Trump's Victory? Predicting the 2020 Presidential Election Using Logistic Regression and Post-stratification

Jiaheng Li, Yuechen Shen, Yuxin Yang

27/10/2020

Code and data supporting this analysis is available at: https://github.com/yangyu77/STA304_PS3.git

Who will win the election? Trump or Biden? As the U.S. presidential election day approaches, many people are increasingly concerning about the election results. We are interested in seeing if Dodald Trump would win the 2020 re-election race with Democrat Joe Biden, thus we choose to use a regression model to predict the popular vote outcome of the 2020 U.S. election based on the Democracy Fund + UCLA Nationscape data and the American Community Surveys (ACS) data.

Overview: Data Sets After Data Cleaning

The Democracy Fund + UCLA Nationscape data (`survey_data`):

Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from [https://www.voterstudygroup.org/downloads?key=ae440528-c8e4-488f-a153-58e9244de17e].

gender	age	race_ethnicity	employment	state_name	vote_trump	vote_biden
Female	49	White	Employed	Wisconsin	1	0
Female	39	White	Employed	Virginia	0	0
Female	46	White	Employed	Virginia	1	0
Female	75	White	Unemployed	Texas	1	0
Female	52	White	Not in labor force	Washington	1	0
Female	44	White	Unemployed	Ohio	0	0

The Democracy Fund + UCLA Nationscape data set was collected from a public opinion survey project. The original data set contains variables for the political news sources, the political views, the vote choice for 2016 election, the basic demographic information, the employment status, and the state each respondent lives in. Note that the original data set does not contain any observations from the District of Clumbia.

We pick the gender, the age, the race, the employment status, and the state as the predictor variables, and "whether vote for Trump" as well as "whether vote for Biden" as the response variables to build models. The cleaned data set `survey_data` contains 6,445 observations (rows) in total. We note that the original Nationscape data set also includes the vote choice for 2016 election, which could be a good predictor for the current election voting. However, we will not include this variable in the models, because the ACS data set does not include vote choice variable while we need the two data sets to match exactly for model prediction.

The 2018 American Community Surveys (`census_data`):

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

state_name	gender	age	race_ethnicity	employment	n
Alabama	Female	16	Black or African American	Employed	6
Alabama	Female	16	Black or African American	Not in labor force	57
Alabama	Female	16	Black or African American	Unemployed	4
Alabama	Female	16	East Asian	Not in labor force	3
Alabama	Female	16	Other	Not in labor force	5
Alabama	Female	16	Other Asian or Pacific Islander	Not in labor force	1

The original ACS data set contains a comprehensive set of variables for demographic information, including age, sex, race, employment status, educational attainment, and the name of the state. It also incudes information in household level. The raw data set contains 3,214,539 observations, thus it is representative for the entire U.S. population.

In order to make prediction using the fitted model, we need to pick variables that match the predictor variables in the model. Therefore, we pick the equivalent predictor variables as the `survey_data`, which are the gender, age, race, employment status, and the state name. We modify the categories for some variables to get an exact match of variables from the two data sets. We also exclude respondents aged less than 16, because they are ineligible for voting up until 2020.

The cleaned data set `census_data` is already post-stratified. The column variable `n` represents the number of respondents within the same group of state, gender, age, race, and employment status. The details of poststratification will be discussed later.

Model and Specifications

We are interested in predicting the popular vote outcome of the 2020 American federal election (include citation). To do this we will first build *two* binary logistic regression models based on the Democracy Fund + UCLA Nationscape data (the `survey_data`), and apply the fitted models to the ACS data set (the `census_data`) to make predictions for proportion of voters who advocate Donald Trump/Biden, in another words, the probability of Trump/Biden wining the election. We will also utilize a post-stratification technique to account for any under-represented groups in the Nationscape data set. In the following sub-sections I will describe the model specifics and the post-stratification calculation.

We use the binary logistic regression model, because our response variable `vote_trump` is binary, representing whether or not the respondent will vote for Trump. It is exactly the model to be used when the dependent variable is binary (0/1, True/False, Yes/No). Moreover, the logistic regression is a classification algorithm which enable us to find the probability of event to be a success or a failure. It supports categorizing data into discrete classes by studying the relationship from a given set of labelled data.[1]

Model Specifics

For the explanatory variables, we will be using age, which is recorded as a numeric variable, and gender, race, employment status, states as categorical variables to model the probability of voters voting for Donald Trump or Joe Biden.

The five predictor variables are:

x_{age}: a numeric variable, representing the age of the respondent.

x_{gender}: the gender identity of the respondent in the Nationscape survey study, which is either male or female.

$x_{race\_ethnicity}$: the single race that the respondent belongs to in terms of anthropological concept. This variable has 6 categories:

1. White;
2. Black or African American;
3. American Indian or Alaska Native;
4. East Asian;
5. Other Asian or Pacific Islander;
6. Other.

$x_{employment}$: the employment status of the respondent. This variable has 5 categories:

1. Employed;
2. Unemployed;
3. Not in labor force;
4. Student;
5. Other. P.S. Homemaker, retired people, and disabled individuals are categorized as "not in labor force".

x_{state_name}: the state that the respondent live in. This variable includes all US states except the District of Colombia and Puerto Rico.

The logistic regression models we are using are:

