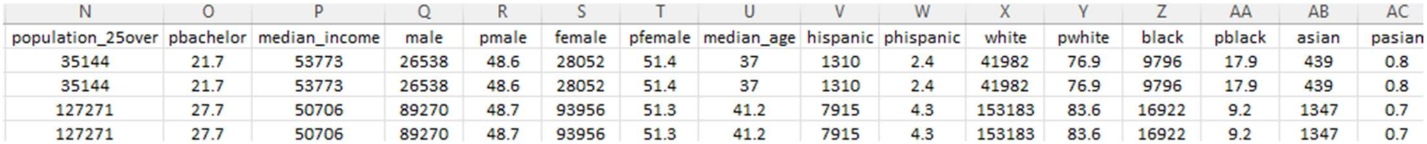
Table 1. Key variables for election data and Census data

Columns 1-13:



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	county_fips	year	state	state_po	county_name	candidate	party	votes	pvotes	total_population	native	foreign	pnative
2	1001	2012	ALABAMA	AL	AUTAUGA	MITT ROMNEY	REPUBLICAN	17379	73.1994	54590	53834	756	98.6151
3	1001	2012	ALABAMA	AL	AUTAUGA	BARACK OBAMA	DEMOCRAT	6363	26.8006	54590	53834	756	98.6151
4	1003	2012	ALABAMA	AL	BALDWIN	MITT ROMNEY	REPUBLICAN	66016	78.181	183226	176391	6835	96.2696
5	1003	2012	ALABAMA	AL	BALDWIN	BARACK OBAMA	DEMOCRAT	18424	21.819	183226	176391	6835	96.2696

Columns 14-29:



N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
population_25over	pbachelor	median_income	male	pmale	female	pfemale	median_age	hispanic	phispanic	white	pwhite	black	pblack	asian	pasian
35144	21.7	53773	26538	48.6	28052	51.4	37	1310	2.4	41982	76.9	9796	17.9	439	0.8
35144	21.7	53773	26538	48.6	28052	51.4	37	1310	2.4	41982	76.9	9796	17.9	439	0.8
127271	27.7	50706	89270	48.7	93956	51.3	41.2	7915	4.3	153183	83.6	16922	9.2	1347	0.7
127271	27.7	50706	89270	48.7	93956	51.3	41.2	7915	4.3	153183	83.6	16922	9.2	1347	0.7

Table 2. Excerpt of merged data from 2012

Difference data: For years 2012 and 2016, and 2016 and 2020, we merged the two datasets, including only data for the Democratic candidates. For each attribute in each county, we calculated the 2020 minus 2016, and 2016 minus 2012 differences, which represent the incremental changes between two consecutive election years. This difference data will be utilized later in some of the regression models.

Exploratory Data Analysis (EDA)

Given that our aim is to investigate the correlation between demographic factors and presidential election voting outcomes, we first created a correlation plot of all variables that would be utilized in later models (**Figure 1**). This included the response variable (*pvotes*) and the predictor variables (*total_population*, *pnative*, *median_income*, *pbachelor*, *median_age*, *pmale*, *phispanic*, *pwhite*, *pblack*, and *pasian*). We used data from the election year 2020 to create the correlation plot shown on the right as this was the most recent dataset available, and this data was filtered to show results for the Democratic party only. We can see from the first row of the plot that there does appear to be a correlation between *pvotes* and each of the predictors, although to varying degrees. In particular, *pbachelor*, *pblack*, and *pasian* demonstrate a relatively stronger positive correlation to *pvotes* while *pwhite* shows a stronger negative correlation with *pvotes*.

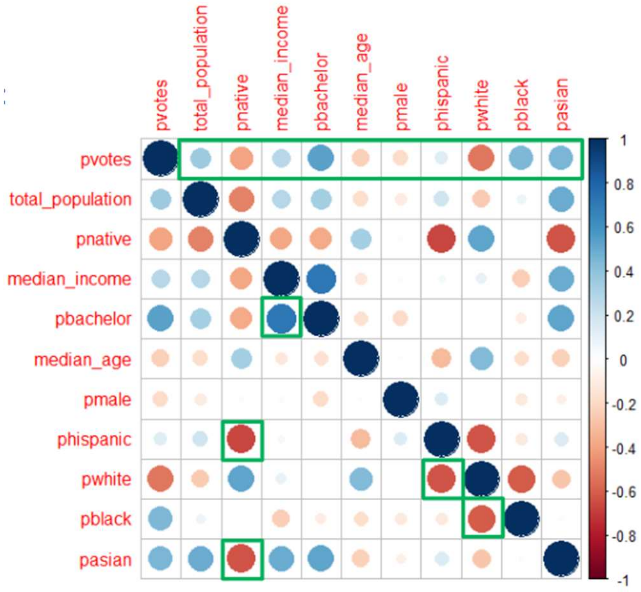


Figure 1. Correlation plot of key variables from 2020 dataset

After further EDA, we uncovered demographic differences between Democratic-won counties and Republican-won counties that were consistent with the literature. Again using 2020 data, we found that Democrats tend to win far fewer counties than Republicans overall, but the counties won by Democrats tend to be much more populous than Republican-won counties. Democratic-won counties, on average, tend to have a population with higher median incomes, a larger proportion of foreigners, and a larger proportion of individuals with bachelor's degrees or higher (that are aged 25 and over). For Republican-won counties, the margin of victory is larger than that of Democratic-won counties. These differences were fairly consistent across election years 2012, 2016, and 2020. Presented in **Table 3** are average values for these metrics for the election year 2020.

