Predicting Election Outcomes: 2016 Primary Elections

By: Eve Dean, Jacob Brandis, and Eric Vandament

Understanding underlying trends in voting patterns and election outcomes at various spatial scales has become an essential part of the campaign cycle and election process. Voting trends, which vary per county, state, and region, are influenced by a variety of factors including economic conditions and demographic factors, among others. We wanted to see if certain parameters were able to accurately predict the number of votes cast, as well as predict the winner of the democratic and republican primaries.

Historically, access to voting has been limited for certain groups, including women and people of color. We wanted to see if using demographic features like the racial breakdown of a county, could be used to predict the number of votes as well as the election outcome. To predict the outcome of the democratic and republican primaries, the racial demographics per county were input into a logistic regression model to predict whether Hillary Clinton won or lost the primary election. The overall accuracy of this model was **79%** The confusion matrix for this model shows that the model was fairly accurate, predicting true-positives **80%** of the time, and true-negatives **97%** of the time. This model also had the highest roc auc score of 0.80. Overall, this model was effective at predicting if Hillary Clinton won or lost the primary in a given county. When we used the same inputs to predict outcomes of the republican primaries, the model was far less effective. Despite having an **80%** overall accuracy, the logistic model for predicting if Donald Trump won or lost in each county was not as effective overall. It was worse at predicting true-positives (only **21%**), and slightly better at predicting true-negatives (**97%**). The roc auc score of this model was much lower, at 0.50.

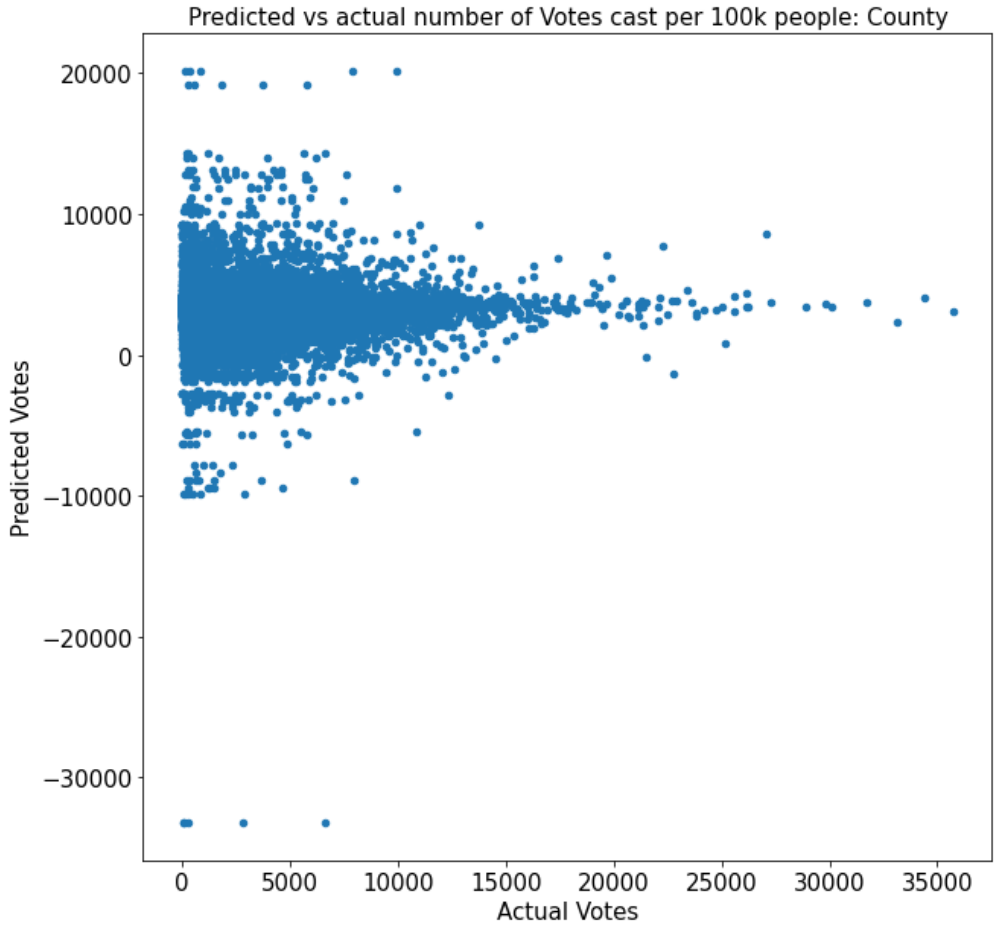
In addition to the racial demographics of the county, we wanted to see if areas with a higher percentage of women in the population would be more likely to vote for Hillary Clinton. This model did not prove to be very effective or accurate, as it had the lowest accuracy (**60%**) and roc auc score (0.5). It was more effective at correctly predicting when Hillary lost, scoring a true-negative **96%** of the time.

