Predicting the 2020 Election

Broad Impact Statement

Our research questions are as follows: Does accounting for various demographic, legal, and behavioral variables allow a model to predict the 2020 presidential and federal House of Representative election with higher accuracy than models did the 2016 election? What factors are most valuable in predicting election results?

A model with a high level of predictive power may potentially have far-reaching implications across the United States. Each governing party will favor the success of different industries, meaning that knowing who will be elected next may improve market stability and reduce economic uncertainty for investors. Furthermore, such predictions can significantly improve political strategy and decision-making, especially regarding allocation of funds; any party that has access to such a model will have an advantage over their peers who do not. Individual civilians with access to predictions may also use them to decide whether they should buy or sell assets, or even migrate, before a potential unfavorable change in government or policy. Lastly, such models may help illuminate to a political party why they lost the election and smoothen the transition of power.

Unfortunately, powerful forecasting of the future can be exploited. Governments who are aware of an incoming loss may rush out more radical policies in preparation for being ousted from office, or even attempt to change the constitution to bend the election in their favor. Beyond this, political strategy may grow increasingly predatory as an organization learns more about voters, with political groups potentially attempting to suppress specific groups' abilities to vote.

It is also important in general to carry out machine learning research carefully to ensure as accurate results as possible. For example, it is essential to ensure that the demographic data being trained on is complete; a model that only accounts for certain racial classes would likely result in unfair predictions. It is also critical to ensure that a model does not capture spontaneous features which may drive predictions in an incorrect direction. For example, during the 2020 election year, not only did COVID-19 permeate throughout the United States, but there were also many incidences of police brutality and a resulting distrust of the government. These spontaneous factors were not included in our model and could have caused it to give poor predictions, especially on the swing states.

Overall, our model did a good job despite uncertainties of uncovering some important features for one of the most important political events in the United States. As critical thinkers, we may be able to learn from it, and think about future elections from the perspectives of demographics, economics, financials, voting resources, and behaviors.

Data Description

Outcome variable for the presidential election

We aim to predict the ratio of Democratic party votes over Republican party votes at the county level in equation (a):

$$Predicted\ vote\ ratio\ per\ county\ = \frac{Votes\ for\ Democrats\ per\ county}{Votes\ for\ Republicans\ per\ county} \qquad \text{(a)}$$

From the predicted voting ratio per county, we convert decimals to factors to derive the percentage of votes for Democrats and votes for Republicans per county in equation (b):

$$Predicted\ \%\ votes\ for\ Democrats/Republicans\ per\ county = \frac{Numerator\ of\ fraction\ (Predicted\ vote\ ratio\ per\ county)}{Numerator\ of\ fraction\ (Predicted\ vote\ ratio\ per\ county) + Denominator\ of\ fraction\ (Predicted\ vote\ ratio\ per\ county)} \qquad \text{(b)}$$

By multiplying percentage votes for Democrats and Republicans by county population and summing up votes for over all counties in a state, we predict the number of votes for each party per state in equation (c):

$$Predicted\ votes\ for\ Democrats/Republicans\ per\ state\ =$$
$$\sum_{i=1}^{N} Predicted\ \%\ votes\ for\ Democrats/Republicans\ per\ county\ in\ State_k \times County\ population\ in\ State_k$$

for i = 1,...,N (number of counties per State) (c)

To make the final election prediction, we calculate the percentage of Democrats and Republicans votes per state. By the winner-take-all system, we multiply electoral votes per state by indicator function I (Predicted % votes for party per state >50). If the percentage vote in a state is higher than 50% for a certain party, they are assigned all of its electoral votes.

$$Electoral\ vote\ for\ Democrats/Republicans =$$
$$\sum_{i=1}^{K} I(Predicted\ \%\ votes\ for\ Democrats/Republicans\ per\ State_i > 50) \times Electoral\ college\ per\ State_i$$ for i= 1,...K=51

(number of States) (d)

In summary, our presidential predicted model enables three levels of prediction: 1) County level votes per party 2) State level votes per party and 3) overall electoral votes. We believe that using the county-level votes allows us to have a detailed understanding of the underlying demography, socioeconomic factors, and regulations that drive presidential election results. Our model is trained on 2016 presidential election results by county retrieved from opendatasoft (https://public.opendatasoft.com/explore/dataset/usa-2016-presidential-election-by-county).

Outcome Variable for the House of Representatives Election

We aim to predict the percentage of votes to the Democratic party, Republican party, and other parties for the house of representatives per county. We summarize precinct-level votes to county-level, using 2016 precinct-level data from the MIT election lab (https://electionlab.mit.edu/data). Each predicted percentage of votes per county is multiplied by county population to estimate votes for Democrats, Republicans, and other parties per county. Summing up all estimated votes per county for each party within a state yields the total votes for each party per state. See equation (e).

$$Predicted\ votes\ for\ Democrats/Republicans/Other\ per\ state\ =$$
$$\sum_{i=1}^{N} Predicted\ \%\ votes\ for\ Democrats/Republicans/Other\ per\ county\ in\ State_k \times County\ population\ in\ State_k$$

for i = 1,...,N (number of counties per State) (e)

For each state, we multiply the percentage of votes for each party by the allocated number of house of representatives per state to yield the final prediction of number of seats in the house of representatives for each party. See equation (f).

$$\#\ Seats\ for\ Democrats/Republicans/Other =$$
$$\sum_{i=1}^{K} Predicted\ \%\ votes\ for\ Democrats/Republicans/Other\ per\ State_i \times \#\ House\ seats\ allocated\ per\ State_i$$

for i = 1,...K = 51 (number of States) (f)

Predictor Variable: Voting Consistency (source: open source opendatasoft)

Different states often have the tendency to vote for a particular political party which is likely to carry into future elections, making voting consistency important to consider when predicting 2020 election results. As the 2020 election had not yet finished during the time when this model was trained we decided to use 2012/2008 and 2016/2012 as our training and testing set. To approximate county-level data, we calculated the Democratic and Republican voting percentage ratio as a continuous variable. If the ratio is smaller than 1, it means there is a decreased voting consistency for that political party from 2008 to 2012. If the ratio is greater than 1, then that particular party has retained its voting habit for those years.

Predictor Variable: Base demographics

We included 5 categories of demographic predictors, using 2016 data for our training set and 2018-2019 data for our test set. Numeric values were converted to a percentage of the county population to make different sized counties comparable.

1. Race, ethnicity, gender (source: County Population by Characteristics: 2010-2019):
    a. 10 continuous numeric variables: total male, total female, White American male/female, Black or African American male/female, Black or African American male/female, American Indian and Alaska Native male/female, and Native Hawaiian and Other Pacific Islander male/female
2. Education status (source: Economic Research Service- Educational attainment for the U.S., 1970-2018):
    a. 5 continuous numeric variables describing education status per county: % of population with high school diploma, at least bachelor's degree, graduate degree, and school enrollment.
3. Poverty (source: same as race, ethnicity, gender category):
    a. 2 continuous numeric variables: estimated population of poverty and median household income in dollars.
4. Insurance status (source: 2008 - 2018 Small Area Health Insurance Estimates using the American Community):
    a. 2 continuous numeric variables per county: numbers of uninsured, numbers of insured per county
5. County business pattern (source: County Business Patterns: 2016 and 2018):
    a. 19 Continuous numeric variables per county for total numbers of establishments in the following business categories: undefined, Administrative, Agriculture, Construction, Educational Services, Finance, Food services, Health care, Information, Management, Manufacturing, Mining, Other, Professional Service, Real Estate, Recreation, Retail, Transportation, Utilities, and Wholesale.

