Final Project

Predicting the 2020 U.S. Election

Introduction to Data Science (Fall 2020)

Janet Li, Philip Bell, Kacper Krasowiak

Group: 200

Workstream 2: 2020 U.S. Senate Elections

Janet Li

Objective

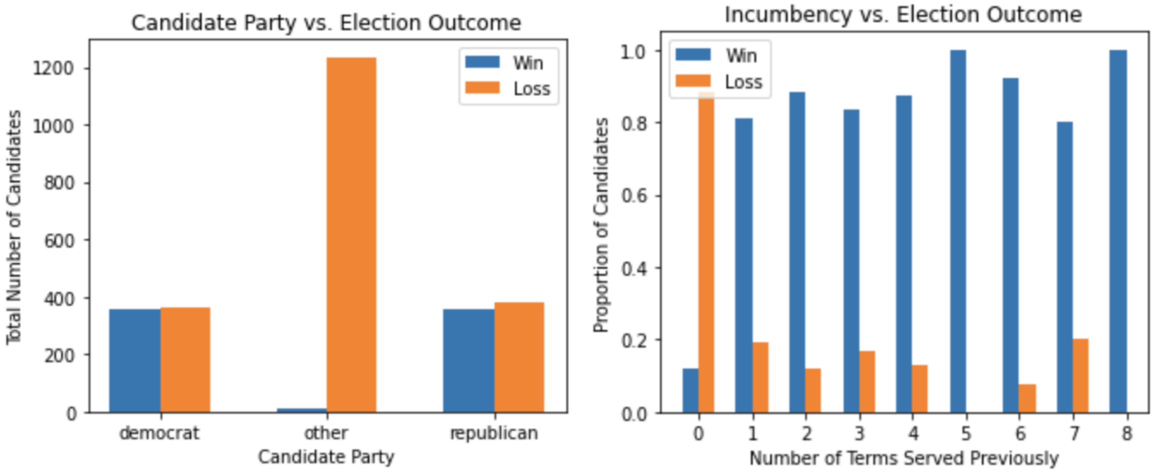
To predict which candidate will win each Senate election, and as a result, which political party will have majority control of the Senate.

Data Description

All data was recorded at the state-level. Historical Senate election results from 1976-2018 were downloaded from MIT Election Data Science Lab [8], and 2020 results were scraped from Wikipedia [9] (except the Georgia special elections, for which there will be a run-off in January 2021). The economic data from 1980-2020 was downloaded from the U.S. Bureau of Economic Analysis [10, 11] and U.S. Bureau of Labor Statistics [12]. The demographic data consisted of census data from 1980-2010, downloaded from the National Historical Geographic Information System [13], and the American Community Survey estimates from 2005-2019 [14], which was scraped using the U.S. Census Bureau API. We used linear regression to impute demographic information for non-survey and non-census years. Incumbency data was calculated from lists of former and current Senators on Wikipedia [15, 16], and voter turnout data from the 1980-2020 elections was obtained from the US Election Project [4].

The data was reformatted and combined into a single table, where the rows are each of the candidates for each state Senate election in each year, from 1980-2020, the columns are the predictors (election, economic, demographic, incumbency, and voter turnout information), and the outcome is a binary indicator, 1 if the candidate won the election. All numeric predictors were scaled, and categorical predictors were one-hot-encoded. From the full data set, the 1980-2018 election data was split 80-20 into training and validation sets, and the 2020 election data was reserved for the test set.

EDA and Feature Selection



Figures 4 and 5. Relationship of political party (left) and incumbency (right) with election outcome.

From Figure 4, we can see that third-party candidates win much less frequently than Democrat or Republican candidates, and from Figure 5, we can see that experienced candidates who have served at least one previous term as Senator are much more likely to be re-elected than new candidates. As a result, we would expect whether a candidate is a third-party candidate (“party_other”) and the number of terms a candidate has previously (“incumbent”) served to be strong predictors of the election outcome. Additionally, from graphs of the distributions of voter turnout and economic and demographic features in Appendix 2, we will select additional features that appear important in the graphs to train our model.