To dive more into these relationships we decided to do a logistic regression model to see if you could predict if a candidate was going to win or lose based on economic factors. Once again doing this by each party, we fit the model to predict if the top candidate were to win or lose each county during the primary elections. To find out the effectiveness of these models we calculated the accuracy by comparing the outcome of the model to the actual outcome of each county. For

the democratic party the model had a **63.02%** accuracy when predicting who would win each county. Surprisingly the model for the republican party was high, with it predicting the outcome of each county correctly **77.39%** of the time.

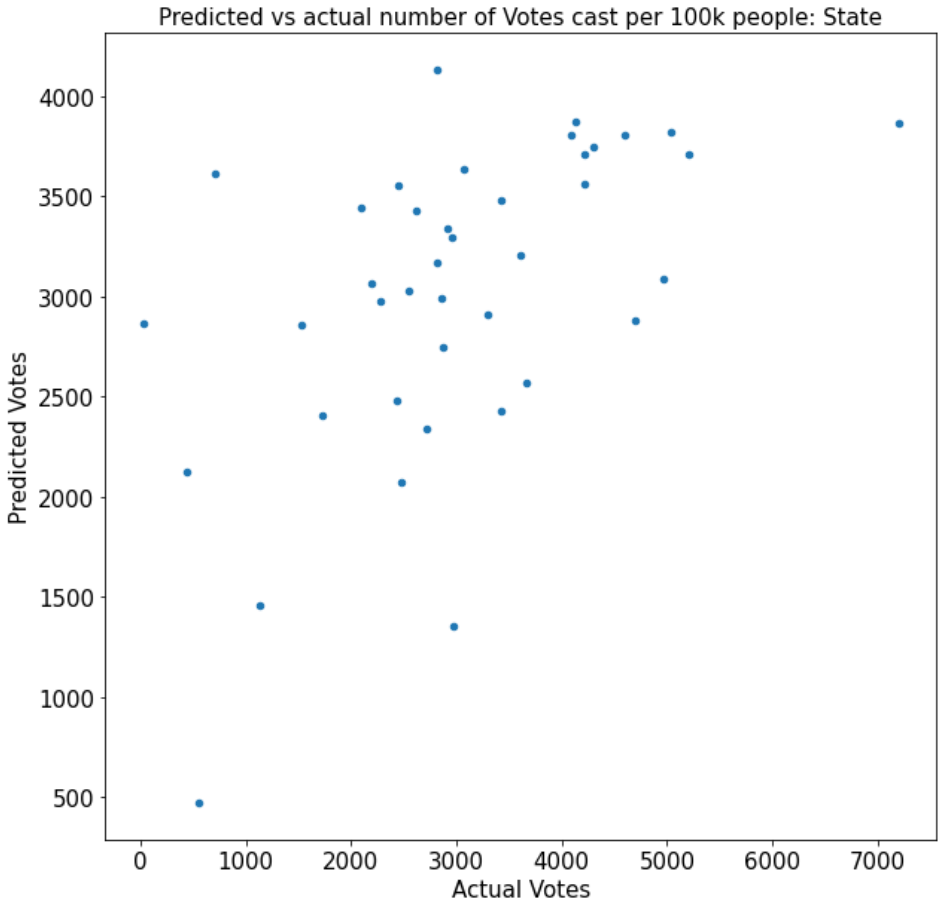
In addition to predicting outcomes of elections, knowing how many people vote in what areas is extremely valuable. Knowing which areas have higher or lower voter turnout can be used in political campaigns to gather support more effectively or to promote increased access to voting services. Once again, we used both demographic and economic data to predict the number of votes cast per state and per county.

