Relationship between race/gender diversity and Democratic/ Republican votes in Senate Election

Fatemeh Asgarinejad (PID: A59006734)

ABSTRACT

In this study I explore US Senate Votes within years and distribution of US race/gender diversities within the states and study the relationship between distribution of the so-called different communities in each particular state and the winner of the Senate election in that state. The study consists of several pre-processing steps and ultimately concludes in prediction. Finally, implementing classification methods including K Nearest Neighbors and Random Forest, We can reach to a promising winner prediction accuracy.

1. INTRODUCTION

Although many states/counties are considered safe for political parties, some others are not reliable at all. Such that, previously claimed Democratic-held seats turn to be be Republican and vice versa. The race between two parties in a state/county is sometimes too close that follows a runoff. Moreover, some states and counties gradually lean towards another party. This gradual change which leads to flipping of a state/county blue or red, is affected by numerous measurable and immeasurable factors, one of which could be distribution of different races. Based on Migration Policy Institute (MPI) report, as of 2018, 44.7 million immigrants live in USA, who come with different races and ethnicities, many of whom will become citizens and eligible to vote within a few years. [5]

Also, there's a common sense between many immigrants that the states with more immigrants and mixed raced communities are more leaning towards a particular group (Democrats) like California and New York. My goal is evaluating the accuracy of this common sense.

While a few hundreds of votes can flip a county or even an state, it worth studying whether increase in size of this significant groups (races and/or genders) is correlated with a state's leaning towards a particular party or not, since candidates of different parties mostly often have some strict immigration policies. In this project I will study the effect of race/gender distributions and republican/democratic votes. I also do a lot of exploration over the datasets.

2. DATASETS

The dataset which are used in this study are from US government census website and Harvard dataverse. "Annual State Resident Population Estimates for 6 Race Groups (5 Race Alone Groups and Two or More Races) by Age, Sex, and Hispanic Origin" [1][2]
Which includes the population based on Age, Sex and other attributes is obtained from US government dataset and U.S. Senate votes 1976–2018 [3] from Harvard dataverse website. Each of the datasets is thoroughly explored in the attached Jupyter Notebook. I need to add that I could just find datasets of race/generation distribution for year 2010. Hence, the investigation based on race/gender are for year 2010 and other until 2018.

3. DATA ANALYSIS AND VISUALIZATION

Each of the states in the United States hosts a unique distribution of races due to various reasons including economy, weather, educational institution, etc. Based on [1], We can divide our race/gender combinations into 12 categories. Races are "White Alone", "Black or African American Alone", "American Indian or Alaska Native Alone", " Female Asian Alone", "Female Native Hawaiian and Other Pacific Islander Alone" and "Two or more races" and genders in the dataset are "men" and "women" constituting twelve pairs of race/gender categories. Figure 1 depicts distribution of Female Black of African American Alone across the country.

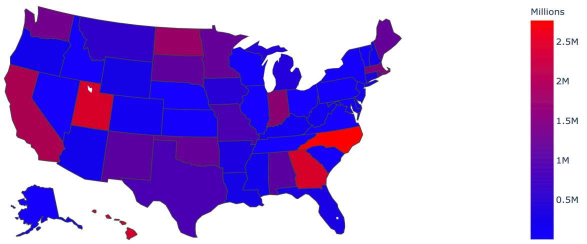

Figure 1: distribution of female black or African Americans Alone in 2010

Figure 2 depicts the ratio of votes for Democrats, Republican and other parties in year 2018 over the whole country. As shown in the figure, 6.3% of the US citizens have voted for other parties than Democrats and Republicans and Democrats have gained over 16 percent higher number of votes than Republican party Candidates. [3]

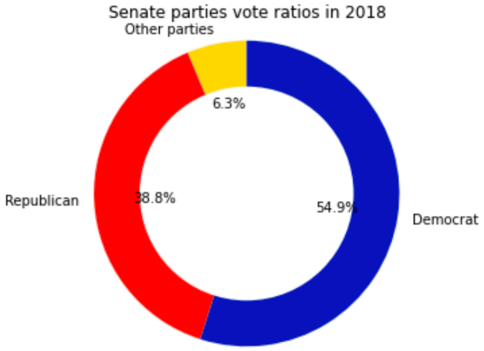

Figure 2: Senate parties votes ratios in 2018

There are 165 different parties who have participated in US Senate election from 1976 to 2018. From which Democrats and Republicans possess the highest ration. This ratio is depicted in figure 3. Ratio of Republican and Democratic Votes doesn't change intensively within years but votes for other parties seems to elevate in recent Senate Elections. [1][2]