Model Selection

The modeling was split into 4 sections, each of which comprised a different category of models: (1) Logistic Regression, (2) Support Vector Machines Classifier, (3) k-Nearest Neighbors Classifier, and (4) Decision Trees, Random Forests and Boosting.

Each of these models was evaluated using the classification accuracies on the train and validation sets, as well as a classification accuracy on adjusted predictions for the validation set, which we will call the “adjusted accuracy.” In the standard model predictions, if a candidate’s predicted probability of victory is greater than 0.5, the candidate will be predicted to win the election. When calculating the adjusted validation accuracy, the predictions are adjusted so that the candidate with the highest predicted probability of victory in each state election will be predicted to win, and all other candidates in that election will be predicted to lose, even if their predicted probability of victory is greater than 0.5. This ensures that exactly one candidate wins each state’s election.

For the logistic regression models, we first fit 3 models: a baseline logistic regression model (without regularization), then added Lasso regularization and Ridge regularization. From the baseline model coefficients, we can see that the most important features are “party_other” and “incumbent” with coefficients of -4.09, and 1.32, respectively. In addition, in the logistic regression model with Lasso regularization, all but two of the coefficients were shrunk to 0: “party_other” and “incumbent” had coefficients of -1.016 and 0.624, respectively. This means that third-party candidates are predicted to have lower chances of winning the election, while incumbent candidates are predicted to have higher chances of winning the election, which was also reflected in the EDA.

To account for collinearity between predictors, we added interaction terms, as well as quadratic terms, then re-fit the 3 logistic regression models on the data with polynomial features: one model without regularization, one model with Lasso regularization, and one model with Ridge regularization. However, the best performing logistic regression model (highest adjusted validation score and highest validation score) was the logistic regression model trained on the original predictors, without polynomial features, with Lasso regularization, with an adjusted validation score of 0.8097, a slight increase compared to the baseline adjusted validation score of 0.8019.

The second group of models consisted of 6 support vector machines (SVM) classifiers, with linear, polynomial (degree 3), or radial basis function kernels and with or without polynomial features. A SVM model represents samples as points in space, and creates a line or hyperplane that separates the samples into classes, such that that samples in separate classes are as far apart as possible. This model can be used even when the samples are not linearly separable, by specifying different kernel functions (e.g. linear, polynomial, radial basis functions) [17]. SVM is also effective in high dimensional spaces, even when the number of features is greater than the number of samples, so it would be a good model to use on our training set with all quadratic and interaction terms added. We found that the best SVM classifier used a

RBF kernel and was fit on the original predictors, with parameters *C* = 10 and *gamma* = 0.01 determined by an exponentially-spaced grid search with 3-fold cross-validation. This model had the same train, validation, and adjusted validation scores as the best-performing logistic regression model, the logistic regression model with Lasso regularization.

For our third group of models, k-nearest neighbors classifiers, we found the 3-fold cross-validated classification accuracies of k-nearest neighbors classifiers, varying *k* from 1 to 100, as shown in Figure 6 below. We chose *k* = 35, to balance the train, validation, and adjusted validation accuracy scores. However, the validation and adjusted validation scores of the k-nearest neighbors model with *k* = 35 had lower train, validation, and adjusted validation scores compared to the best logistic regression and SVM classifier models, as well as the baseline model.