To predict the number of votes cast, we used a linear regression model. The first model used the same demographic inputs as before to predict the number of votes cast per 100K people. We looked at the relationship between the number of votes and demographic information at various spatial scales, starting with counties.



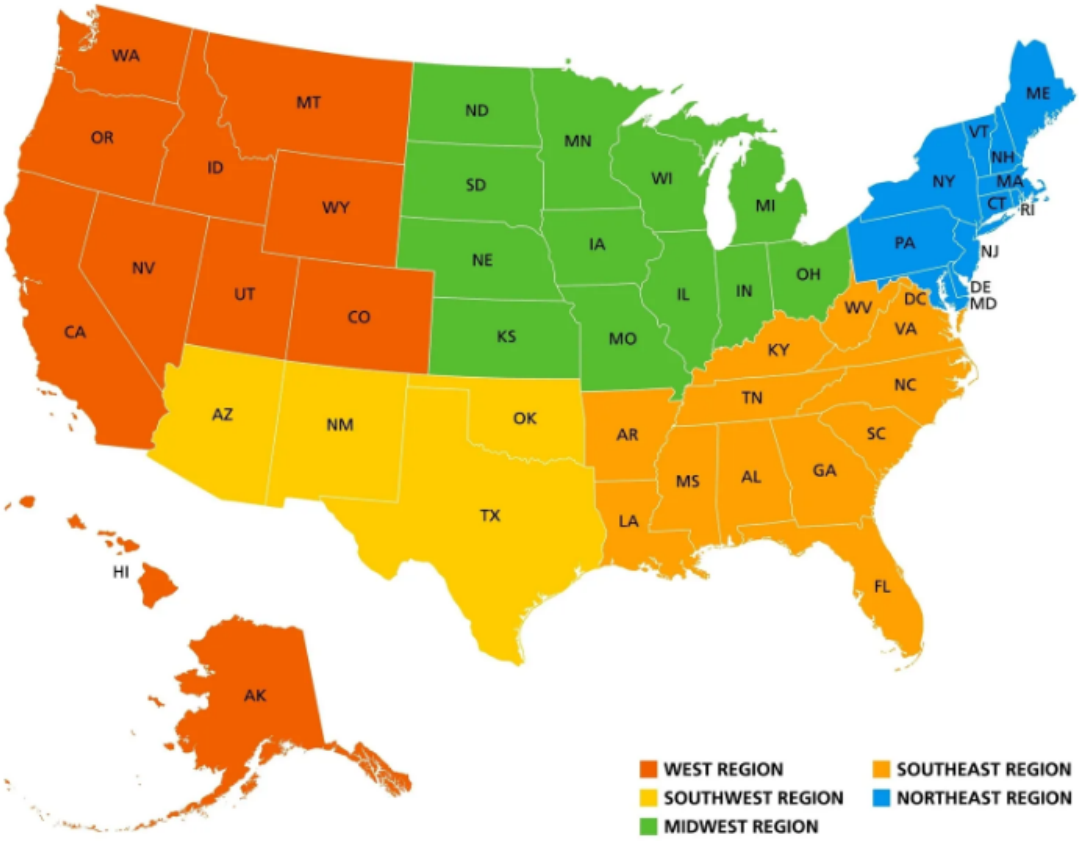
As you can see, the county's model did not have very accurate predictions. This model had an R^2 value of -0.25 and a MSE value of 12614795. In the county model, the percentage of biracial and

hawaiian people had the most significant effect on the model with coefficients of 72.6 and -456.9, respectively.