- Donald Trump:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_{male} + \beta_2 X_{age} + \beta_3 X_{BlackorAfricanAmerican} + \beta_4 X_{EastAsian} + \beta_5 X_{raceOther} +$$
$$\beta_6 X_{OtherAsianorPacificIslander} + \beta_7 X_{White} + \beta_8 X_{NotInLaborForce} + \beta_9 X_{employmentOther} + \beta_{10} X_{Student} +$$
$$\beta_{11} X_{Unemployed} + \beta_{12} X_{Alaska} + \beta_{13} X_{Arizona} + \beta_{14} X_{Arkansas} + \beta_{15} X_{California} + \beta_{16} X_{Colorado} +$$
$$\beta_{17} X_{Connecticut} + \beta_{18} X_{Delaware} + \beta_{19} X_{Florida} + \beta_{20} X_{Georgia} + \beta_{21} X_{Hawaii} + \beta_{22} X_{Idaho} + \beta_{23} X_{Illinois} +$$
$$\beta_{24} X_{Indiana} + \beta_{25} X_{Iowa} + \beta_{26} X_{Kansas} + \beta_{27} X_{Kentucky} + \beta_{28} X_{Louisiana} + \beta_{29} X_{Maine} +$$
$$\beta_{30} X_{Maryland} + \beta_{31} X_{Massachusetts} + \beta_{32} X_{Michigan} + \beta_{33} X_{Minnesota} + \beta_{34} X_{Mississippi} + \beta_{35} X_{Missouri} +$$
$$\beta_{36} X_{Montana} + \beta_{37} X_{Nebraska} + \beta_{38} X_{Nevada} + \beta_{39} X_{NewHampshire} + \beta_{40} X_{NewJersey} + \beta_{41} X_{NewMexico} +$$
$$\beta_{42} X_{NewYork} + \beta_{43} X_{NorthCarolina} + \beta_{44} X_{NorthDakota} + beta_{45} X_{Ohio} + \beta_{46} x_{Oklahoma} + \beta_{47} X_{Oregon} +$$
$$\beta_{48} X_{Pennsylvania} + \beta_{49} X_{RhodeIsland} + \beta_{50} X_{SouthCarolina} + \beta_{51} X_{SouthDakota} + \beta_{52} X_{Tennessee} + \beta_{53} x_{Texas} +$$
$$\beta_{54} X_{Utah} + \beta_{55} X_{Vermont} + \beta_{56} X_{Virginia} + \beta_{57} X_{Washington} + \beta_{58} X_{WestVirginia} + \beta_{59} X_{Wisconsin} + \beta_{60} X_{Wyoming} + \epsilon$$

Where p represents the probability of voters to vote for Donald Trump in the 2020 U.S. Presidential Election. $vote\_trump = 1$ denotes "will vote for Trump", while $vote\_trump = 0$ denotes "will not vote for Trump".

β_0 represents the intercept of the model.

β_1 represents the relationship between whether vote for Trump and the voter's gender. For every male voter, we expect a β_1 increase in the probability of voting for Donald Trump.

β_2 represents the slope of the model. So, for everyone one unit increase in age, we expect a β_2 increase in the probability of voting for Donald Trump.

β_3 to β_6 represents the relationship between whether vote for Trump and the voter as different race (Black or African American, white, East Asian, Other Asian or Pacific Islander, Other)

β_7 to β_{10} represents the relationship between whether vote for Trump and the voter as different employment statuses (not in labor force, not employment,student, Other)

β_{11} to β_{60} represents the relationship between whether vote for Trump and the different states the voters live in, when all other predictors hold as the same.

- Joe Biden:

The model is the same as the model for Trump, except that the response variable is `vote\_biden`, where p represents the probability of voters to vote for Joe Biden in the 2020 U.S. Presidential Election. $vote\_biden = 1$ denotes "will vote for Biden", while $vote\_trump = 0$ denotes "will not vote for Biden".

Post-Stratification:

Post-Stratification is the process that we use models, which create based on the sample population, to estimate our target population. We use multilevel logistic regression with poststratification to estimate the result of the 2020 U.S. presidential election by using the logistic model which is created based on the smaller-sized survey data. The technique of Post Stratification is very useful, because firstly "it allows the estimating of preference within a specific locality based on a survey taken across a wider area that includes relatively few people from the locality in question, or where the sample may be highly unrepresentative"[2], in addition, we can analyze and build models in a small but representative data set, and then use these to make predictions in a large data set, which can make the statistical analysis less expensive and not take too much time.

We split the observations into strata based on gender, age, race, employment, states. First, the reason why we choose gender is that, after learning some news about the election, we saw reports that some of the actions of the Trump team were against feminism. Besides, Biden has picked a female, Klama Harris, as his running mate. It is reasonable that the female voters would have lower probability to vote for Trump and advocate Biden more. Therefore, we think the variable "gender" is a significant factor affecting the results of the election.

Secondly, we choose "age" because we found that older Americans are more likely to approve and support Trump's campaign philosophy. Meanwhile, "Biden has coalesced support among young voters, and he has become more popular with them. The IOP survey found that Biden had a 34% approval rating among young voters in the spring 2020 poll, but that number has risen to 47% among all youth and 56% among likely voters in that age group. [https://www.usnews.com/news/elections/articles/2020-10-26/biden-bolsters-lead-over-trump-among-young-voters]

Moreover, we pick the variable "race" as we believe that it will be one of the hot topics in the 2020 election. We know the police killing of George Floyd in Minneapolis on June 4. This incident directly caused the U.S. people to march in the streets to resist the unfair treatment of black people in the United States. This incident has once again pushed the opposition to racial discrimination to a climax. Therefore, we think the race of voters will affect the voting during the 2020 presidential election.

In addition, during the Trump administration, he has been committed to increasing employment opportunities in the United States and providing more job opportunities for the people. For example, he has been asking American companies to move production plants from developing countries back to the United States. Hence, Those Americans who are unemployed at home are more likely to support Trump. On the contrary, the top executives and bosses of those companies may not support Trump because their costs are forced to increase a lot.

Finally, we also choose "state", because, in previous general elections, presidential candidates have gone back to fight for the support of the states, especially some major states with more voice, such as California, because the support of these major states may influence the outcome of the entire election. Therefore, we select the variable "state" to be in our cell splits of this post-stratification.

Results

Model Fitting

We use R function glm() for creating the multilevel binary logistic regression models.

$model = glm(vote\_trump \sim gender+age+race\_ethnicity+employment+state\_name, data = survey\_data, binomial)$

$model = glm(vote\_biden \sim gender+age+race\_ethnicity+employment+state\_name, data = survey\_data, binomial)$

Then we apply the fitted model to census_data, and get the log odds estimates. Since we are using logistic models, the estimate that we get directly from the model prediction are the log transformation of the success-failure probability ratio. We need to do further calculation to get the true probability of Trump winning the election.

Next, we calculate the probability of voting for Trump/Biden for each individual. We have constructed strata in the `census_data` using the poststratification technique. It basically divides individual respondents into different groups (strata) based on certain characteristics, such as the sociodemographic information. This technique allows us to take the groups that are under-represented into account and minimize the effects brought by the over-represented groups in the `survey_data`.

At the final step, we sum the probability of voting for Trump of individual U.S. residents and divide the sum by the total number of the residents that are included in the census data set. The final number we get is the prediction of the popular vote outcome.