Also race/gender populations of year 2010 for each particular state depicted in Figure 4. In this figure, as mentioned before, each combination of race/gender is considered as different categories.[1]

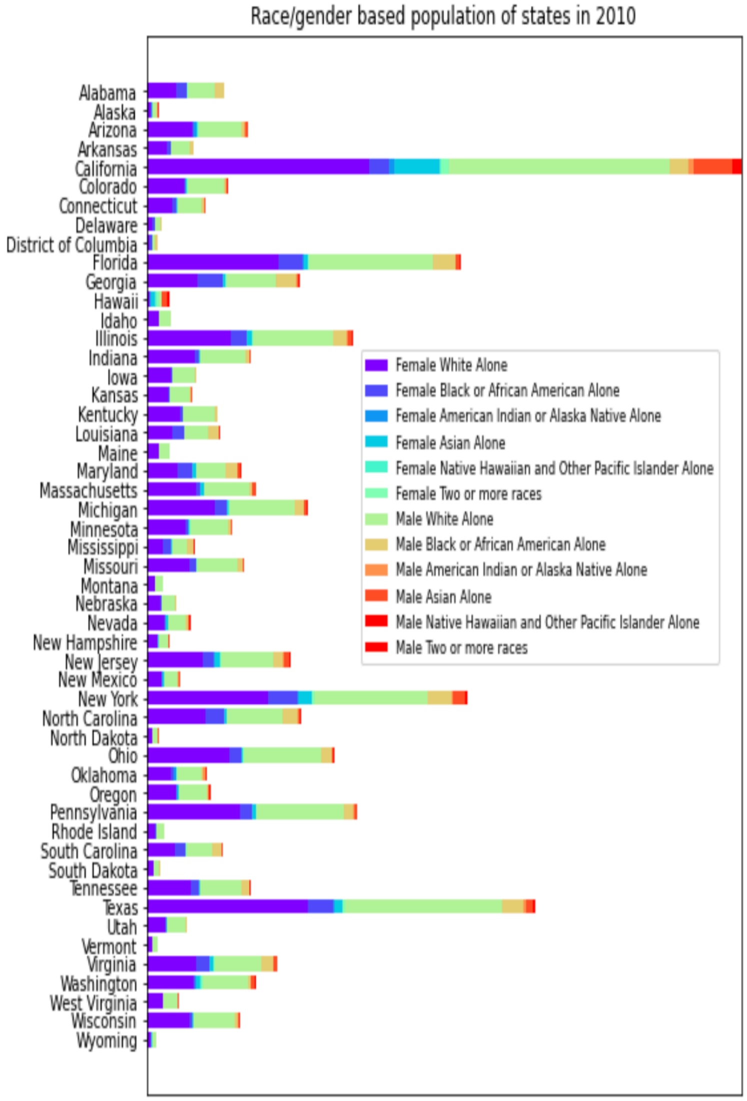

Figure 3: Distribution of race/gender categories over all US states

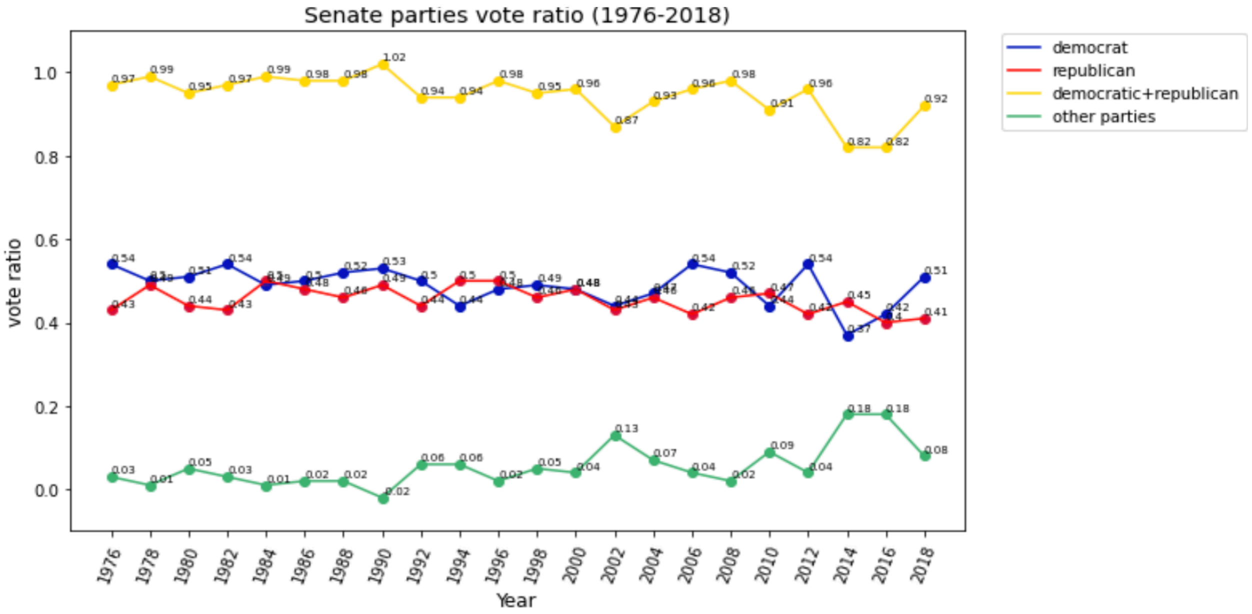

Figure 4: Senate parties vote rratio (1976-2018)

Figure 5 shows the the Minimum, Maximum, Standard Deviation and average of each race population over all states. [1] Obviously, White men and women has the highest population in all states, also their minimum population often proceeds even the maximum population of other races.

Also as expected, number of women proceeds men population in all states. This is depicted in figure 6. Based on figure 7 which shows number of votes of each Democrat and Republican Candidate within 1976 to 2018, it seems that outliers (high values) usually happen for Democrats.