	Democrat-won counties	Republican-won counties
Number	537	2574
Percent of votes received by winning party	62%*	72%*
Population	364,834*	50,472*
Number of foreigners	67,793*	2,977*
Median income	61.5k* ± 21.8k	53.5k* ± 12.2k
Education attainment	32.2%*⧫ with bachelor or higher	20.6%*⧫

** values based on averages from 2020 data*
⧫ percent of individuals aged 25 and over

Table 3. 2020 election statistics

Using the difference data between consecutive election years, we uncovered consistent trends in certain demographic factors since 2012: *median_income*, *median_age*, *pbachelor*, and *phispanic* kept increasing and *pnative* and *pwhite* kept decreasing. **Figure 2** shows the histograms of the 2020-2016 difference data.

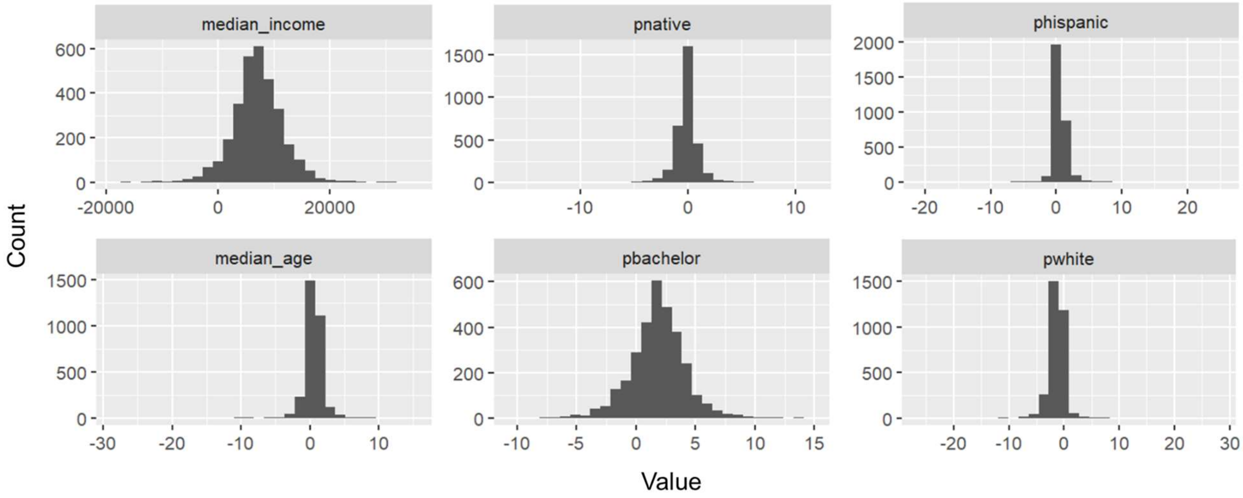


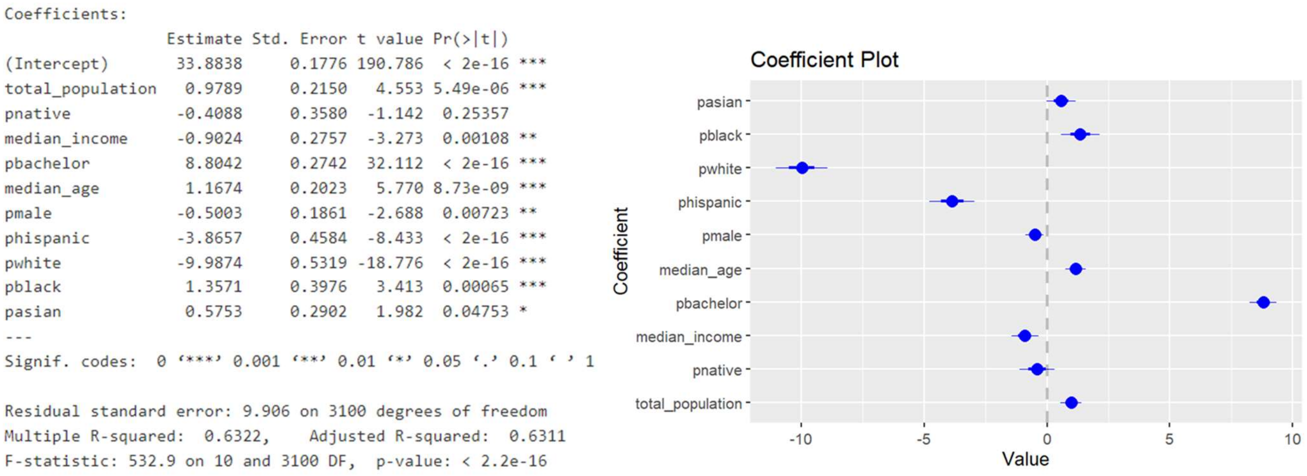
Figure 2. Histograms of difference data between election years 2020 and 2016