- **Donald Trump**

Table 1: Trump

term	estimate	std.error	statistic	p.value
(Intercept)	-0.6492806	0.3339287	-1.9443693	0.0518509
genderMale	0.4454792	0.0560760	7.9442091	0.0000000
age	0.0122127	0.0020245	6.0323875	0.0000000
race_ethnicityBlack or African American	-1.8870944	0.2622758	-7.1950774	0.0000000
race_ethnicityEast Asian	-0.8422991	0.3342970	-2.5196132	0.0117484
race_ethnicityOther	-0.5406057	0.2548387	-2.1213645	0.0338911
race_ethnicityOther Asian or Pacific Islander	-0.3804502	0.2792351	-1.3624729	0.1730487
race_ethnicityWhite	0.1953967	0.2308344	0.8464797	0.3972852
employmentNot in labor force	-0.1795127	0.0719856	-2.4937319	0.0126408
employmentOther	-0.3005900	0.2320852	-1.2951708	0.1952613
employmentStudent	-0.9909052	0.1726032	-5.7409420	0.0000000
employmentUnemployed	-0.2941316	0.0945193	-3.1118675	0.0018591
state_nameAlaska	0.2750975	0.7660270	0.3591224	0.7195035
state_nameArizona	-0.3431047	0.2846715	-1.2052652	0.2281010
state_nameArkansas	0.2115537	0.3778759	0.5598497	0.5755820
state_nameCalifornia	-0.7149893	0.2467851	-2.8972139	0.0037649
state_nameColorado	-0.4381730	0.3142268	-1.3944481	0.1631823
state_nameConnecticut	-1.3731229	0.3735276	-3.6760948	0.0002368
state_nameDelaware	-0.7422013	0.4811063	-1.5426970	0.1229043
state_nameFlorida	-0.3300534	0.2502888	-1.3186905	0.1872726
state_nameGeorgia	0.0575415	0.2854721	0.2015660	0.8402561
state_nameHawaii	-0.4265231	0.4833514	-0.8824286	0.3775451

Table 1: Trump *(continued)*

term	estimate	std.error	statistic	p.value
state_nameIdaho	0.0278877	0.4455127	0.0625969	0.9500875
state_nameIllinois	-0.5536205	0.2648420	-2.0903805	0.0365836
state_nameIndiana	-0.3870087	0.3013969	-1.2840500	0.1991245
state_nameIowa	-0.5552702	0.3640547	-1.5252383	0.1271997
state_nameKansas	-0.1593926	0.3738775	-0.4263229	0.6698726
state_nameKentucky	-0.1482389	0.3170735	-0.4675222	0.6401263
state_nameLouisiana	-0.0309083	0.3349787	-0.0922695	0.9264839
state_nameMaine	-0.7669753	0.5159274	-1.4865954	0.1371217
state_nameMaryland	-0.5192201	0.3229371	-1.6078056	0.1078778
state_nameMassachusetts	-1.1905839	0.3216911	-3.7010161	0.0002147
state_nameMichigan	-0.6222260	0.2818975	-2.2072777	0.0272947
state_nameMinnesota	-0.2763696	0.3392789	-0.8145793	0.4153132
state_nameMississippi	0.0068912	0.4056684	0.0169873	0.9864468
state_nameMissouri	-0.3953003	0.3012032	-1.3124040	0.1893838
state_nameMontana	-0.3288614	0.5446584	-0.6037940	0.5459806
state_nameNebraska	-0.5182162	0.5005256	-1.0353441	0.3005082
state_nameNevada	-0.1887446	0.3442934	-0.5482087	0.5835486
state_nameNew Hampshire	-0.6496419	0.5282166	-1.2298779	0.2187428
state_nameNew Jersey	-0.5135485	0.2765275	-1.8571335	0.0632922
state_nameNew Mexico	-1.2741014	0.5278387	-2.4138084	0.0157868
state_nameNew York	-0.5322835	0.2505080	-2.1248167	0.0336019
state_nameNorth Carolina	-0.2508184	0.2750830	-0.9117916	0.3618784
state_nameNorth Dakota	-0.3650177	0.8099723	-0.4506545	0.6522386
state_nameOhio	-0.4979497	0.2633697	-1.8906867	0.0586662
state_nameOklahoma	-0.2221229	0.3500280	-0.6345860	0.5256984
state_nameOregon	-0.7067324	0.3210145	-2.2015589	0.0276965
state_namePennsylvania	-0.3337919	0.2642137	-1.2633404	0.2064669
state_nameRhode Island	-1.0887752	0.7258833	-1.4999315	0.1336322
state_nameSouth Carolina	0.0964605	0.3133629	0.3078235	0.7582166
state_nameSouth Dakota	-0.2501554	0.5617994	-0.4452754	0.6561208
state_nameTennessee	-0.0091854	0.3024147	-0.0303736	0.9757691
state_nameTexas	-0.1303765	0.2515014	-0.5183927	0.6041843
state_nameUtah	-0.5114795	0.3730953	-1.3709085	0.1704035
state_nameVermont	-2.6199277	1.0690032	-2.4508136	0.0142534
state_nameVirginia	-0.5210443	0.2787507	-1.8692124	0.0615933
state_nameWashington	-0.6594050	0.3005752	-2.1938103	0.0282491
state_nameWest Virginia	-0.0802239	0.4009345	-0.2000923	0.8414084
state_nameWisconsin	-0.7731062	0.3056861	-2.5290851	0.0114360
state_nameWyoming	-1.7801849	1.1432785	-1.5570877	0.1194497

The probability of voters to vote for Donald Trump in the 2020 U.S. Presidential Election is :

```
## [1] 0.4135419
```