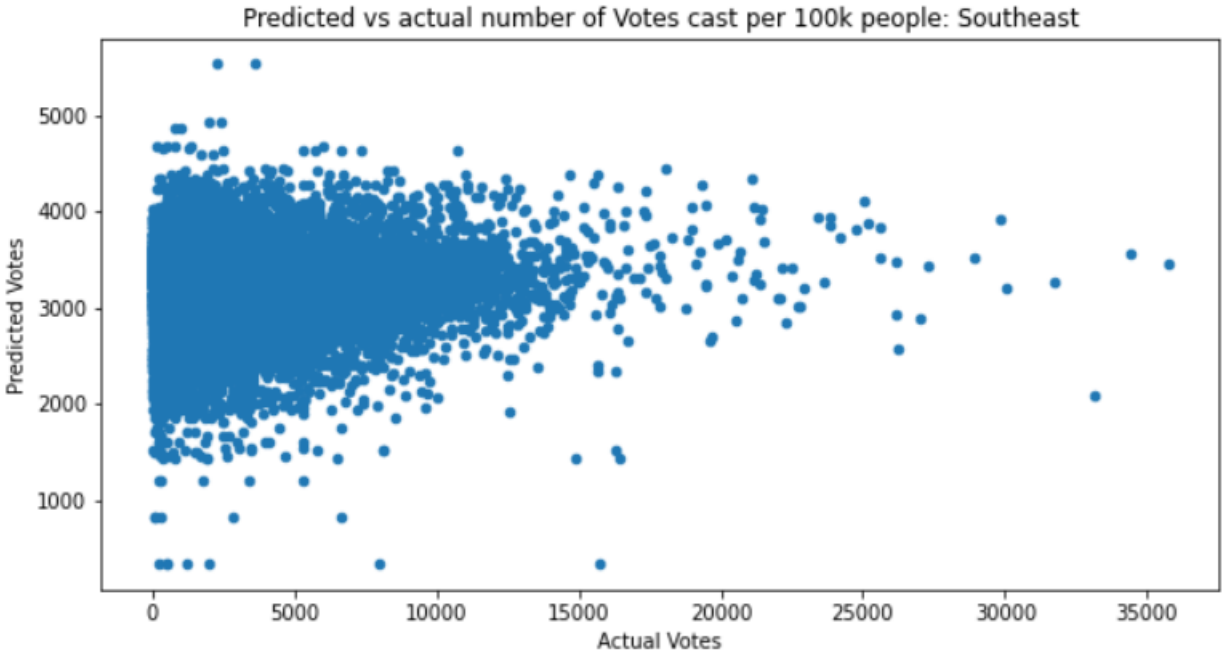


When grouping the data by state, the model was slightly more effective, but still not very useful. The state model had an R^2 value of 0.29 and a MSE value of 1460084. Additionally, compared to the county model, the coefficients were significantly smaller indicating that each input had a smaller overall effect on the number of votes.

Another factor that we looked into was if the region had an effect on the number of votes when looking at demographics. We grouped up the states into five regions: the northeast, the southeast, the midwest, the southwest, and the west.