Results

Regression models

To examine the correlation between demographic factors and 2020 presidential election voting patterns in each county, we built a multiple linear regression model with the 2020 data. The independent variables were standardized (centered and scaled) in order to compare the coefficients on the same scale. As shown in **Figure 3**, all coefficients are significant except *pnative*. The value of each coefficient is better illustrated in the coefficient plot. We can see that keeping all else constant, an increase in *pbachelor*, *median_age*, *total_population*, *pblack*, or *pasian* is associated with an increase in *pvotes*, whereas an increase in *pwhite*, *phispanic*, *pmale*, *median_income*, or *pnative* is associated with a decrease in *pvotes*. Although this seems to be in conflict with the study indicating Biden counties are more likely to be home to a younger population with higher income [6], one has to consider that it is a different method that is used to reach that conclusion (conditional probability in lieu of linear regression), and that in multiple linear regression, each coefficient is adjusted holding all other independent variables constant. When taking the absolute values of the coefficients, we see the features that have the biggest effects are *pwhite* and *pbachelor*.

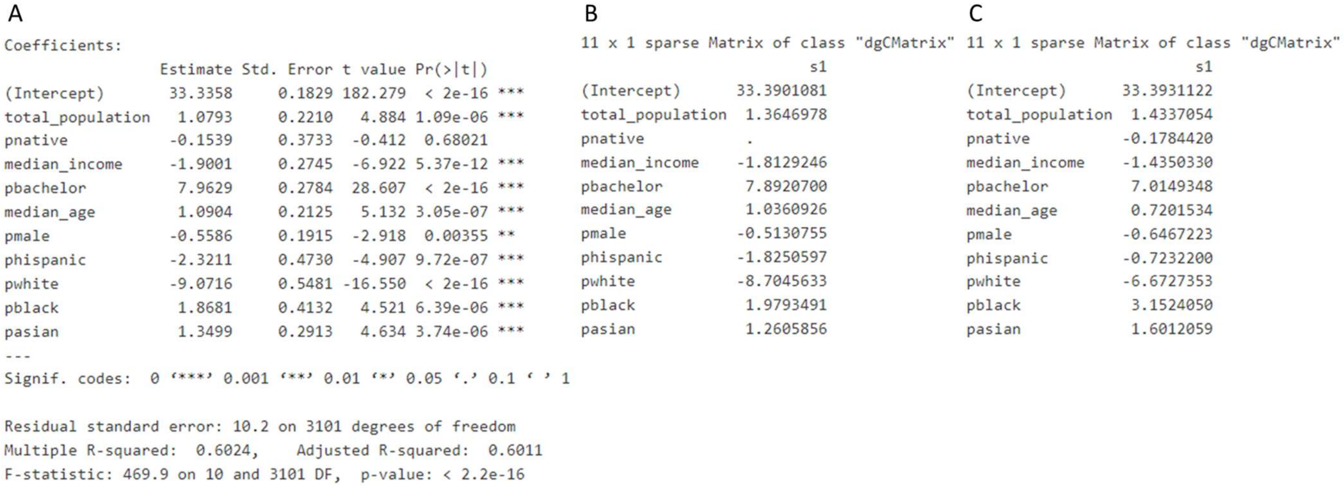
To test the predictability of presidential election voting outcomes using demographic factors, we started by using a basic linear regression model trained on 2016 data to predict 2020 *pvotes* (**Figure 4A**). The model included all independent variables (*total_population*, *pnative*, *median_income*, *pbachelor*, *median_age*, *pmale*, *phispanic*, *pwhite, pblack*, and *pasian*), and achieved a test RMSE of 10.05. Next, to improve model performance, we explored three different approaches. First, we used regularization to avoid overfitting after a 10-fold cross validation to find the best hyperparameter lambda (**Figure 4B and C**). This resulted in a test RMSE of 10.07 for Lasso (L1 regularization) and 10.17 for Ridge (L2 regularization) regression. It is interesting to note that Lasso removed the only insignificant variable *pnative* from the basic linear regression model. Second, we used natural log transformation on the variables that are severely right skewed, specifically *median_income* and *total_population* (**Figure 5A**). With this transformation, the test RMSE came down to 9.48. Last but not least, we used the 2016-2012 difference data to predict *pvotes* in the 2020-2016 difference data (**Figure 5B**). It has proven to be the best improvement of the three as the test RMSE came down to 7.14. The summary of the above linear regression models, including the best hyperparameters and the test RMSE, is shown in **Table 4A**.



```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       33.8838     0.1776 190.786  < 2e-16 ***
total_population   0.9789     0.2150   4.553 5.49e-06 ***
pnative           -0.4088     0.3580  -1.142  0.25357    
median_income     -0.9024     0.2757  -3.273  0.00108 ** 
pbachelor          8.8042     0.2742  32.112  < 2e-16 ***
median_age         1.1674     0.2023   5.770 8.73e-09 ***
pmale             -0.5003     0.1861  -2.688  0.00723 ** 
phispanic         -3.8657     0.4584  -8.433  < 2e-16 ***
pwhite            -9.9874     0.5319 -18.776  < 2e-16 ***
pblack             1.3571     0.3976   3.413  0.00065 ***
pasian             0.5753     0.2902   1.982  0.04753 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.906 on 3100 degrees of freedom
Multiple R-squared:  0.6322,	Adjusted R-squared:  0.6311 
F-statistic: 532.9 on 10 and 3100 DF,  p-value: < 2.2e-16
```