- **Joe Biden**

Table 2: Biden

term	estimate	std.error	statistic	p.value
(Intercept)	-1.1663784	0.3392925	-3.4376781	0.0005867
genderMale	-0.3232322	0.0539366	-5.9928125	0.0000000
age	0.0046503	0.0019492	2.3858013	0.0170420
race_ethnicityBlack or African American	1.5798438	0.2526904	6.2520923	0.0000000
race_ethnicityEast Asian	1.0360951	0.3082926	3.3607525	0.0007773
race_ethnicityOther	0.6621276	0.2584933	2.5614879	0.0104225
race_ethnicityOther Asian or Pacific Islander	0.6535450	0.2776531	2.3538186	0.0185817
race_ethnicityWhite	0.3093289	0.2422903	1.2766872	0.2017127
employmentNot in labor force	-0.0919034	0.0699966	-1.3129696	0.1891932
employmentOther	-0.0474187	0.2227509	-0.2128777	0.8314224
employmentStudent	0.4691719	0.1272224	3.6878097	0.0002262
employmentUnemployed	-0.0897320	0.0883349	-1.0158160	0.3097170
state_nameAlaska	-0.4123839	0.8532812	-0.4832919	0.6288885
state_nameArizona	0.2244413	0.2813030	0.7978631	0.4249499
state_nameArkansas	-0.5831093	0.4080513	-1.4290097	0.1530014
state_nameCalifornia	0.5488378	0.2409090	2.2781957	0.0227149
state_nameColorado	0.2269322	0.3112388	0.7291257	0.4659248
state_nameConnecticut	0.9078406	0.3347180	2.7122555	0.0066827
state_nameDelaware	0.7673920	0.4496918	1.7064844	0.0879179
state_nameFlorida	0.3008269	0.2456853	1.2244401	0.2207862
state_nameGeorgia	0.0215407	0.2760776	0.0780240	0.9378090
state_nameHawaii	0.6352692	0.4504160	1.4104053	0.1584200
state_nameIdaho	-0.5266261	0.4776940	-1.1024339	0.2702731
state_nameIllinois	0.3941843	0.2579773	1.5279809	0.1265173
state_nameIndiana	0.1665012	0.2991791	0.5565268	0.5778508
state_nameIowa	0.5075401	0.3569225	1.4219895	0.1550293
state_nameKansas	-0.0597037	0.3828231	-0.1559563	0.8760675
state_nameKentucky	0.4272844	0.3142523	1.3596862	0.1739293
state_nameLouisiana	0.1658326	0.3274897	0.5063749	0.6125935
state_nameMaine	0.8480462	0.5038888	1.6830027	0.0923746
state_nameMaryland	0.4261920	0.3074642	1.3861514	0.1657007
state_nameMassachusetts	0.7804286	0.2970383	2.6273667	0.0086049
state_nameMichigan	0.5744213	0.2739542	2.0967787	0.0360132
state_nameMinnesota	0.7325906	0.3329068	2.2005874	0.0277652
state_nameMississippi	-0.0784495	0.3828352	-0.2049172	0.8376368
state_nameMissouri	0.2151628	0.2956248	0.7278237	0.4667215
state_nameMontana	0.4475222	0.5445975	0.8217485	0.4112201
state_nameNebraska	0.0018448	0.5009597	0.0036825	0.9970618
state_nameNevada	0.1189812	0.3352190	0.3549358	0.7226377
state_nameNew Hampshire	0.4777688	0.5211792	0.9167074	0.3592960
state_nameNew Jersey	0.3708143	0.2699761	1.3735078	0.1695945
state_nameNew Mexico	0.6330143	0.4473561	1.4150121	0.1570650
state_nameNew York	0.4117725	0.2456383	1.6763371	0.0936722
state_nameNorth Carolina	0.3677467	0.2682277	1.3710244	0.1703673
state_nameNorth Dakota	-1.0566553	1.1086831	-0.9530725	0.3405533
state_nameOhio	0.2603780	0.2584815	1.0073370	0.3137729
state_nameOklahoma	-0.1918654	0.3582577	-0.5355513	0.5922687

Table 2: Biden *(continued)*

term	estimate	std.error	statistic	p.value
state_nameOregon	0.5112036	0.3119362	1.6388083	0.1012532
state_namePennsylvania	-0.0465002	0.2631128	-0.1767309	0.8597198
state_nameRhode Island	0.7701899	0.6251591	1.2319902	0.2179528
state_nameSouth Carolina	-0.4605716	0.3202774	-1.4380395	0.1504228
state_nameSouth Dakota	0.0679944	0.5869158	0.1158503	0.9077712
state_nameTennessee	-0.2912117	0.3073112	-0.9476118	0.3433271
state_nameTexas	-0.1136761	0.2483637	-0.4577003	0.6471678
state_nameUtah	-0.4027945	0.3961730	-1.0167136	0.3092897
state_nameVermont	1.9871958	0.6929067	2.8679126	0.0041319
state_nameVirginia	0.5129555	0.2697407	1.9016616	0.0572154
state_nameWashington	0.4752414	0.2921007	1.6269779	0.1037418
state_nameWest Virginia	-0.0008599	0.4166993	-0.0020635	0.9983536
state_nameWisconsin	0.5698232	0.2945841	1.9343309	0.0530725
state_nameWyoming	-0.3535680	1.1418215	-0.3096526	0.7568251

The probability of voters to vote for Joe Biden in the 2020 U.S. Presidential Election is :

```
## [1] 0.4082419
```

```
## [1] 280
```

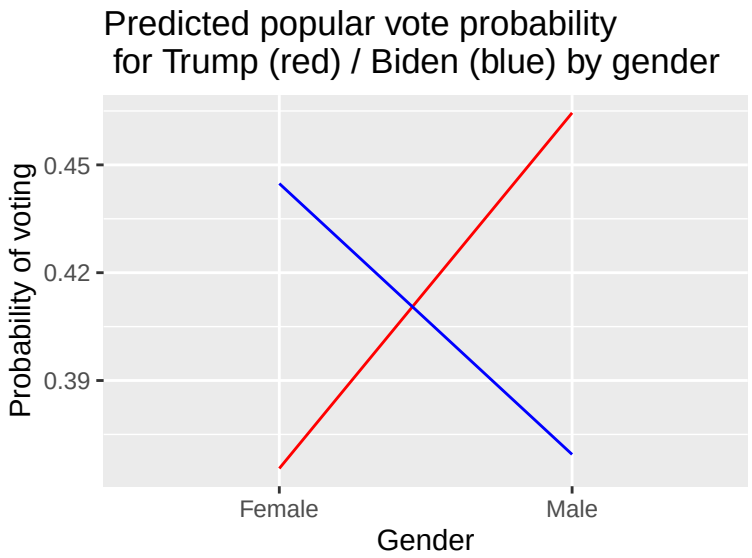
```
## [1] 255
```

We have also predicted the electoral vote outcome based on our models. We predict that Trump will get 280 votes while Biden will get 255 votes. The prediction exclude the District of Columbia, because the data for this region is not in the Nationscape data set, which we used for building models.