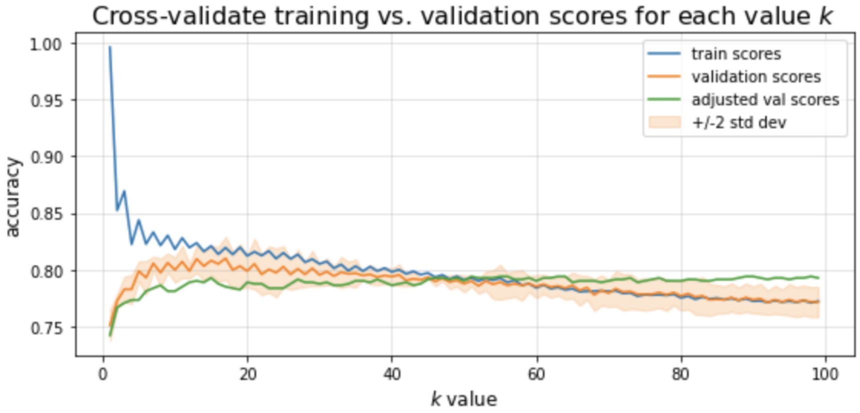
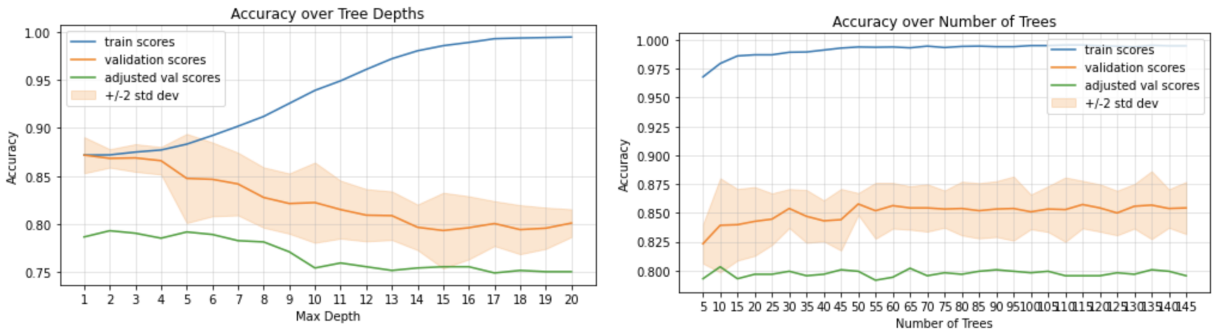


Figure 6. 3-fold cross-validation training, validation, and adjusted validation scores for each k-nearest neighbors model, varying *k* from 1 to 100

Finally, we trained 5 decision trees: a single decision tree and random forest classifiers with and without polynomial features, as well as a boosted decision tree model using AdaBoost. For the single decision trees, we used 3-fold cross-validation to determine the depth that maximized the adjusted validation score, as shown in Figure 7, and found the optimal maximum depth for both the decision tree trained on the original predictors and the decision tree trained on the polynomial features to be depth 2.



Figures 7 and 8. 3-fold cross-validation training, validation, and adjusted validation scores for each decision tree with maximum depths varying from 1 to 20 (left), and random forest with maximum depth 16 and number of trees varying from 5 to 150 (right), trained on the original predictors

Then, we fit random forest classifiers of maximum depth 16 (chosen to overfit the data, from the graph of tree depth vs. accuracy in Figure 7), varying the number of decision trees used in the model from 5 to 150. We found the optimal number of decision trees to be 10 when trained on the original predictors, as shown in Figure 8, and 30 when trained on the polynomial features. Both random forest classifiers had higher adjusted validation scores than the single decision tree classifiers. This is because each individual tree is overfit (so has low bias), and averaging many trees' predictions will reduce the variance. In addition, the random forest classifier de-correlates the trees, further improving performance.

Lastly, we fit a boosted decision tree using AdaBoost. We fit multiple AdaBoost classifiers for 800 iterations, varying the depth of the decision trees, as shown in Figure 9. We found that the model with depth 3 and 50 iterations seemed optimal, since that combination had very similar train and validation scores, indicating that the model has not overfit.

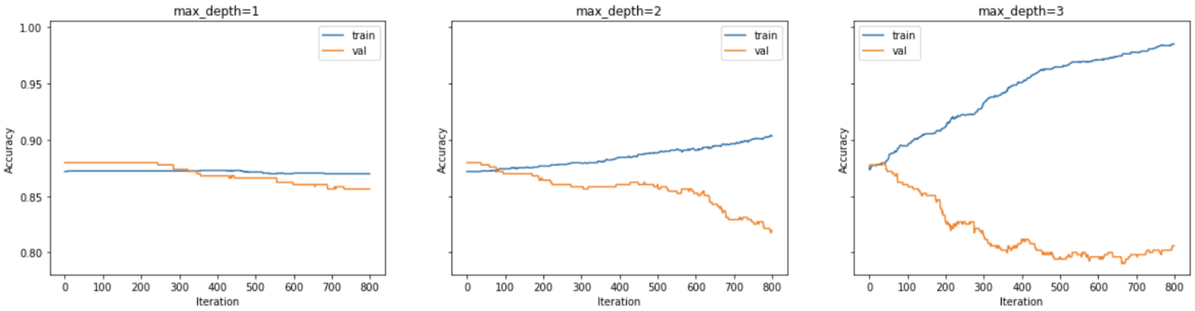


Figure 9. Accuracy of AdaBoost as training progresses, for trees of varying maximum depths