Figure 3. Summary output and coefficient plot from the linear regression model using 2020 data showing the value and p-value of each coefficient.



A

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       33.3358     0.1829 182.279  < 2e-16 ***
total_population   1.0793     0.2210   4.884 1.09e-06 ***
pnative           -0.1539     0.3733  -0.412  0.68021    
median_income     -1.9001     0.2745  -6.922 5.37e-12 ***
pbachelor          7.9629     0.2784  28.607  < 2e-16 ***
median_age         1.0904     0.2125   5.132 3.05e-07 ***
pmale             -0.5586     0.1915  -2.918  0.00355 ** 
phispanic         -2.3211     0.4730  -4.907 9.72e-07 ***
pwhite            -9.0716     0.5481 -16.550  < 2e-16 ***
pblack             1.8681     0.4132   4.521 6.39e-06 ***
pasian             1.3499     0.2913   4.634 3.74e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.2 on 3101 degrees of freedom
Multiple R-squared:  0.6024,	Adjusted R-squared:  0.6011 
F-statistic: 469.9 on 10 and 3101 DF,  p-value: < 2.2e-16
```

B

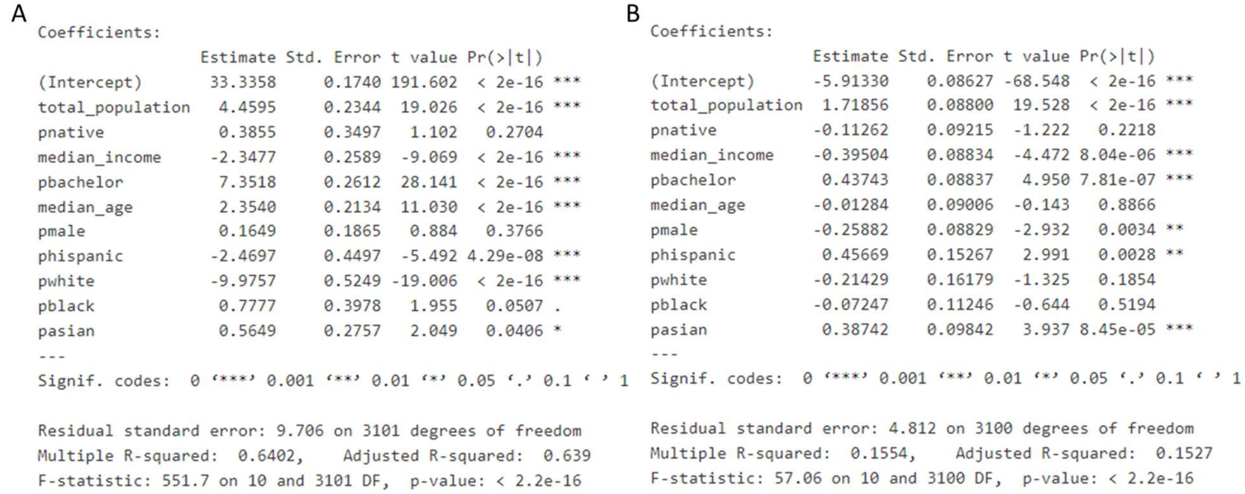
```
11 x 1 sparse Matrix of class "dgCMatrix"
                         s1
(Intercept)      33.3901081
total_population  1.3646978
pnative           .        
median_income    -1.8129246
pbachelor         7.8920700
median_age        1.0360926
pmale            -0.5130755
phispanic        -1.8250597
pwhite           -8.7045633
pblack            1.9793491
pasian            1.2605856
```

C