Predictor Variable: Financial data (source: FEC https://www.fec.gov/data/browse-data/?tab=candidates)

The amount of money a candidate spends on their campaign may predict their likelihood of winning, since candidates that spend more can reach more voters. Because there is no state or county level data of presidential spending for the election, we will be using the total spending within each state by house representatives as an approximation. Specifically, we are measuring the amount each candidate received and spent during their campaign, and what they had left afterwards.

Predictor Variable: Polling Data (source: FiveThirtyEight 2020 https://projects.fivethirtyeight.com/polls/; 2016 https://projects.fivethirtyeight.com/2016-election-forecast/)

Although polling data has had difficulty predicting the elections in the past, we believe that the polls potentially still have valuable information and decided to include it in our model. We used the mean to assess the average of results for each party within each state, the median to control for outliers, and we want to take into consideration the uncertainty of the polls by adding in the standard deviation. We find the results at the state level and give each county within that state these values.

Predictor Variable: Voting accessibility

Accessibility to voting, or level of voter suppression, could help predict the election. Populations that are unable to vote in certain states tend to have shared demographic or socioeconomic traits, and thus be more likely to vote for a certain candidate. This inability to vote may result from restrictions placed on the degree of identification necessary to cast a vote, or how complex the process is to send in a mail-in ballot. To represent these predictors, we used the following data:

1. Laws for Voter ID, Voter Registration, Felony Disenfranchisement, and Early Voting (source: The Guardian)
    a. Each law is represented with an ordinal categorical variable coded to represent strictness
2. Rates of early votes (source: United States Elections Project)
    a. This data is only available on the state level, so it is estimated by calculating the percentage of the population that voted early at a state level.

Predictor Variable: Economic uncertainty (source: U.S. Bureau of Labor Statistics).

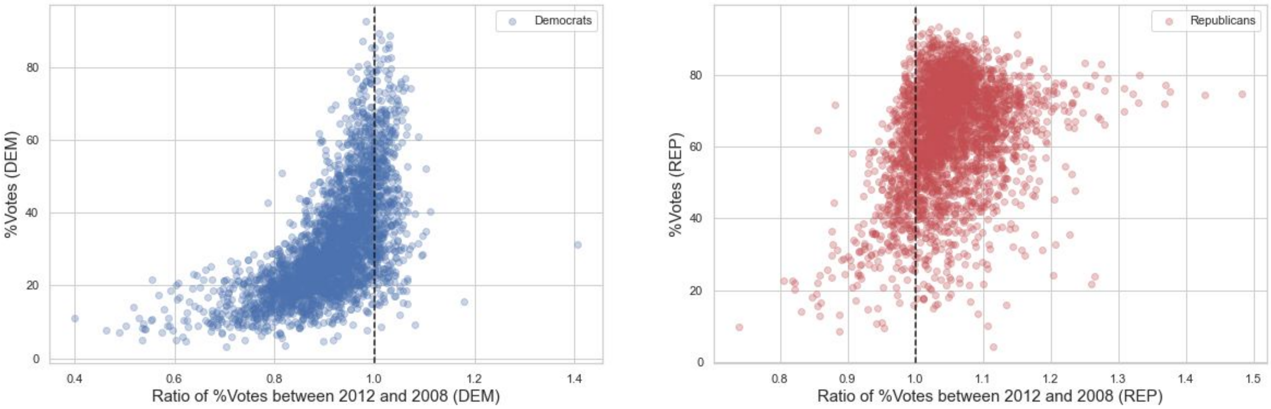
Economic uncertainty could impose a big influence on presidential and federal house representative elections. Here, in order to delineate the relationship between economic uncertainty and election results, we decided to look into the unemployment rate (continuous variable). The unemployment rate data of 2016 will be used for the training set and unemployment rate of 2019-2020 will be used for the testing set.

Visualizations and findings from EDA

The most interesting plots are shown here. The visuals for the remaining figures can be accessed in our Jupyter notebook.

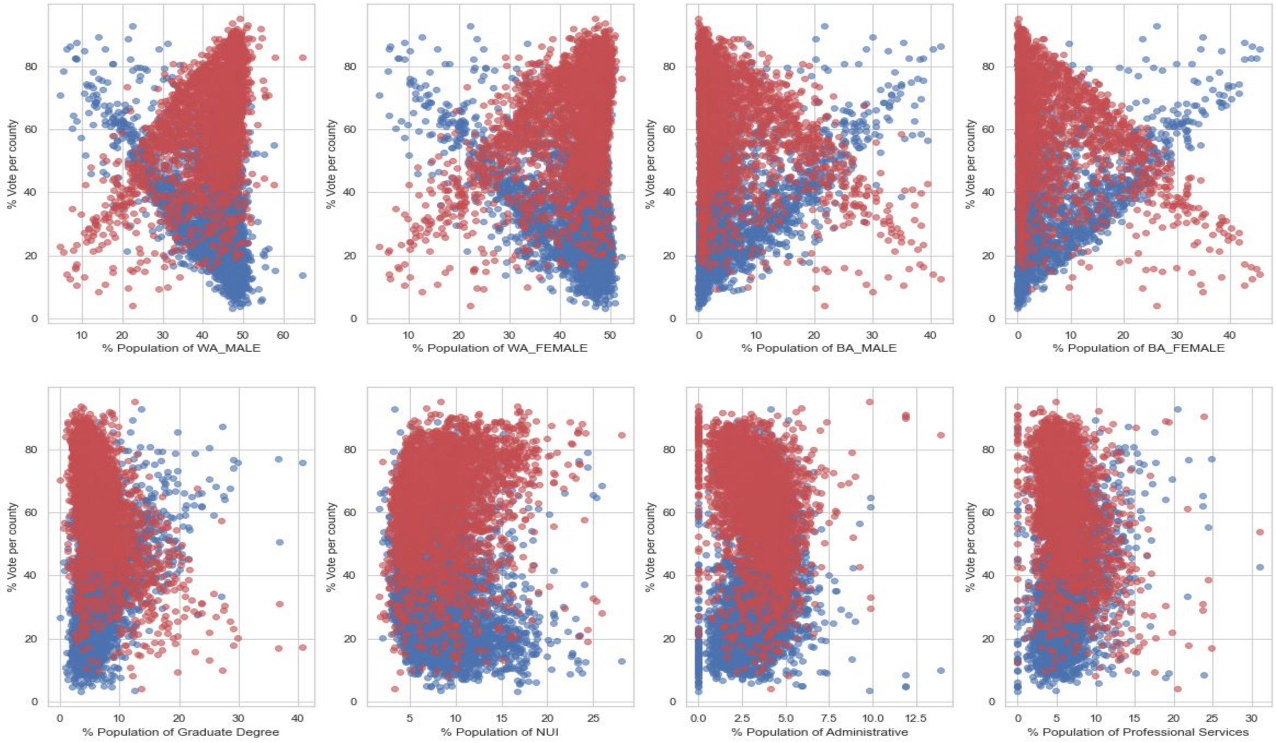
Scatter Plot of % Votes Ratio versus Voting Consistency

The majority of Republican counties were persistent with their voting habits from 2012 to 2016 as most of the ratios fell above 1, while a large proportion of the Democratic counties shifted their voting behavior to Republican parties in 2016. This is in line with the fact that Republican candidate won the presidential election in 2016 (reference figure 4 in the notebook).