Of the 5 decision tree models, AdaBoost performed the best, with training, validation, and adjusted validation scores of 0.8825, 0.8718, and 0.8097, respectively. This is unsurprising because AdaBoost has the advantages of bagging, while also learning from misclassified samples in previous bootstrapped iterations.

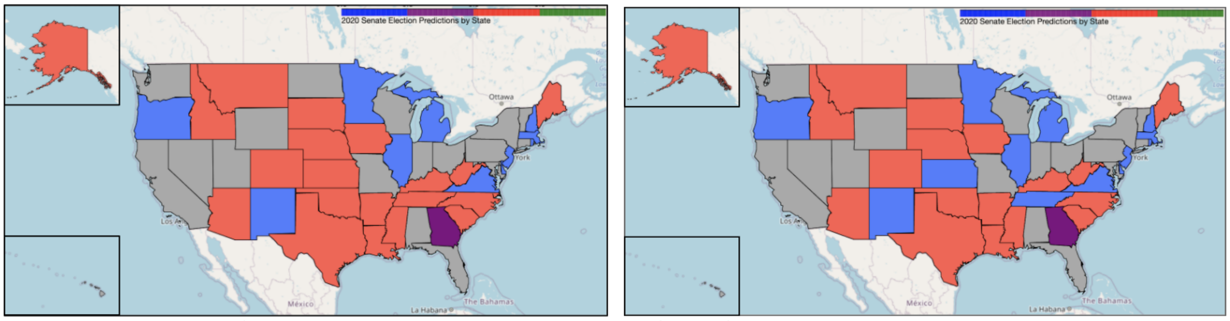
Results and Conclusions

Table 1. Comparison of best model in each of the 4 groups.

Model	Train	Validation	Adj. Validation
Logistic Regression (baseline)	0.8630	0.8524	0.8019
Logistic Regression (Lasso)	**0.8723**	**0.8796**	**0.8097**
SVM (RBF kernel)	**0.8723**	**0.8796**	**0.8097**
k-NN (k = 35)	0.8135	0.8136	0.7864
AdaBoost	0.8825	0.8718	0.8097

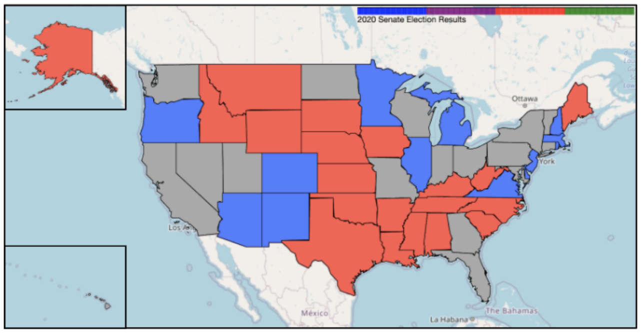
As shown in Table 1, the best models were the logistic regression with Lasso regularization and the SVM classifier with RBF kernel. Although both models had the same train, validation, and adjusted validation scores, the logistic regression model performed much better on the 2020 data, with an adjusted test accuracy of 0.9538, while the adjusted test accuracy of the SVM classifier was 0.9077.

As shown in the map in Figure 10, the logistic regression predicts that the Democrats will win 13 seats and the Republicans will win 22 seats, so the Republicans will maintain control of the Senate, with a seat distribution of 46 Democrat, 52 Republican, and 2 Independent. On the other hand, as shown in the map in Figure 11, the SVM model predicts that the Democrats will win 16 seats and the Republicans will win 19 seats, so the Democrats will just barely gain control of the Senate (since Independents usually caucus with the Democrats), with a seat distribution of 49 Democrat, 49 Republican, and 2 Independent.