```
11 x 1 sparse Matrix of class "dgCMatrix"
                         s1
(Intercept)      33.3931122
total_population  1.4337054
pnative          -0.1784420
median_income    -1.4350330
pbachelor         7.0149348
median_age        0.7201534
pmale            -0.6467223
phispanic        -0.7232200
pwhite           -6.6727353
pblack            3.1524050
pasian            1.6012059
```

Figure 4. Summary output from the linear (A), Lasso (B), and Ridge (C) regression models using 2016 data.

While linear regression models are fast with simple interpretations, random forest regression may be more robust in general. Therefore, we also used random forest regression to predict the election outcomes. With a similar workflow, we trained the model on the 2016 data to predict 2020 *pvotes*. With just the default setting, random forest gave a test RMSE of 7.22. To improve the interpretability of the model, we performed permutation tests while building the model to calculate the increase in mean squared error of the predictions (estimated with out-of-bag samples) as a result of variables being permuted (values randomly shuffled). The higher the number, the more important the variable. As shown in **Figure 6A**, the top important features are *pwhite* and *pbachelor*, similar to the basic linear regression model. We also used the 2016-2012 difference data to train the model and then predicted *pvotes* in the 2020-2016 difference data. The test RMSE did not improve. The permutation importance plot showed *total_population* is the top important feature (**Figure 6B**). The summary of the above random forest regression models, including the best hyperparameters and the test RMSE, is shown in **Table 4B**.



```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      33.3358     0.1740 191.602  < 2e-16 ***
total_population  4.4595     0.2344  19.026  < 2e-16 ***
pnative           0.3855     0.3497   1.102   0.2704    
median_income    -2.3477     0.2589  -9.069  < 2e-16 ***
pbachelor         7.3518     0.2612  28.141  < 2e-16 ***
median_age        2.3540     0.2134  11.030  < 2e-16 ***
pmale             0.1649     0.1865   0.884   0.3766    
phispanic        -2.4697     0.4497  -5.492 4.29e-08 ***
pwhite           -9.9757     0.5249 -19.006  < 2e-16 ***