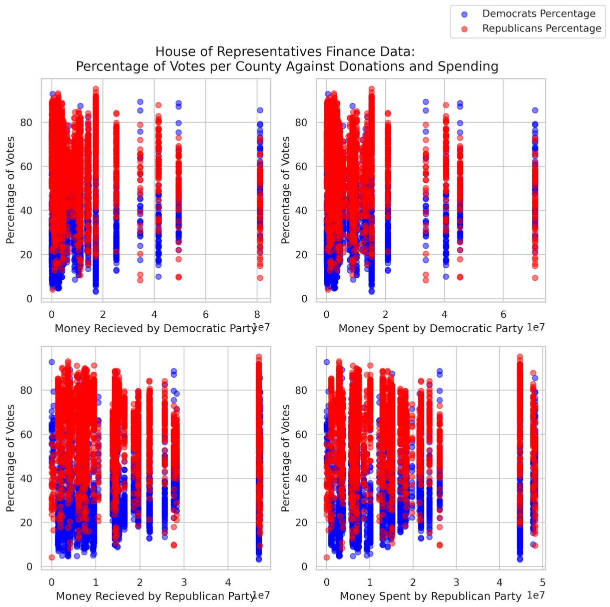
Base Demographics and Business

The scatter plots of selected demographic features demonstrate how the percentage of different characteristics determined the turnout for each party in the 2016 election. The percentage of White Americans in each county highly predicted the percentage vote for Republicans, while the Black American population per county correlated to the Democrat vote. Counties with more individuals with graduate degrees also seem to favor voting Democrat. The percentage of people without insurance (NUI) is low among all counties, and these people seem to be equally likely to vote for each party. Lastly, more administrative and professional services in a county correlates with a Democratic leaning in that county (reference figure 7 in the notebook).



Financial and Polling Data

There is no clear trend for any financial predictors on each of the different voting response variables [reference figure 10 in the notebook]. The polling data consists of the average, median, and standard deviation of the polls in each state. We can see that as the average, median, and standard deviation of the polls increase for a party, the higher the percentage that party wins counties (reference figure 11 in the notebook).



House of Representatives Finance Data:
Percentage of Votes per County Against Donations and Spending

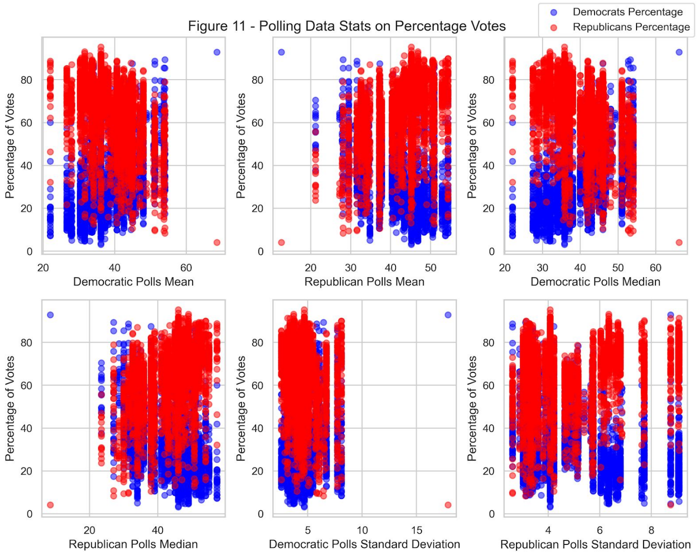
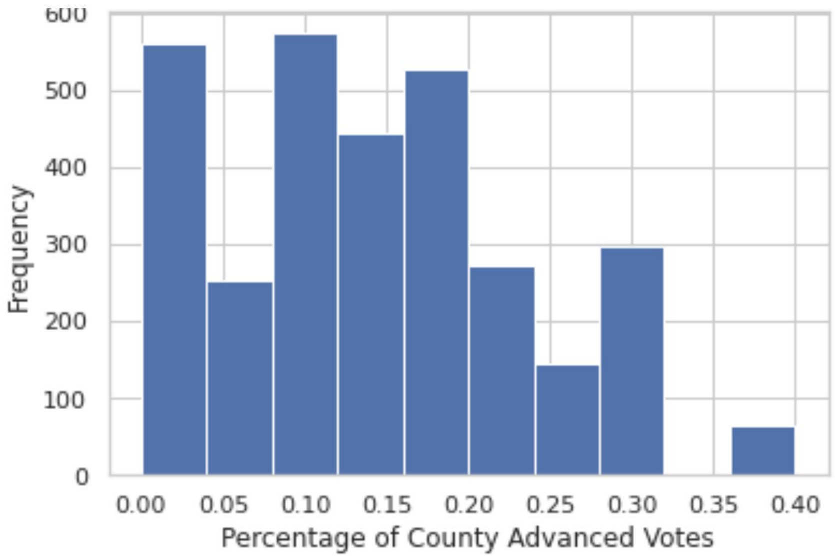


Figure 11 - Polling Data Stats on Percentage Votes

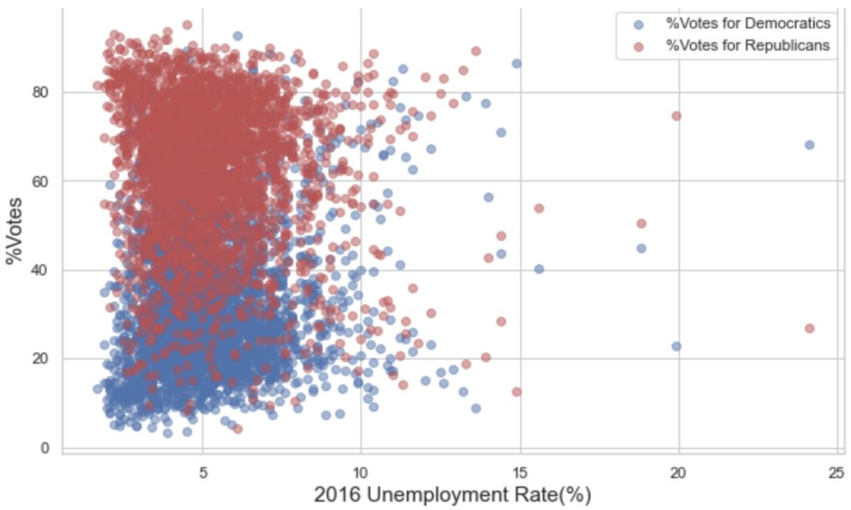
Histogram of County Advanced Voting Frequencies in 2016

While there does not seem to be a clear distribution, the majority of states in 2016 had 20% or less of their population vote in advance, either through mail or through early in-person voting (reference figure 15 in the notebook).