2020 Senate Election Predictions by State
Ottawa
York
Los A
México
La Habana
The Bahamas
2020 Senate Election Predictions by State
Ottawa
York
Los A
México
La Habana
The Bahamas

Legend: blue: Democrat; red: Republican; green: Independent; purple 1 Democrat, 1 Republican

Figures 10 and 11. Map of the winning political party in each Senate race, as predicted by the logistic regression model (left) and the SVM classifier (right).



Legend: blue: Democrat; red: Republican; green: Independent; purple 1 Democrat, 1 Republican

Figure 12. Map of the true winning political party in each 2020 US Senate race

Limitations

As shown in Figures 10, 11, and 12, both the logistic regression and SVM classifier incorrectly predicted that Arizona and Colorado would be won by Republican candidates, when the Democrats were actually able to flip the Arizona and Colorado seats [9]. This suggests that our model may be overfitting on the incumbency information.

We can improve our models by incorporating an additional binary feature, indicating whether an incumbent Senator had previously been appointed through a special election (in which case, that candidate may be more likely to lose the election than other incumbents). For example, the incumbent Republican Arizona Senator Martha McSally had been elected through a special election in 2018 [9]. Ideally, we would also incorporate polling data, but we were unable to find enough historical polling data to use in our model (we were only able to find raw polling data from 2018 and 2020, from FiveThirtyEight [3]).

References

[1] FiveThirtyEight Election Forecast, 2020. https://projects.fivethirtyeight.com/2020-election-forecast/. Accessed November 3, 2020.

[2] Wikipedia contributors. (2020, December 12). 2020 United States elections. In *Wikipedia, The Free Encyclopedia*. Retrieved December 13, 2020, from https://en.wikipedia.org/w/index.php?title=2020_United_States_elections&oldid=993755354

[3] FiveThirtyEight Polling Database, 2020. https://projects.fivethirtyeight.com/polls/. Accessed November 3, 2020.

[4] McDonald, Michael P. 2020. "Voter Turnout." United States Elections Project. Accessed November 3, 2020, from http://www.electproject.org/home/voter-turnout/voter-turnout-data

[5] MIT Election Data and Science Lab, 2017, "U.S. President 1976–2016", https://doi.org/10.7910/DVN/42MVDX, Harvard Dataverse, V5, UNF:6:Mw0hOUHAijKPTVRAe5jJvg==

[6] "2020 Live Election Results", *New York Times,* https://github.com/favstats/USElection2020-NYT-Results/blob/master/data/2020-11-07%2014-15-14/results_president.csv. Accessed November 7, 2020.

[7] Hatley, N. and Kennedy, C. 'A Resource for State Preelection Polling', 2020, *Pew Research Centre.*

[8] MIT Election Data and Science Lab, 2017, "U.S. Senate 1976–2018", https://doi.org/10.7910/DVN/PEJ5QU, Harvard Dataverse, V4, UNF:6:WzSZLQX8O9Nk6RKWwkjx9g==

[9] Wikipedia contributors. (2020, December 13). 2020 United States Senate elections. In *Wikipedia, The Free Encyclopedia.* Retrieved November 18, 2020, from https://en.wikipedia.org/w/index.php?title=2020_United_States_Senate_elections&oldid=993938893

[10] U.S. Bureau of Economic Analysis, 2019, "Annual State Personal Income and Employment." Accessed November 18, 2020, from https://apps.bea.gov/regional/docs/DataAvailability.cfm

[11] U.S. Bureau of Economic Analysis, 2019, "Annual Gross Domestic Product by State." Accessed November 18, 2020, from https://apps.bea.gov/regional/docs/DataAvailability.cfm

[12] U.S. Bureau of Labor Statistics, 2020, "Local Area Unemployment Statistics." Accessed November 18, 2020, from https://beta.bls.gov/dataQuery/find?fq=survey:%5Bla%5D&s=popularity:D

[13] Steven Manson, Jonathan Schroeder, David Van Riper, Tracy Kugler, and Steven Ruggles. IPUMS National Historical Geographic Information System: Version 15.0. Minneapolis, MN: IPUMS. 2020. http://doi.org/10.18128/D050.V15.0

[14] U.S. Census Bureau. "American Community Survey 1-Year Data (2005-2019)." *The United States Census Bureau*, www.census.gov/data/developers/data-sets/acs-1year.html. Accessed October 15, 2020.