pblack            0.7777     0.3978   1.955   0.0507 .  
pasian            0.5649     0.2757   2.049   0.0406 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.706 on 3101 degrees of freedom
Multiple R-squared:  0.6402,	Adjusted R-squared:  0.639 
F-statistic: 551.7 on 10 and 3101 DF,  p-value: < 2.2e-16
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -5.91330    0.08627 -68.548  < 2e-16 ***
total_population  1.71856    0.08800  19.528  < 2e-16 ***
pnative          -0.11262    0.09215  -1.222   0.2218    
median_income    -0.39504    0.08834  -4.472 8.04e-06 ***
pbachelor         0.43743    0.08837   4.950 7.81e-07 ***
median_age       -0.01284    0.09006  -0.143   0.8866    
pmale            -0.25882    0.08829  -2.932   0.0034 ** 
phispanic         0.45669    0.15267   2.991   0.0028 ** 
pwhite           -0.21429    0.16179  -1.325   0.1854    
pblack           -0.07247    0.11246  -0.644   0.5194    
pasian            0.38742    0.09842   3.937 8.45e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.812 on 3100 degrees of freedom
Multiple R-squared:  0.1554,	Adjusted R-squared:  0.1527 
F-statistic: 57.06 on 10 and 3100 DF,  p-value: < 2.2e-16
```

Figure 5. Summary output from the linear regression models using natural log transformed variables in the 2016 (A) and the 2016-2012 difference data (B).

A

Model	Best hyperparameter	Test RMSE
Linear	N/A	10.04772
Lasso	lambda = 0.037263	10.07186
Ridge	lambda = 0.901823	10.17126
Linear-log	N/A	9.477201
Linear (difference data)	N/A	7.138449

B

Model	Best hyperparameter	Test RMSE
Random forest	default	7.216822
Random forest (difference data)	default	7.488857

Table 4. Summary of the regression models.

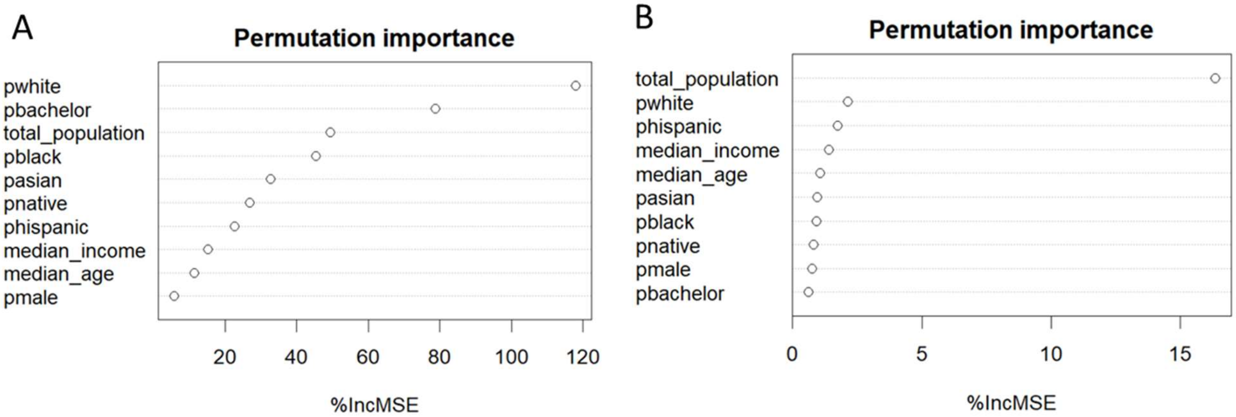
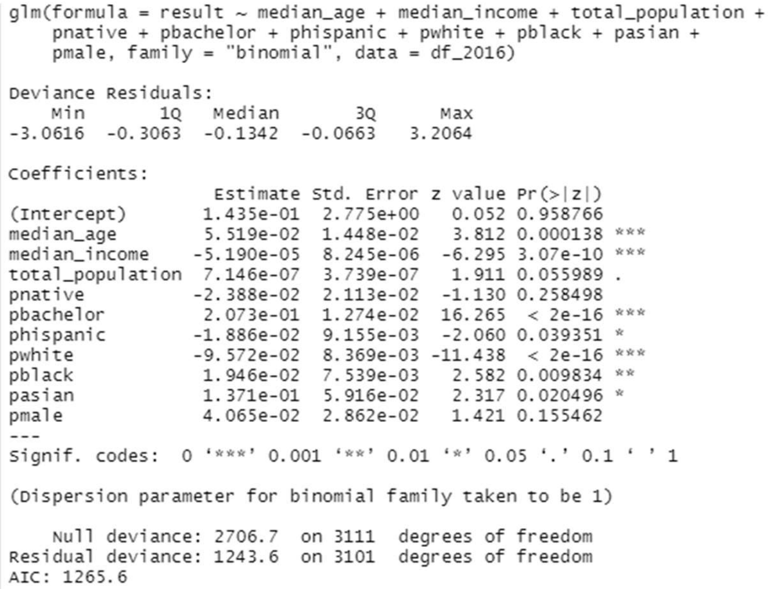


Figure 6. Permutation importance plots of the random forest regression models using 2016 data (A) and 2016-2012 difference data (B).

Classification models

After regression analysis, we moved on with classification models. We started off with a logistic binary classification model, which had 91.3% accuracy and 94.2% AUC **(Figure 7)**. All of our features are statistically significant except *pmale* and *pnative*. The result is very similar to the linear regression model done in regression analysis.