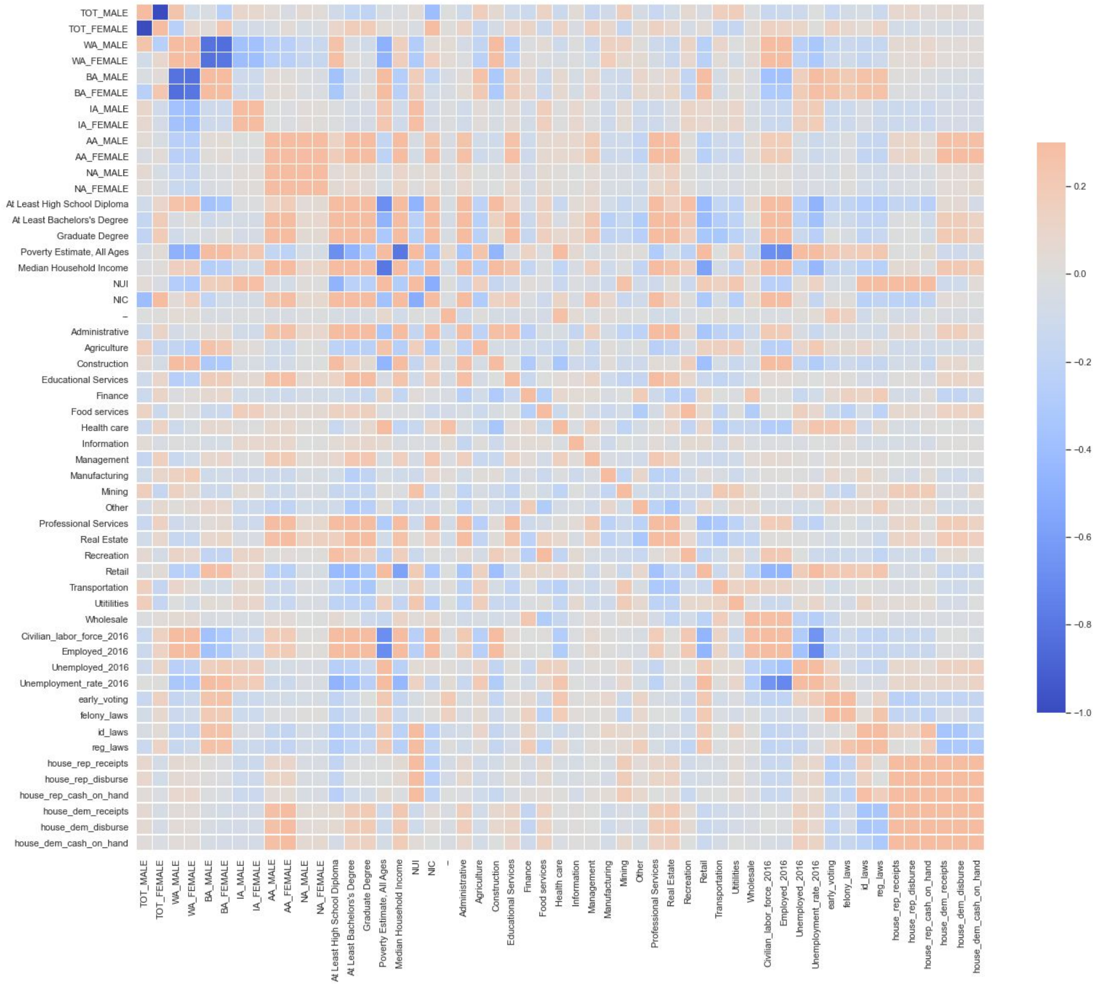
Scatter Plot of Unemployment Rate (%) vs %Votes

There is no observable relationship between percentage votes and unemployment rate. Overall, there are just higher percentage votes for Republican Party in 2016 (reference figure 13 in the notebook).



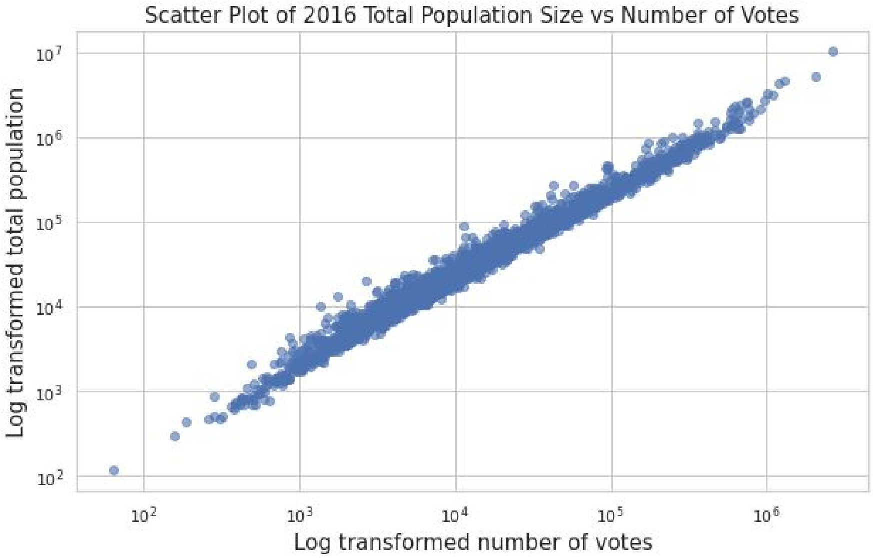
Heatmap of All Features

The range of correlations between our features is low to moderate, with Pearson's correlation scores ranging from 0.2 to -1, suggesting that our model can benefit from regularization (reference figure 16 in the notebook).



Assumption Plot

When this model was built, we did not have the access to the number of people who actually voted in the 2020 election. Therefore, we used the overall population size as a proxy for the number of voters for both 2016 and 2020. Pearson's correlation between the number of votes and total population in 2016 yields a high value of 0.972, indicating that it is safe to assume the total population in a county can be used as a proxy for the number of votes in an election (reference figure 17 in the notebook).