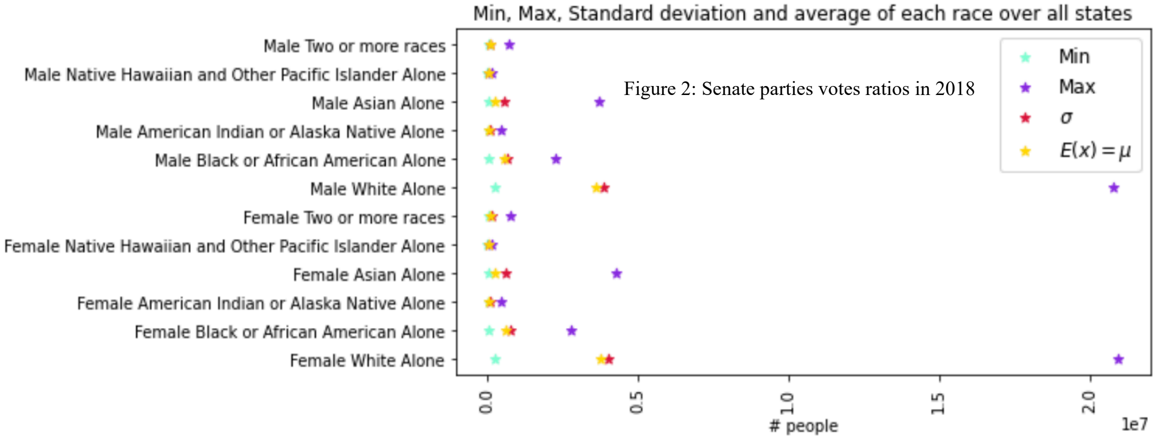

Figure 2: Senate parties votes ratios in 2018

Figure 5: Min, Max, Standard deviation and average of each race over all states

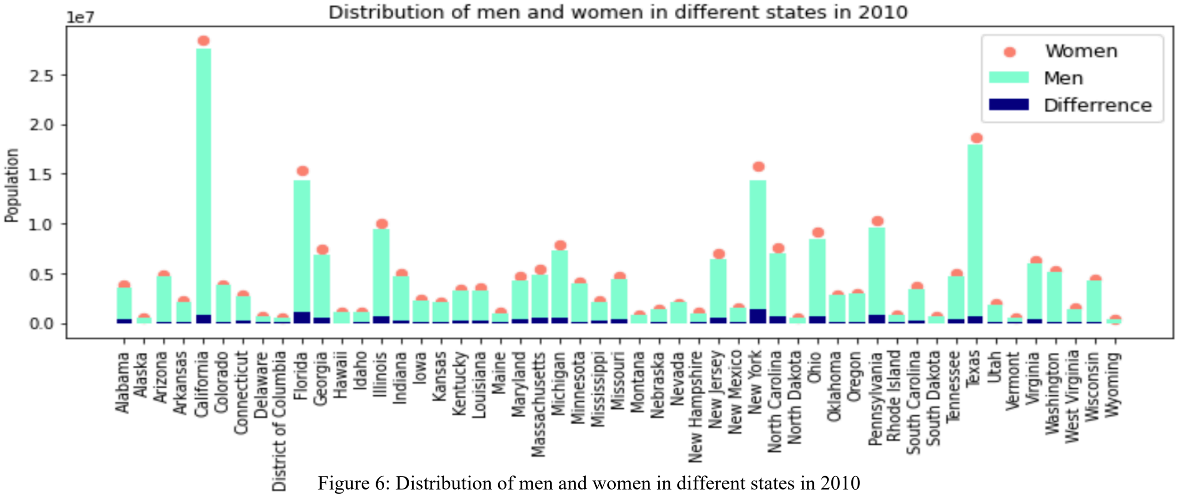

Figure 6: Distribution of men and women in different states in 2010

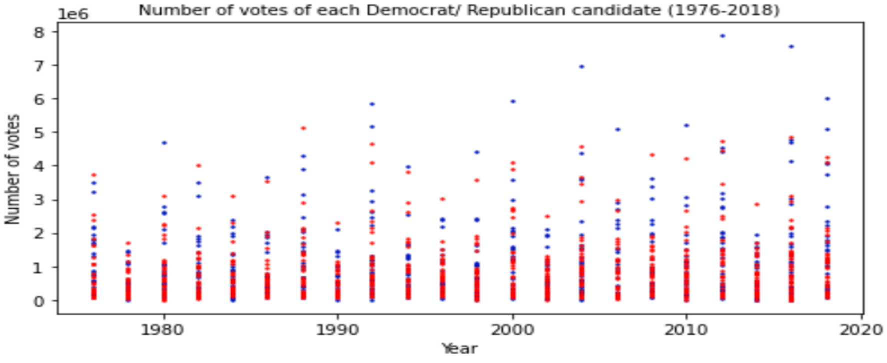

Figure 7: Number of votes of each Democrat/Republican candidate (1976-2018)

4. PREDICTION

The goal of this study is predicting the winner of each state based on the population of each race/gender category. After preprocessing and merging datasets from [1] and [3], I have a dataset which consists of the following columns “state” “race category 1”, …, “race category 12”, “Democrat votes”, “Republican votes” and “winner”. Winner column which is the dependent column here, is 1 when Democrat votes proceeds Republicans and -1 otherwise. Having the aforementioned classes -1 and +1 and using KNN, Support Vector Machine and Random Forest classification methods, the prediction of the winner is done. I have experimented with different test ratios and have iterated each experiment 5 times. Results are shown in figure 8. For test ratio=0.1, all classification methods yield an accuracy of 100%.

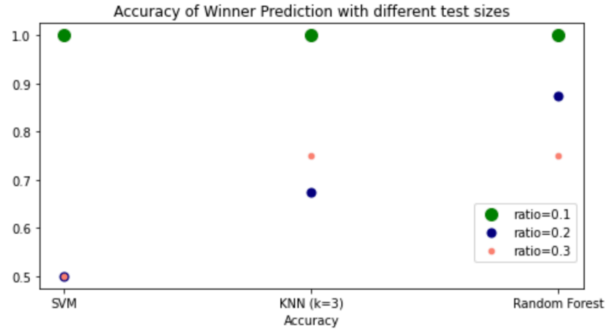

Figure 8: Winner Prediction accuracy with 3 test sizes via SVM, KNN and RF

5. CONCLUSION

Using the distribution of race/gender categories within the states, we can predict with a high accuracy that whether the winner of the Senate Election is a Republican candidate or Democratic. Thereby, it seems there’s a high correlation between race/gender diversities within a state and whom the state chooses for the Sensate.

6. FUTURE WORK

As mentioned before, the dataset is limited to a particular year and when US government census releases an updated dataset of racial distribution, we can predict the election result for the upcoming election.

7. REFERENCES

[1] Annual State Resident Population Estimates for 6 Race Groups dataset: https://www2.census.gov/programs-surveys/popest/tables/2010-2019/state/asrh/sc-est2019-alldata6.csv

[2] Annual State Resident Population Estimates for 6 Race Groups Description file: https://www2.census.gov/programs-surveys/popest/technical-documentation/file-layouts/2010-2019/sc-est2019-alldata6.pdf

[3] U.S. Senate votes 1976–2018 Version 4.0: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PEJ5QU

[4] USA government Website for vote rules: https://www.usa.gov/voter-registration-age-requirements

[5] Five Thirty Eight website: https://projects.fivethirtyeight.com/2016-swing-the-election/

[6] States information (FIPS, etc) for plotting map: https://raw.githubusercontent.com/plotly/datasets/master/2011_us_ag_exports.csv