```
glm(formula = result ~ median_age + median_income + total_population +
    pnative + pbachelor + phispanic + pwhite + pblack + pasian +
    pmale, family = "binomial", data = df_2016)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0616  -0.3063  -0.1342  -0.0663   3.2064

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.435e-01  2.775e+00   0.052 0.958766
median_age        5.519e-02  1.448e-02   3.812 0.000138 ***
median_income    -5.190e-05  8.245e-06  -6.295 3.07e-10 ***
total_population  7.146e-07  3.739e-07   1.911 0.055989 .
pnative          -2.388e-02  2.113e-02  -1.130 0.258498
pbachelor         2.073e-01  1.274e-02  16.265  < 2e-16 ***
phispanic        -1.886e-02  9.155e-03  -2.060 0.039351 *
pwhite           -9.572e-02  8.369e-03 -11.438  < 2e-16 ***
pblack            1.946e-02  7.539e-03   2.582 0.009834 **
pasian            1.371e-01  5.916e-02   2.317 0.020496 *
pmale             4.065e-02  2.862e-02   1.421 0.155462
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2706.7  on 3111  degrees of freedom
Residual deviance: 1243.6  on 3101  degrees of freedom
AIC: 1265.6
```

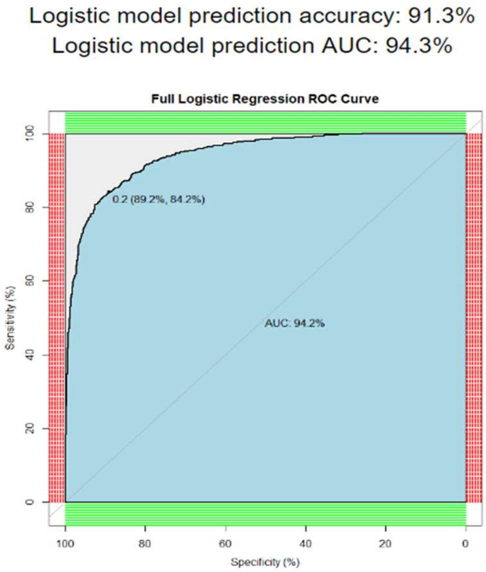


Figure 7. Logistic classification model statistics and ROC graph.

So far none of our models included interaction variables. To better analyze how features interact with each other, we decided to use decision trees as our next model. The performance is very good with 91.6% accuracy and 88.4% AUC (**Figure 8**). Although it has a lower AUC than logistic regression, the ROC curve has shifted to the left compared to the logistic classification model, which indicates a higher specificity and lower sensitivity. This implies that our decision trees were performing better in correctly identifying positives, which in our case referring to correctly identifying Democrats winning.

Figure 9 shows the feature importance and decision tree diagram. *pwhite* is ranked at the top of the feature importance chart, followed by *pbachelor*, *pblack*, *phispanic*, *pasian*, *median_income*, *pnative*, *total_population*, *median_age*, and *pmale*. The feature importance chart is generally consistent with our previous analysis. However, *median_age* and *total_population* that were considered significant in our previous models are ranked very low. After removing them from our model, there were no improvements. We think features which have a low feature importance may still add predictive power to our decision tree model, because unlike logistic regression without interaction terms, decision trees could benefit from combining their information together with information of other features.

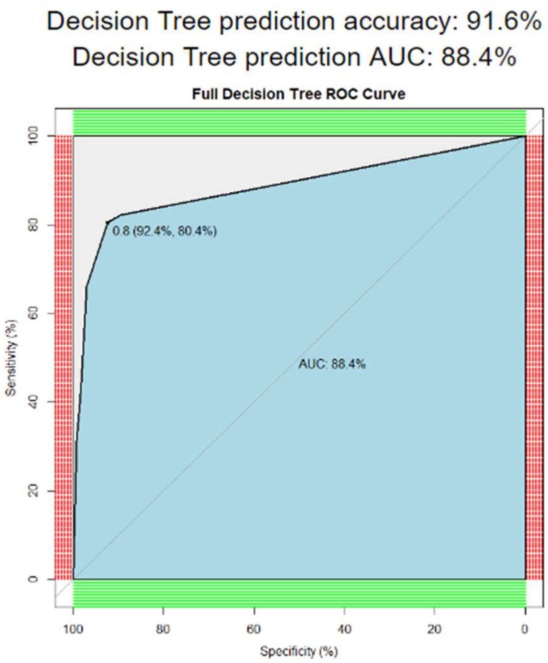


Figure 8. Decision ROC graph

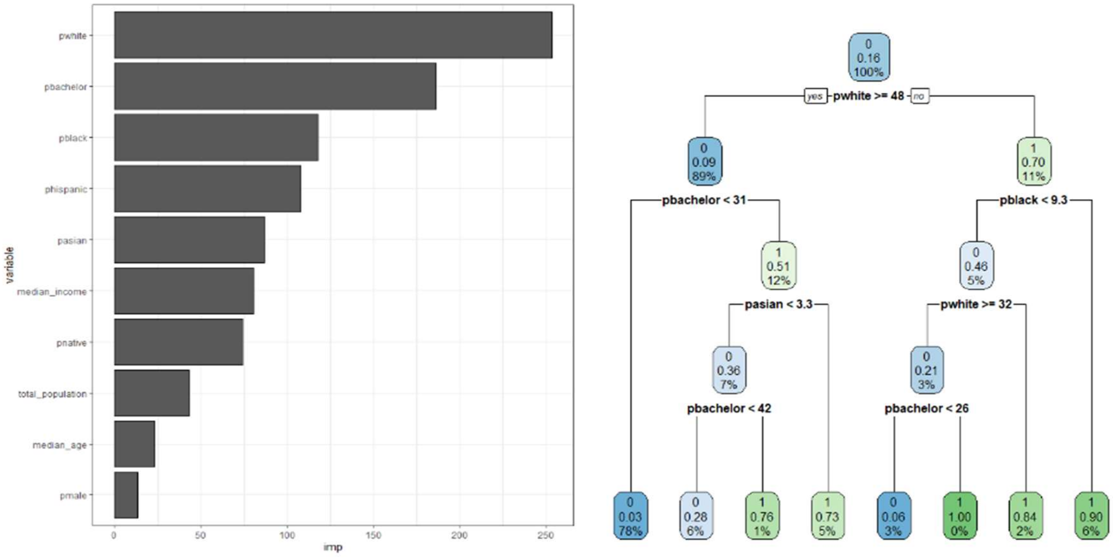


Figure 9. Feature importance chart and decision tree diagram

Conclusion

Overall, all of our models were able to perform fairly well using demographic factors to predict presidential election voting patterns. By examining the coefficients and feature importance, we can conclude that most demographic features are correlated with the election outcomes. Knowing the results of our analysis may affect how campaign managers spend their advertising money. If certain counties or states are identified to be “indecisive” when voting, it may be beneficial to spend more money in those regions to try and win their votes. Marketing decisions will also be impacted, as having information on the demographic factors impacting presidential voting outcomes may allow campaigns to be more targeted towards certain groups over others. Having this targeted information allows campaign managers to be more strategic with their advertisements, and they can potentially save money by focusing their resources on counties with indecisive voters.

pwhite (percentages of white individuals) and *pbachelor* (percentages of individuals who has a bachelor’s degree) are two very dominant features in both classification and regression models. Take decision trees as an example, if we go back to our decision tree graph (**Figure 9**), we can see that the first decision node was based on whether *pwhite* is greater than or equal to 48. Then the left child node shows that the average election outcome is only 0.09 which strongly favors Republican winning. If we go down the same branch, the next decision node asks if *pbachelor* is less than 31, if yes, then the average election outcome is only 0.03. Our decision tree model using only these two features was able to narrow down part of the election results very quickly. This supports our finding from the literature review, which said that compared to Trump voters, Biden voters were younger, more racially and ethnically diverse, more likely to have at least a college degree, and less likely to live in rural areas [4].