[15] Wikipedia contributors. (2020, December 13). List of former United States senators. In *Wikipedia, The Free Encyclopedia*. Retrieved November 18, 2020, from https://en.wikipedia.org/w/index.php?title=List_of_former_United_States_senators&oldid=993991258

[16] Wikipedia contributors. (2020, December 7). List of current United States senators. In *Wikipedia, The Free Encyclopedia*. Retrieved November 18, 2020, from https://en.wikipedia.org/w/index.php?title=List_of_current_United_States_senators&oldid=992916830

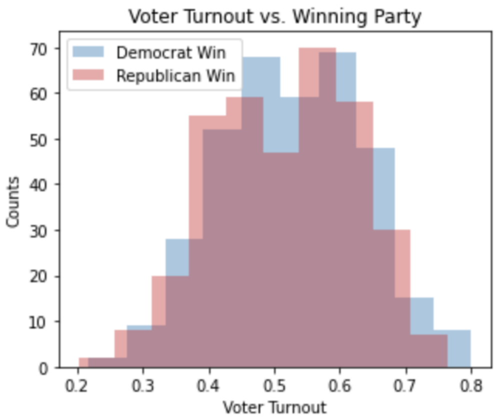
[17] Berwick, R. “An Idiot’s guide to Support vector machines (SVMs).” MIT 6.034, November 10, 2011, MIT. PDF.

[18] MIT Election Data and Science Lab, 2017, "U.S. House 1976–2018", https://doi.org/10.7910/DVN/IG0UN2, Harvard Dataverse, V8, UNF:6:p05gglERZ/Fe5LP4RarxeA==

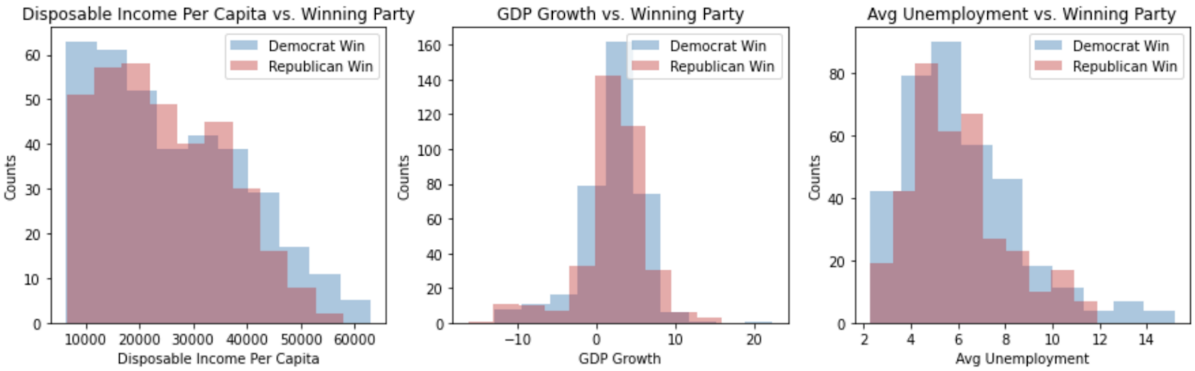
[19] Wikipedia contributors. (2020, December 13). 2020 United States House of Representatives elections. In *Wikipedia, The Free Encyclopedia*. Retrieved 00:12, December 14, 2020, from https://en.wikipedia.org/w/index.php?title=2020_United_States_House_of_Representatives_elections&oldid=993939371

[20] Lichtman, Allan. “The Keys to the White House: Forecast for 2020 · 2.4.” *Harvard Data Science Review*, PubPub, 27 Oct. 2020, hdsr.mitpress.mit.edu/pub/xhgpcyoa/release/2.