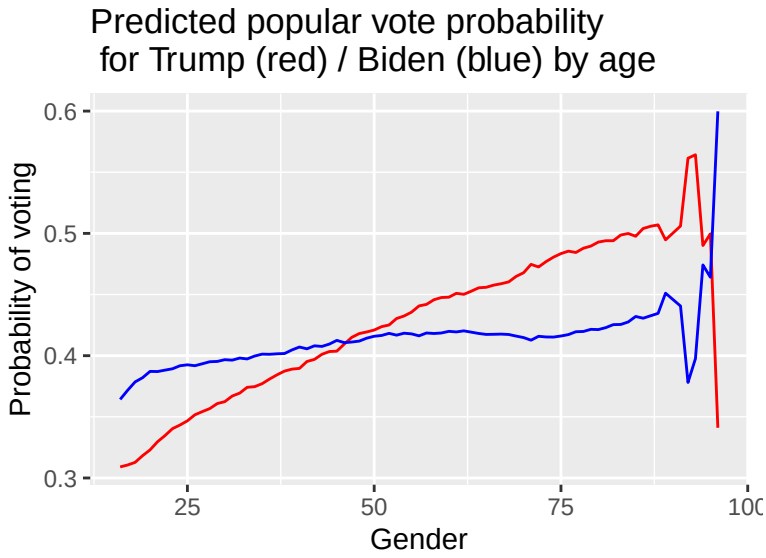
In general, we use multilevel logistic model with Post Stratification which account for gender, age, race, employment and states to estimate that the proportion of voters in favour of voting for Donald Trump to be 41.4% and we also use the same way to estimate that the proportion of voters in favour of voting for Biden to be 40.8%. The total result is that Donald Trump will win in the 2020 presidential election.

Graphs showing the popular vote outcome by groups after post-stratification

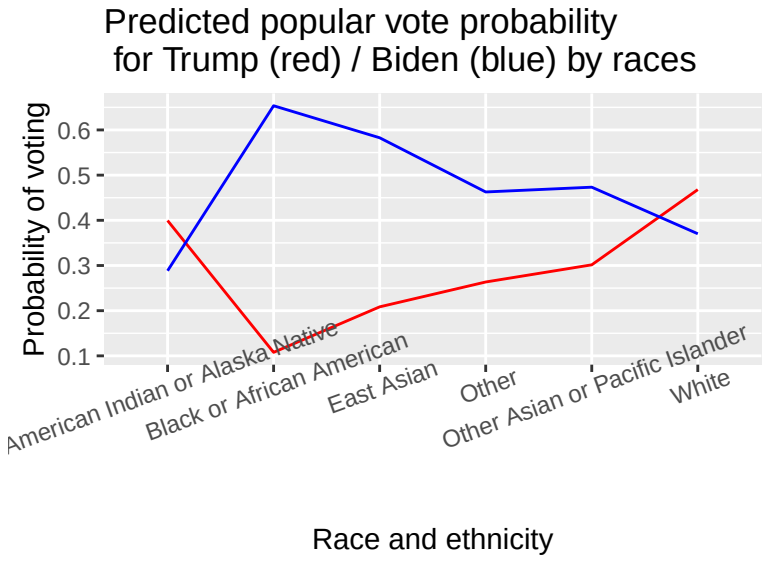
We plot the predicted popular votes based on different group characteristics.



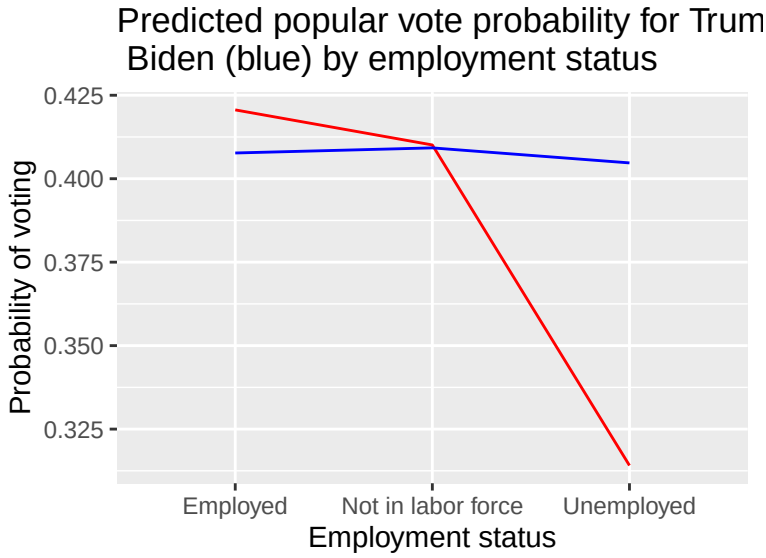
From the above plot, we see that the female voters do support Biden significantly more than Trump, while the male voters do the opposite.



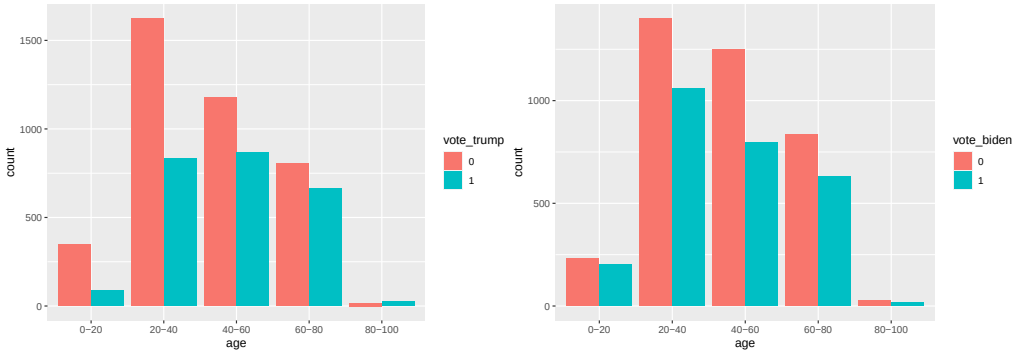
From the above plot, we can see that the young population support Biden more yet they have overal lower voting probability than the older population. While senior people advocate Trump, the oldest population show strong preference for the Democrats.