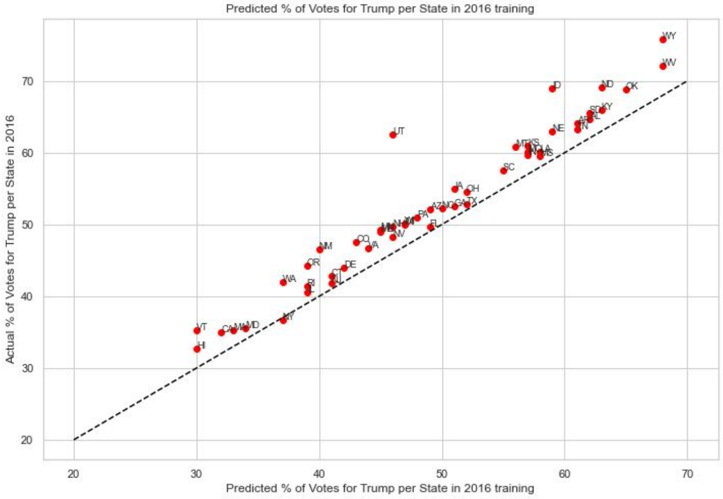


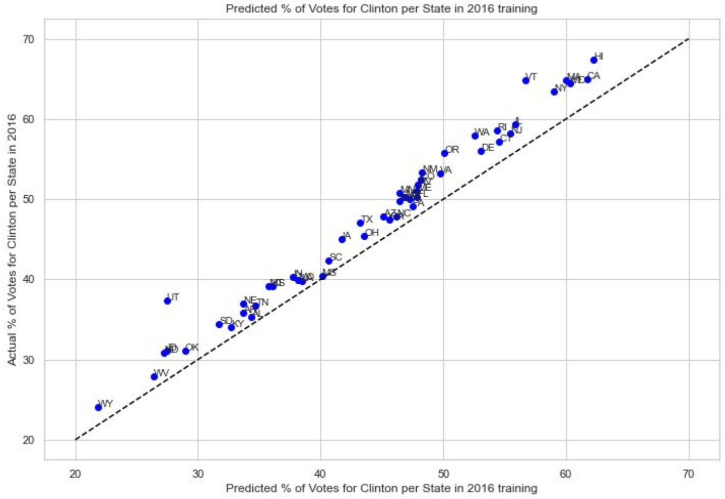
Presidential Election Model

Random Forest

Given the complexity of the model and promising model evaluation of the tree-based regression models in our baseline model construction (with best validation R squared = 0.96), we decided to build our presidential prediction model with random forest to decorrelate features and to reduce overfitting. We tuned hyperparameters of the random forest model by the randomized random forest cross validation method (sklearn.RandomizedSearchCV). This function performs similarly to a grid search, both of which optimize the validation score by a range of given parameters. The differences between these two methods are computational time, as cross-validated random search only tries a fixed number of parameters settings sampled from specified distributions over numbers of iterations. Documentations have shown that both randomized search and grid search perform similarly and with randomized search being more computationally efficient.

We performed a randomized search over 100 iterations and 3-fold cross validation for four parameters: number of estimators for the base random forest regressor model, maximum features at each node, maximum depth for each tree, and a boolean parameter for bootstrapping to control if a subsample is selected to build each tree. Fitting 3 folds for each candidate, we maximized the validation R squared at 0.73 with the following parameters: 800 estimators, 15 maximum features, a maximal depth of 24, and no bootstrap to fit each tree. The model is fitted on a standardized training set. The training score for this best model is 0.99 for the county level votes ratio. The R squared for the training set in terms of percentage votes for Republicans and Democrats per state is 82.5% and 87.2%, respectively. The scatter plots below illustrate how positively correlated our predicted percentage vote on a state values are to the 2016 values.





We then predicted the 2020 presidential election predicted votes ratio per county (equation a) by the tuned random forest regressor. We utilized 2018-2019 values to estimate test-set features corresponding to features in the training set for the 2016 presidential election. Testing set is first standardized from the standard scaler fitted on the training set.

Results and Discussion

To interpret the model, we investigated both feature and permutation importance. The feature importance (left) of the random forest regression model evaluates how removing each feature decreases the explained variance in the model. The top features are the percentages of White American females and males per county, followed by Democrat and Republican voting consistency. The fifth most important feature is the percentage of the population with a graduate degree. Percentages of Asian and Black American women per county also show high weights in the model. Education level and percentage of education services also show fairly high influence in our predictive model. In comparison, permutation importance (right) demonstrates high concordance for overall rank of feature importances, with more predictive weight put on the consistency of voting behavior for the Democratic party.

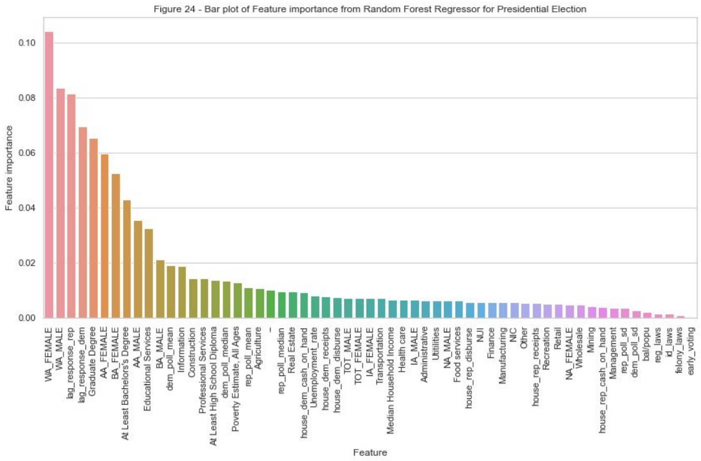
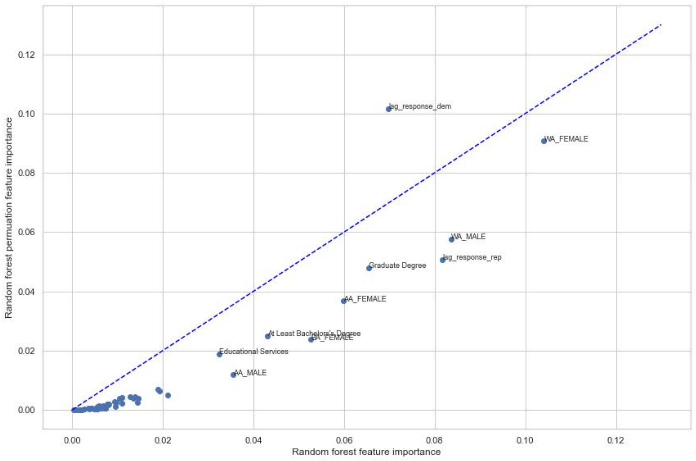
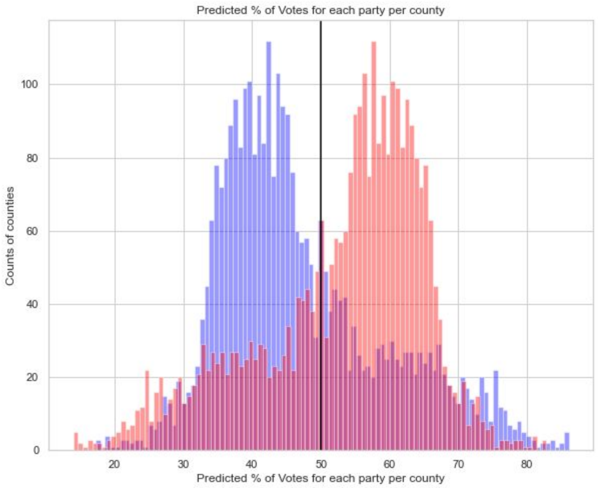
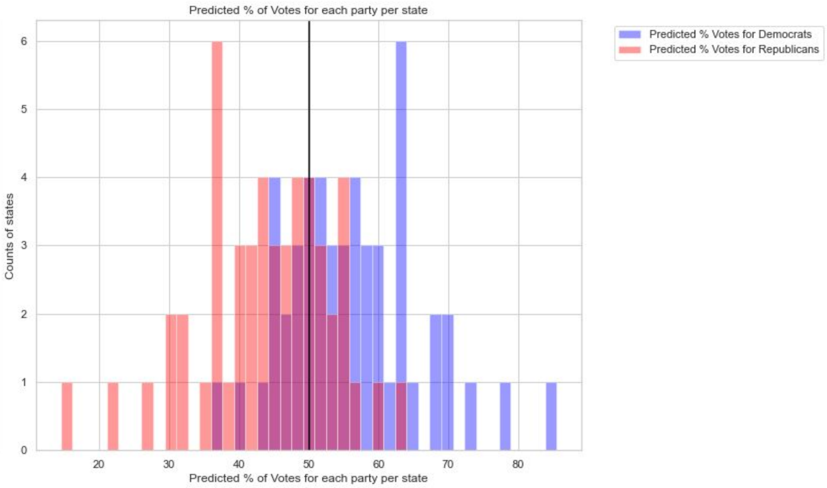