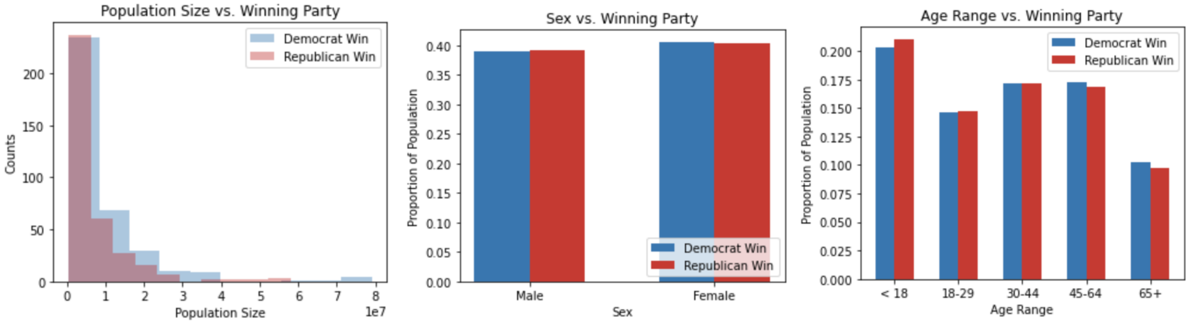
Appendix 2: Comparisons of distributions of economic and demographic features and voter turnout when Democrats win vs. when Republicans win



In the above graph of voter turnout, we can see that voter turnout is slightly lower in states when Republicans win than when Democrats win.



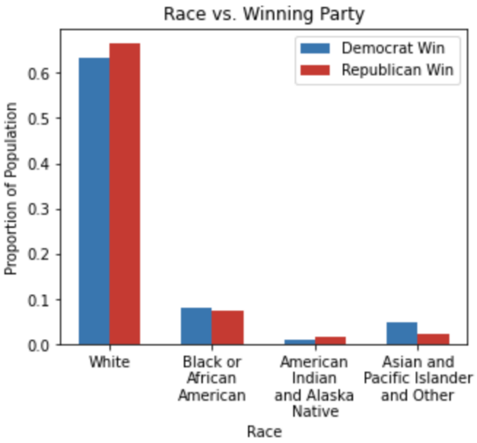
In the above graph of economic features, we can see that disposable income per capita and average unemployment are both usually higher in states when Democrats win the election, so they may be good predictors of the election outcome. On the other hand, the distribution of GDP growth is similar between Democrat wins and Republican wins, so we do not expect GDP growth to be a strong predictor, and will not use it as a predictor in our model.

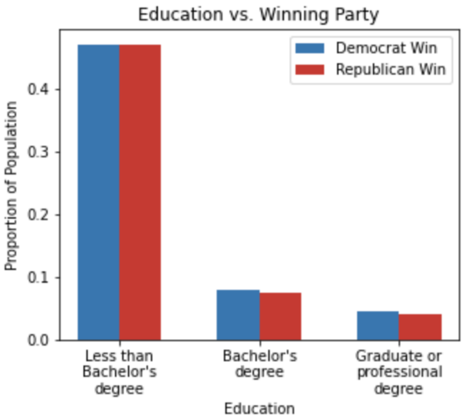


From the above graphs, left and center, of demographic features, we can see that Democrats tend to win in states with larger populations, but the proportion of males and females in the state's population when

Democrats win is very similar to the proportion when Republicans win, so we will use population as a predictor, but not the proportion of males or females.

In the graph above, on the right, we can see that the proportion of the population that is under 18 years is lower, while the proportion of the population that is over 45 years old is higher in states when Democrats win the election. As a result, we will use those two age ranges as predictors in our model.





From the graph above, on the left, the proportion of the population who identify as White, American Indian, or Alaska Native, is higher in states when Republicans win, while the proportion of the population who identify as Black, Asian, Pacific Islander, or any other race is higher in states when Democrats win, so we will use the proportion of all four racial categories as predictors.

From the graph above, on the right, the proportion of the population with at least a Bachelor's degree is higher in states when Democrats win, while the proportion of the population with less than a Bachelor's degree is higher in states when Republicans win, so we will use the proportion of the population with at least a Bachelor’s degree as a predictor.