Future improvement

Moving forward, there is still room for improvements. For example, sentiment analysis has been widely used for election predictions, so we could collect Twitter data by counties to conduct sentiment analysis and incorporate the results as additional features into our existing models. Also, our EDA shows that some features are highly correlated such as education and income. Multicollinearity could cause bad performances for linear and logistic regression, both of which require variable independence assumptions. In the future, we should re-engineer these highly correlated features such as education and income. Additionally, our model is unbalanced, so we could utilize re-sampling techniques to up-sample or create synthetic data.

References

[1] Andreas Graefe. “Predicting elections: Experts, polls, and fundamentals.” *Judgment and Decision Making,* Vol. 13, No.4, 2018. https://journal.sjdm.org/18/18124/jdm18124.pdf. Accessed 27 Oct. 2022.

[2] Frey, Willian H. “Biden-won counties are home to 67 million more Americans than Trump-won counties.” *Brookings,* 21 Jan. 2021. https://www.brookings.edu/blog/the-avenue/2021/01/21/a-demographic-contrast-biden-won-551-counties-home-to-67-million-more-americans-than-trumps-2588-counties/. Accessed 29 Oct. 2022.

[3] Hudak, John and Christine Stenglein. “How demographic changes are transforming U.S. elections.” *Brookings,* 13 Sept. 2016, https://www.brookings.edu/blog/fixgov/2016/09/13/how-demographic-changes-are-transforming-u-s-elections/. Accessed 20 Oct. 2022.

[4] Igielnik, Ruth, et al. “Behind Biden’s 2020 Victory.” *Pew Research Center*, 30 Jun. 2021. https://www.pewresearch.org/politics/2021/06/30/behind-bidens-2020-victory. Accessed 21Oct. 2022.

[5] *Office of Legislative Research*. Understanding Voter Turnout, 2022. https://www.cga.ct.gov/2022/rpt/pdf/2022-R-0154.pdf. Accessed 23 Oct. 2022.

[6] *United States Census Bureau.* Voting and Registration Visualizations, 2021. https://www.census.gov/topics/public-sector/voting/library/visualizations.html. Accessed 26 Oct. 2022.

Appendix. Explanation of the variables

Variable Name	Data Type	Brief Description of Variable
county_fips	chr	Unique identifier code for each county
year	chr	Year data was collected (2012, 2016, or 2020)
state	chr	State in which data was collected
state_po	chr	2-letter abbreviation of state name
county_name	chr	County in which data was collected
candidate	chr	Name of presidential candidate (2012 - Mitt Romney or Barack Obama, 2016 - Donald Trump or Hillary Clinton, 2020 - Donald Trump or Joe Biden)
party	chr	Party of candidate (Democrat or Republican)
votes	int	Number of votes received by candidate (by county)
pvotes	num	Percent of votes received by candidate (by county)
total_population	int	Total population of county
native	int	Number of individuals born in the U.S. (by county)
foreign	int	Number of individuals born outside the U.S. (by county)
pnative	num	Percent of native-born individuals in a given county
population_25over	int	Number of individuals aged 25 and over (by county)

pbachelor	num	Percent of individuals that attained a bachelor’s degree or higher (by county)
median_income	num	Median household income (by county)
male	int	Number of males in a given county
pmale	num	Percent of males (by county)
female	int	Number of females in a given county
pfemale	num	Percent of females (by county)
median_age	num	Median age of population within a county
hispanic	int	Number of Hispanic/Latino individuals in a given county
phispanic	num	Percent of Hispanic/Latino individuals (by county)
white	int	Number of white individuals in a given county
pwhite	num	Percent of white individuals (by county)
black	int	Number of black individuals in a given county
pblack	num	Percent of black individuals (by county)
asian	int	Number of Asian individuals in a given county
pasian	num	Percent of Asian individuals (by county)