Figure 24 - Bar plot of Feature importance from Random Forest Regressor for Presidential Election

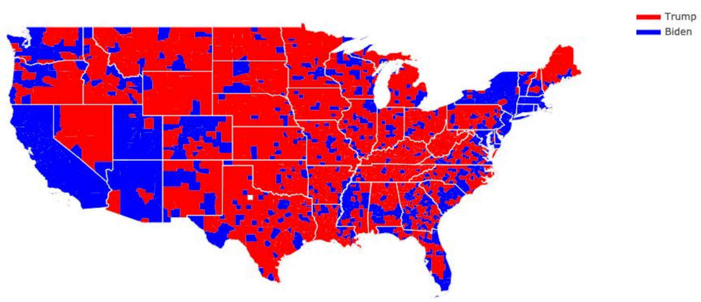


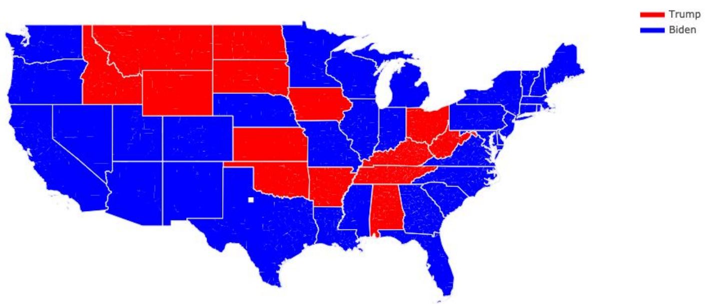
The predicted percentage of votes for Republicans and Democrats at a county level shows that more counties favored Trump over Biden than vice versa (left). However, after calculating the electoral college votes for each candidate (equations (a) to (d)), it is revealed that Biden won the majority of votes (right). This result indicates that smaller counties are predicted to heavily favor Trump while larger counties are predicted to have moderate support for Biden.



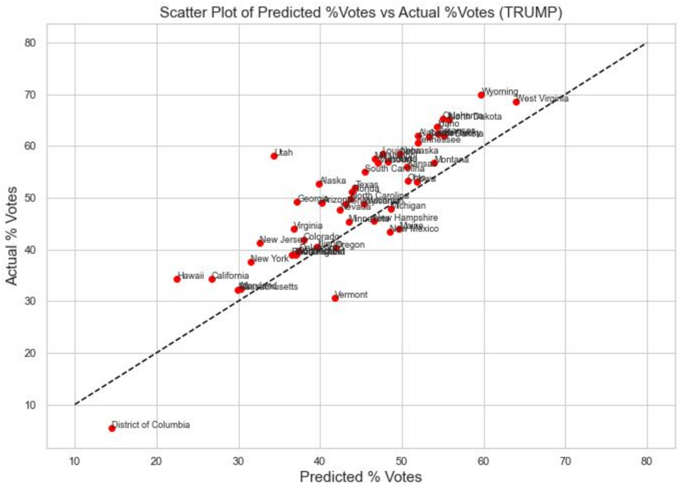


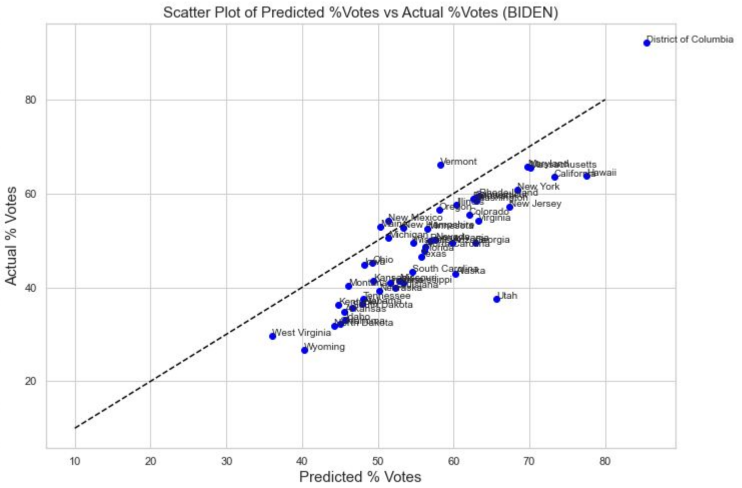
These results are presented on the maps below at a county level on the left and a state level on the right. If a county or state favors Trump, it is colored red; if it favors Biden, it is colored blue. The mapping of voting prediction per state shows that we predict Alabama, Arkansas, Idaho, Iowa, Kentucky, Kansas, Montana, North Dakoda, Ohio, Oklahoma, South Dakoda, Tennessee, West Virginia, and Wyoming to vote for Trump with 122 electoral votes. The rest of the states are predicted to vote for Biden with 416 electoral votes.





After we obtained the percentage votes for each party by each state, we then evaluated our model with the 2020 election results. We calculated the R-squared score for each party, finding relatively low scores of 54.8% for Democrats (right) and 35.2% for Republicans (left). This suggests that our model is flawed and not very effective. As the plots below demonstrate, our model underpredicts the percentage of votes for Republicans and overpredicts that of the Democrats.