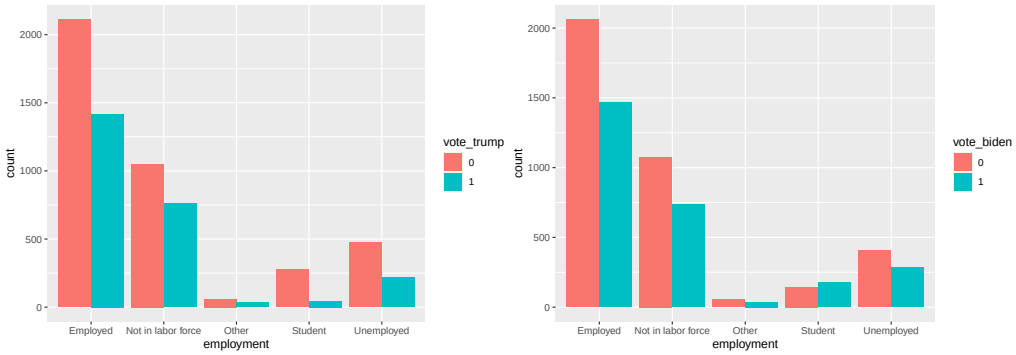
From the above plot, we see that White people and American Indian or Alaska Native people have a higher probability to vote for Trump, while the other races have a higher probability to vote Biden. The Afriican American population has the highest probability among these races.



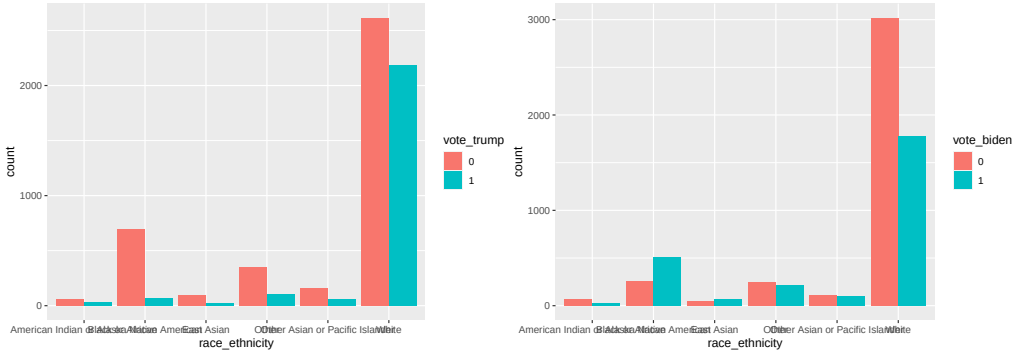
From the above plot, we know that the probability for voting Biden span evenly across groups with different employment status. The probability of the unemployed people voting for Trump is about 10% lower than the other groups.



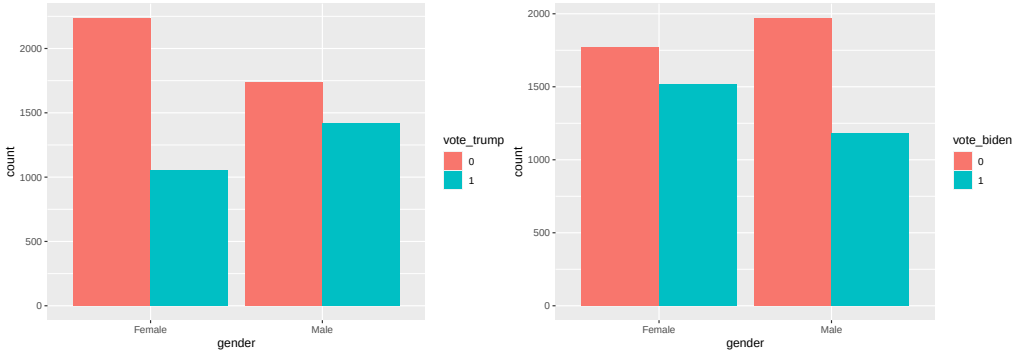
From the above two plots, we can see that among people aged 18-40, Biden's supporters account for the majority. However, Trump has more supporters over 40 than Biden.



It can be seen from these two plots that, except for students, the number of supporters of Trump and Biden in the other employment statuses is similar. For student, most of student choose to vote for Biden.



To start off, the bar graph we have constructed using race ethnicity clearly shows that majority of the race in the U.S is white, followed by the second largest race which is African American, however, we can observe from the bar graph that the total number of white people in the U.S is approximately 6 times the total population of African American, which means that the vote results will rely heavily on the white people population because they are the majority in the U.S. By comparing the blue bars, we see that Trump has a slightly higher bar volume compare to Biden.



It can be seen from these two plots that more women will vote for Biden, while on the other hand, more men will vote for Trump

Discussion

Summary

The final goal of this analysis is to predict and estimate a result of whether Donald Trump or Biden will win the election and become the president of the U.S. In order to come with a valid and meaningful result, we have planned to use 5 independent/explanatory variables for the x axis, the dependent/response variable for the y axis is a binary variable "vote_trump". For the five independent variables, we chose age, gender, race_ethnicity, state and employment, the reason why we value age, gender and race as three important variables are because people with different age, gender and race could have different preferences in picking the president, such as older people might favor more Biden over Trump, whereas younger generation might pick Trump over Biden. In terms of the other two variables, our group thinks that Trump and Biden have their own style and policies, which will affect people in different states and people in different work fields. Our group previously decided to use the variable household_income, however, we later determined that since household_income is a nominal variable that contains lots of different numbers, the binary variable employment is easier to analyze and make conclusions since it only contains the answer unemployed or employed, which makes the variable more straightforward. After finalizing the variables, we have used the sample data to create a logistic model to analyze the relationship between the independent and dependent variables and observe the results. Finally, taking the results and analyzing them from the logistic model, we use post stratification technique which allows us to use these sample data we have and outputs the estimate of who has a higher chance of becoming the president.

Conclusion

RACE

According to the analysis and graphs we have constructed with the 5 variables; it suggests that Trump is likely to have more supports than Biden. One important fact that we obtain from this bar graph is that other race in the U.S votes Biden instead of Trump, which shows the fact that Trump is not being favored in other races, such as African American, Indian, Asian, etc. The reason behind that is because Trump has the idea of white supremacy, according to the report from The New York Times, On July 12th, Trump has verbally attacked four congresswomen saying that they should return to their home country and help with their terrible government. This conflict clearly shows Trump's unfriendly manner towards people that are not born in American and are not white, which is why Trump has very few supporters from other race compare to Biden. Overall, through the analysis of race, we can see that Trump has more white supporters whereas Biden has more supporters from other race as well, by approximately the values of these supporters, Trump and Biden will have very close results but Trump will be a little higher since the majority of the population in the U.S is white.