House of Representative Election Model

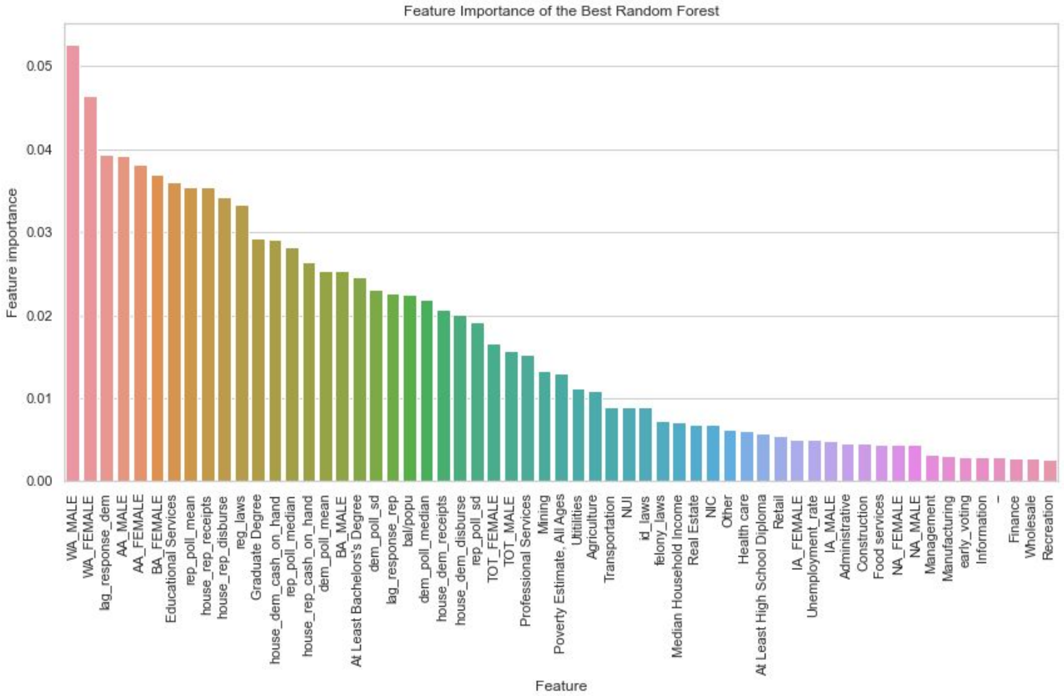
Method 1: Random Forest

Given the initial success we had with a baseline decision tree model, we decided to fit a random forest to the training data and predict on the test data. Notably, candidates who are neither from the Republican or Democrat party have an impact in House of Representatives elections. Therefore, we could not run a random forest with a regression outcome because we could not develop a way to consolidate the votes for three different groups into one regression outcome in the way we combined Democrat and Republican votes for the Presidential election. Therefore, we fit a Random Forest Classifier that predicted the winner of each county with three outcomes: 0 for Republican, 1 for Democrat, and 2 for Other. After this, we predicted who would win each house seat at a district level and worked with a “winner takes all” model such that, if a candidate won a county, they would earn all of the votes from that county. Such a methodology is not perfect as it does not accurately represent the number of votes that the model would predict goes to each political party in each district, but it allowed us to use the same county-level dataset that we used for the presidential election.

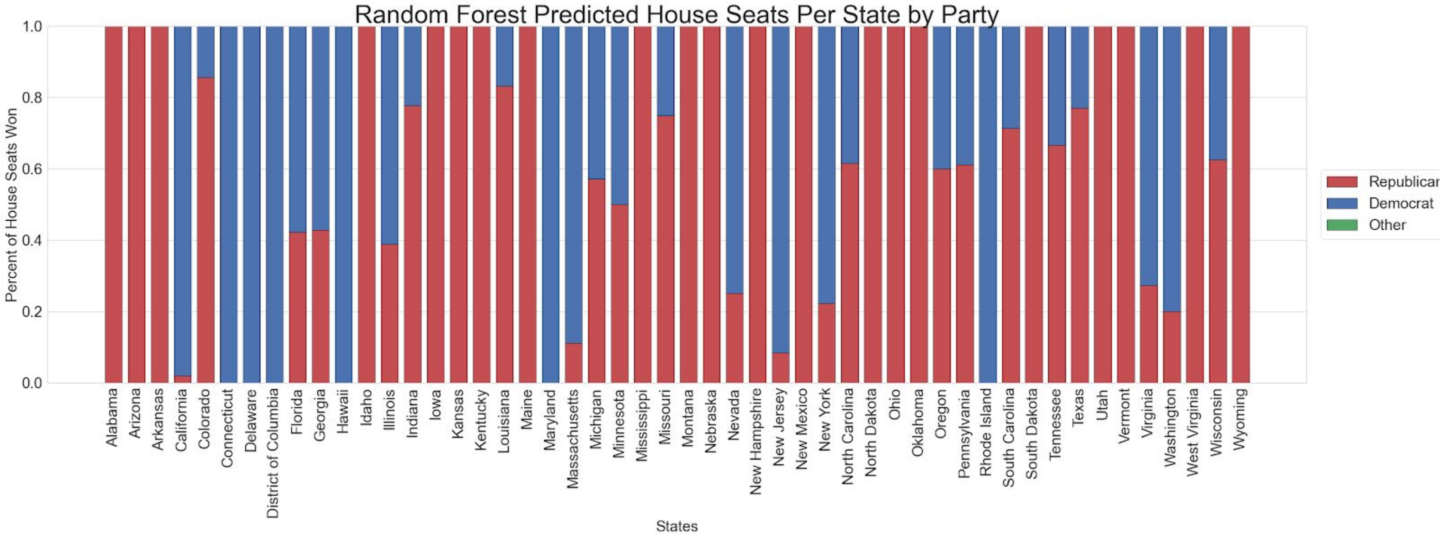
We utilized cross-validation and a Random Grid to tune our Random Forest in the same way we did the Presidential Election model, except using accuracy as our metric to maximize. Our best model had 100 estimators, 10 maximum features, a maximal depth of 90, and no bootstrap to fit each tree to reach an accuracy of 0.8734.

Results and Discussion

To investigate the details of the model, we looked into the random forest feature importances. As the chart below demonstrates, the features that our model took into account the most in our model were the demographic data in the form of the White American population by county, then the Democratic voter consistency, and then the Black American population by county. The white male and female populations by county also carried significant weight in predicting their house voting outcome compared to the other predictors. However, almost all the predictors in our model had a feature importance of below 0.05, which is relatively low.



Our random forest predicted that Republicans will win the majority of house seats with 224, Democrats will win 211 seats, and other parties will win 0. The predictions for each state are below:



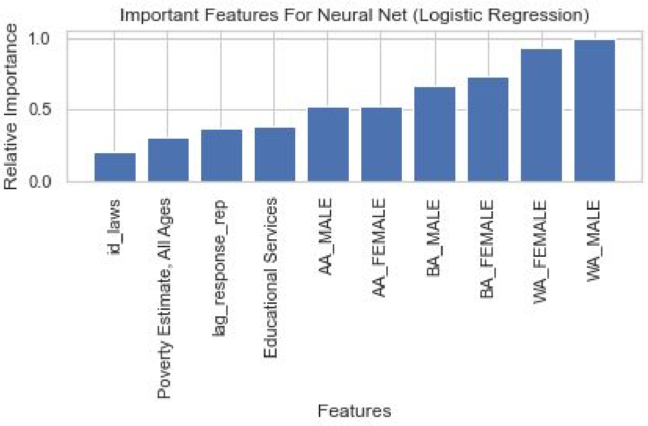
Method 2: Neural Network

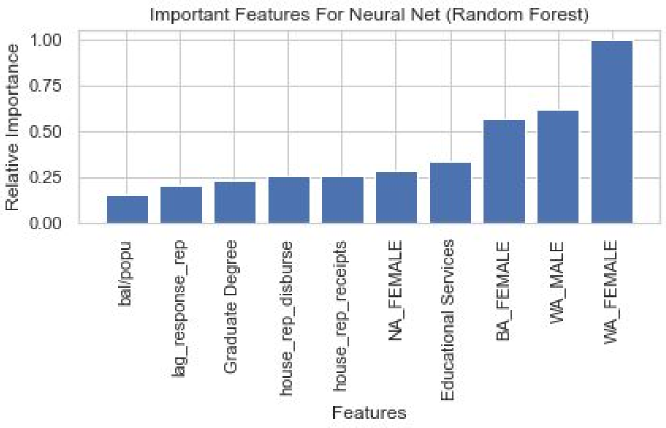
To handle our Random Forest's issues of a categorical response and unrealistic "winner takes all" design, we looked into a method that could handle three different regression outcomes: the percentage of votes for democrats, the percentage for republicans, and the percentage for candidates from other parties. We chose a Neural Network for this purpose—it would achieve three outcome variables while handling a high level of complexity and minimizing overfitting.

We tuned the model's parameters using k-fold cross validation, with the aim of minimizing mean squared error. Our final model had one input layer, one hidden layer with 1000 nodes, and an output layer that predicted three outcome variables. We opted for ridge regularization with a kernel weight of 0.001, a ReLU activation function in our hidden layer, and Sigmoid activation for our output layer. We used a batch size of 10, 150 epochs, a 0.2 validation split, and an Adam optimizer with a 0.001 learning rate. The epoch with the lowest cross-validation loss was chosen. This architecture yielded an R^2 score of 0.632 on the training set.

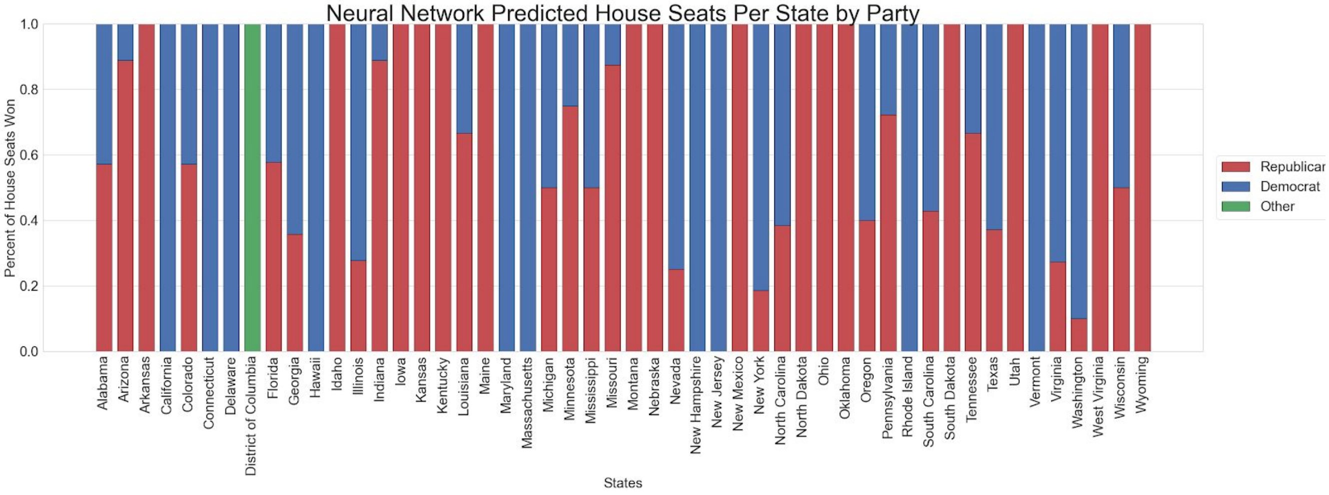
Results and Discussion

We ran permutation importance on our training dataset using both a logistic regression and a random forest. As the charts below demonstrate, the predictors that the neural network considered the most important were the White American populations, the African-American populations, and the degree of educational services available by county.





Unlike the Random Forest model, the Neural Network favored Democrats winning the 2020 election. It predicted that 242 seats would go to the Democrats, 192 would go to Republicans, and 1 seat would go to a third party.



In the actual results of the 2020 election so far, Democrats earned 222 seats, Republicans won 211, and other parties 0. Therefore, our Neural Network model appears to be more accurate than the Random Forest.

Limitations

While our President Random Forest and House Neural Network models predicted the correct overall results, all the constructs of this project had predicted proportions of votes or seats that deviated from the actual results. One possible significant contributor to inaccuracy is that the 2020 election results were affected by variables which are completely unique, such as the pandemic. Our models are weighted heavily on more static features such as voting consistency and racial profiles, and thus cannot predict volatile voting behaviors brought on by COVID-19. Furthermore, there are potential problems with the data. Many of the predictors, such as voter suppression, had no existing true measurements; we were forced to use proxies which may have not been accurate representations of the actual phenomenon. There were also holes in the data which needed to be filled with imputation, potentially leading to inaccurate variation. Beyond this, our observations were not independent, with neighboring counties tending to be similar to one another, violating an assumption that both Random Forests and Neural Networks make. Finally, we only had approximately 3000 observations which is not enough for a Neural Network to effectively train on, as is evident in our relatively low training score.

If we were to attempt again to predict the outcome of the presidential election, we would fit a second random forest model without those highly weighted features, such as racial profiles, education and voting consistency and plot the features importance with the remaining ones. By using this “mix ensemble models”, we can further and more accurately tune the weights for each predictor in the model building. For the House of Representatives election, there are several things we would change. We would likely use a Neural Network, but would have it fit and predict on data at a district rather than a county level because of the nature of the House of Representatives election—each elected individual wins and represents a district, and the model will likely be more accurate and generalizable to reality if post-hoc calculations are not necessary. Notably, this change would contribute more to the other significant problem in our model, which is how we do not have enough observations for a neural network to effectively fit to the training data. To solve this issue, we would also pull data from at least three other elections before 2016, and as a result have an enlarged training set that the model can learn from. Beyond this, we may change our methodology for how we choose the predictors to use and collect data. For many predictors, we had difficulty finding the actual data we wanted and instead used proxies which likely correlated with the values we were actually looking for. Instead, especially for the new data we would have to collect for 2012 and earlier, we would reappraise the variables that we choose to use after finding the most precise form it will take.

Conclusion

Overall, our models predict that Biden will win the presidential election, and the house will be won either by the Republicans by a relatively narrow margin if we use our Random Forest model, or the Democrats strongly if we use our Neural Network. The predictors we found to consistently be important across the two elections are race, particularly the White and Black American populations, the voting consistency, and the degree of education in a county. While our models were not able to strongly predict the election results, they show some promise and, if fine-tuned properly and used on complete and representative data, will likely be able to make accurate predictions.