Employment

During the analysis of the variable employment, we see that the votes for both Trump and Biden are almost identical for employed people and people not in the labor force. The reason behind that is these population does not care too much about who becomes the president because none of the policies nor changes will greatly affect them anyways. However, we see that unemployed population supports Biden instead of Trump, BBC news explains that this is mainly because Trump's action and policy during the Pandemic has caused the highest unemployment rate in the recent 80 years. During the pandemic, Trump has played a poor role in aiding the U.S economic, as a result, many unemployed individuals and teenagers all lost confidence in Trump, which is why we see more votes in Biden than Trump in terms of employment.

Age

From the variable age, we can still see a similar result between both Trump and Biden, the major difference is at the population that is 20-40 years old. This can be related to the employment variable, majority of the population that is 20-40 years old are also in the labor force, as U.S economy is dramatically dropping during the pandemic and Trump seems to lack the ability to aid the economy and slows down the unemployment, younger generation people will turn to Biden instead for another try. Other age group have an even split in both Trump and Biden, due to the increasing younger generation people that votes for Biden, Biden has a slightly higher vote when analyzing the vote rate with age.

Gender

Through the analysis with gender and vote_trump, we clearly see a gender gap, Trump has more male supporters and Biden has more female supporters. Such gap and preference can be explained by the fact that Biden belongs to the Democratic party whereas Trump belongs to the republican party. Democratic party has a core idea of freedom, females is more emotional in general and care more about well-fare, freedom, diversity and care. These preferences will give the Democratic party more female supporters. On the other hand, republican party features power, responsibility and promotes business and free trade, which will be more favorable by males. Since Trump has more male supporters and Biden has more female supporters, the difference almost evens out in the bar graph, which again produces a similar chance of winning between Trump and Biden.

State

In terms of state, states that are along the great lakes appears to be support Trump, such as Pennsylvania, Wisconsin and Ohio. These states are known be U. S's manufacturing state that relies heavily on industries and manufacturing goods, as discussed earlier, the republican party that Trump belongs favors free trade and promotes business, which is beneficial for the workers and people that lives in these states. As both Trump and Bidens are travelling from state to state conducting speeches and polling for votes, we again observe an almost even split in the states that supports each candidate.

We have made our prediction that Trump gets a popular vote of 41.4%, while Biden gets 40.8%. The number of votes are very close with only 0.6% difference. Both presidential nominees gain advantages in certain variables and some disadvantages on the other. Referring to our predictions strictly, then Donald Trump will win the election. There are several reasons to explain. First of all, from the variable race_ethinicity, we see that white people makes up a large portion of the entire U.S population, and a lot of them are supporting Trump rather than Biden. Generally, the Democratic Party has more Black supporters, but the Blacks usually have lower voting rate during the past elections. Although Biden has more supporters in other race, majority could overtake the minority vote. Secondly, Trump has set a few goals back in the 2016 election, such as quiting the Iranian nuclear deal, quiting NATO, reducing corporate income tax, protecting U.S intellectual Property Rights, building the Mexican Wall, etc. Although not all the objectives sound perfect, but Trump has done what he said, Trump did act on what he says which is an important trait as a politician. Thirdly, Trump did try to promote the U.S economy, since 2017, Trump has largely reduced tax for U.S corporates, lowering the cost of trading, protects U.S owned corporates and increasing job opportunities, all of these has result in a low unemployment rate of only 3.5% in the late 2019 before the pandemic. Also, during the George Floyd incident, Trump has strictly ordered the government to reduce the violent action and protest. Based off the estimated proportion of voters in favor of voting for Donald Trump being 41.4%, we predict that Donald Trump may have higher chance to win the election.

However, it is the truth that Trump did not fulfill his duty in controlling the COVID-19 pandemic. Since the overal death surpass 100000, there are increased disapproval sounds towards his response to the pandemic. Given we predicted the vote outcome based on models built with data in June 2020, there are 4 months left to the presidential election day. The unsure voters may make new decisions, which is the part we wouldn't know. Biden will still have possibilities to win the election.

Weakness and Next Step

The weakness of this analysis is the variable part, the variables that we pick are all very relevant and closely related with the response variable, in each bar graph, we do see a difference in supporters of the two candidates, however, the results are too similar in most cases. Which makes it hard to differentiate who has a slight advantage when comparing with a certain variable. Let's take the variable age for example, we have constructed five vertical bars for each candidate, the result turns out that Trump has more volume in two of the bars and Biden also has more volume in the other two bars, so the result evens out, which is hard to come up with a conclusion or to discuss who actually stands out.

In terms of next step, there are two things that can be improved, one is finding a variable that can clearly differentiates a result between the two candidates. The other is about graphing the plot, in this report, for a more direct and simple view of the graph, the proportion of vote and no vote is stacked on each other, which takes some time to compare who has greater value than the other. Next time we will try putting them side by side, although that could take up more space and makes the graph more complex, we think this might be able to provide a straighter forward comparison.

Appendix

R Code

The popular vote outcome for Trump

```
- Creating the Model
model_logit_trump <- glm(vote_trump ~ gender + age + race_ethnicity + employment +
                           state_name, data = survey_data, binomial)

- Model Summary (to be reported in Results section)
kable(broom::tidy(model_logit_trump), "latex", booktabs = T) %>%
  kable_styling(latex_options = c("striped", "scale_down"))

- Apply the fitted model to census_data, and get the log odds estimates
census_data$logodds_estimate <- model_logit_trump%>%
  predict(newdata = census_data)

- Calculate the probability of voting for Trump for individuals within each stratum
census_data$estimate <- exp(census_data$logodds_estimate)/(1+exp(census_data$logodds_estimate))

- Sum up the probability of voting for Trump for each stratum
census_data <- census_data %>%
  mutate(elect_predict_prob = estimate * n)

Trump_prob <- sum(census_data$elect_predict_prob)/sum(census_data$n)
Trump_prob
```

The popular vote outcome for Biden

```
- Creating the Model
model_logit_biden <- glm(vote_biden ~ gender + age + race_ethnicity + employment +
                           state_name, data = survey_data, binomial)

- Model Results (to Report in Results section)
kable(broom::tidy(model_logit_biden), "latex", booktabs = T) %>%
  kable_styling(latex_options = c("striped", "scale_down"))

- Apply the fitted model to census_data, and get the log odds estimates
census_data$logodds_estimate2 <- model_logit_biden%>%
  predict(newdata = census_data)

- Calculate the probability of voting for Biden for individuals within each stratum
census_data$estimate2 <- exp(census_data$logodds_estimate2)/(1+exp(census_data$logodds_estimate2))

- Sum up the probability of voting for Biden for each stratum
census_data <- census_data %>%
  mutate(elect_predict_prob2 = estimate2 * n)

Biden_prob <- sum(census_data$elect_predict_prob2)/sum(census_data$n)
Biden_prob
```

The electoral vote outcome

```
polling <- census_data%>%
  group_by(state_name) %>%
  mutate(state_poll_trump = sum(elect_predict_prob)/sum(n)) %>%
  mutate(state_poll_biden = sum(elect_predict_prob2)/sum(n)) %>%
  dplyr::select(state_name, state_poll_trump, state_poll_biden) %>%
  distinct() %>%
  mutate(trump_victory = ifelse(state_poll_trump > state_poll_biden, 1, 0)) %>%
  mutate(biden_victory = ifelse(state_poll_biden > state_poll_trump, 1, 0))

polling$electoral_votes <- c(9, 3, 11, 6, 55, 9, 7, 3, 29, 16, 4, 4, 20, 11, 6, 6, 8, 8, 4, 10,
                             11, 16, 10, 6, 10, 3, 5, 6, 4, 14, 5, 29, 15, 3, 18, 7, 7, 20, 4,
                             9, 3, 11, 38, 6, 3, 13, 12, 5, 10, 3)

sum(polling$trump_victory * polling$electoral_votes)
sum(polling$biden_victory * polling$electoral_votes)
```

Table: `polling`

state_name	state_poll_trump	state_poll_biden	trump_victory	biden_victory	electoral_votes
Alabama	0.4747927	0.3723978	1	0	9
Alaska	0.5961780	0.2184921	1	0	3
Arizona	0.4545889	0.3698108	1	0	11
Arkansas	0.5591520	0.2311923	1	0	6
California	0.3326466	0.4749932	0	1	55
Colorado	0.4424760	0.3687895	1	0	9
Connecticut	0.2326868	0.5531089	0	1	7
Delaware	0.3378097	0.5390997	0	1	3
Florida	0.4411857	0.4158937	1	0	29
Georgia	0.4615925	0.3920673	1	0	16
Hawaii	0.3560271	0.5384530	0	1	4
Idaho	0.5678420	0.2096760	1	0	4
Illinois	0.3877581	0.4334852	0	1	20
Indiana	0.4441077	0.3655256	1	0	11
Iowa	0.4292020	0.4307467	0	1	6
Kansas	0.5099866	0.3091071	1	0	6
Kentucky	0.5042948	0.4219572	1	0	8
Louisiana	0.4427371	0.4260106	1	0	8
Maine	0.3877485	0.5120006	0	1	4
Maryland	0.3413094	0.4897392	0	1	10
Massachusetts	0.2655390	0.5170077	0	1	11
Michigan	0.3865852	0.4671368	0	1	16
Minnesota	0.4931459	0.4889511	1	0	10
Mississippi	0.4255935	0.3900828	1	0	6
Missouri	0.4421552	0.3790988	1	0	10
Montana	0.4925096	0.4062598	1	0	3
Nebraska	0.4327369	0.3174887	1	0	5
Nevada	0.4626535	0.3694544	1	0	6
New Hampshire	0.4091133	0.4236301	0	1	4
New Jersey	0.3865580	0.4369207	0	1	14
New Mexico	0.2551386	0.4525298	0	1	5
New York	0.3715740	0.4525818	0	1	29
North Carolina	0.4308955	0.4461978	0	1	15
North Dakota	0.4781779	0.1350564	1	0	3
Ohio	0.4127288	0.3947912	1	0	18
Oklahoma	0.4784392	0.2807001	1	0	7
Oregon	0.3846610	0.4331664	0	1	7
Pennsylvania	0.4598106	0.3203429	1	0	20
Rhode Island	0.2934924	0.5090061	0	1	4
South Carolina	0.4928383	0.2810367	1	0	9
South Dakota	0.5034958	0.3204686	1	0	3
Tennessee	0.5091787	0.2861590	1	0	11
Texas	0.4771352	0.3184135	1	0	38
Utah	0.4218472	0.2312843	1	0	6
Vermont	0.0929608	0.7639266	0	1	3
Virginia	0.3759983	0.4775205	0	1	13
Washington	0.3817191	0.4338112	0	1	12
West Virginia	0.5356388	0.3163540	1	0	5
Wisconsin	0.3759813	0.4493848	0	1	10
Wyoming	0.1867113	0.2379598	0	1	3

Reference

[1] GeeksForGeeks, A., AmiyaRanjanRout, & GeeksForGeeks, T. (2020, September 02). Advantages and Disadvantages of Logistic Regression. Retrieved November 01, 2020, from https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

[2] Multilevel regression with poststratification. (2020, October 14). Retrieved November 01, 2020, from https://en.wikipedia.org/wiki/Multilevel_regression_with_poststratification

[3] Biden Bolsters Lead Over Trump Among Young Voters. (n.d.). Retrieved November 01, 2020, from https://www.usnews.com/news/elections/articles/2020-10-26/biden-bolsters-lead-over-trump-among-young-voters

[4] Charles. (2019, July 14). Trump's Tweets Prove That He Is a Raging Racist. Retrieved November 01, 2020, from https://www.nytimes.com/2019/07/14/opinion/trump-twitter-racism.html?_ga=2.154681726.756034816.1604194097-739269052.1604194097

[5] Team, R. (2020, September 23). US 2020 election: The economy under Trump in six charts. Retrieved November 01, 2020, from https://www.bbc.com/news/world-45827430

[6] Sokolove, M. (2020, October 23). Why Does Trump Win With White Men? Retrieved November 01, 2020, from https://www.nytimes.com/2020/10/23/opinion/sunday/gender-gap-2020-election.html

[7] Barry, E. (n.d.). The 2020 Battleground States: Updates on the Swing Voters. Retrieved November 01, 2020, from https://www.nytimes.com/live/2020/battleground-states-2020-election