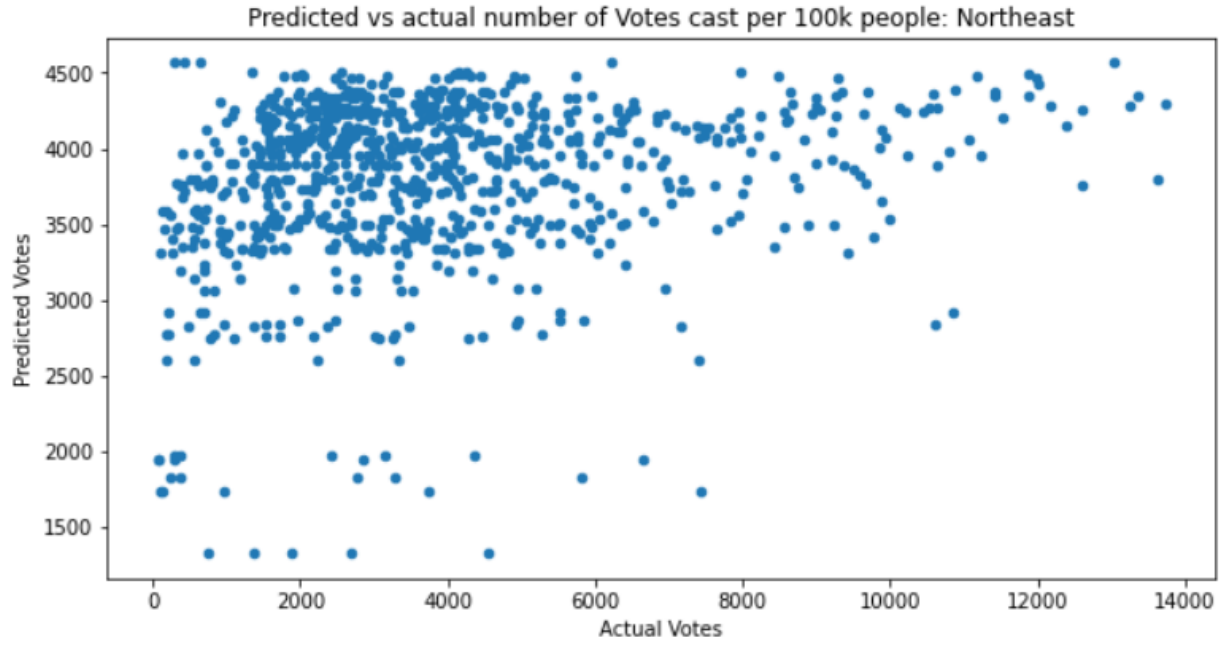


When grouping the states by region and making a linear regression model looking at the county's votes, we noticed that almost all of the regions closely followed the same pattern that appeared when studying all of the country's counties.



Above is a scatter plot for the southeast, but this bulb-like shape was seen in the plots in the southeast, the midwest, the southwest, and the west. However, on major difference was found in the northeast.

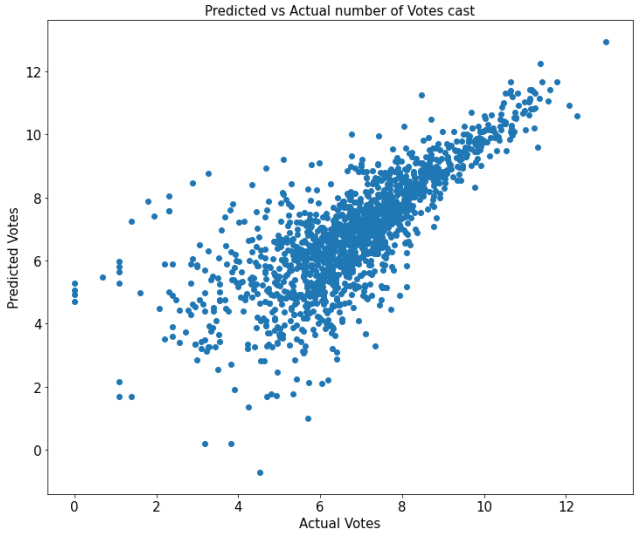
This graph was dramatically different from all of the others, and did not follow the same pattern. Additionally, the coefficients of the factors studies (demographics) were almost all positive, while the other regions had them almost all negative.



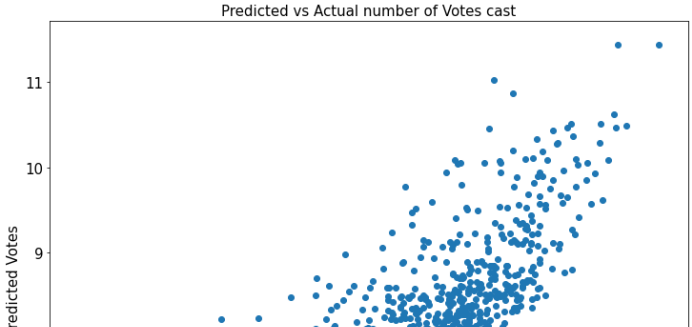
One potential explanation for this is the northeast is a dramatically different region than the rest of the United States. The northeast had a very high percentage of white population, had the highest median income, the highest bachelor's degree or higher rate, highest by population by sq. mile, and the lowest poverty level. There are many factors that can impact voter turnout, and although there was not the strongest correlation between demographics and votes cast, it seems like the northeast was a vastly different. Economic and racial factors could have potentially caused this difference.

We wanted to look at how economic factors might play a role in predicting the amount of votes a candidate will receive, and whether they will win or lose in the primary elections. The economic factors that we used were the amount of manufacturer shipments, merchant wholesaler sales, retail sales, and food services sales per county that were all represented in $1000. These previous mentioned columns were used for our linear regression and logistic regression. Since our data encompasses primary elections we decided to split our candidates up into two dataframes based on their party affiliation.

The first linear regression model that we ran pertained to the democratic party. You can see below that when it comes to the predicted amount of votes a candidate receives compared to

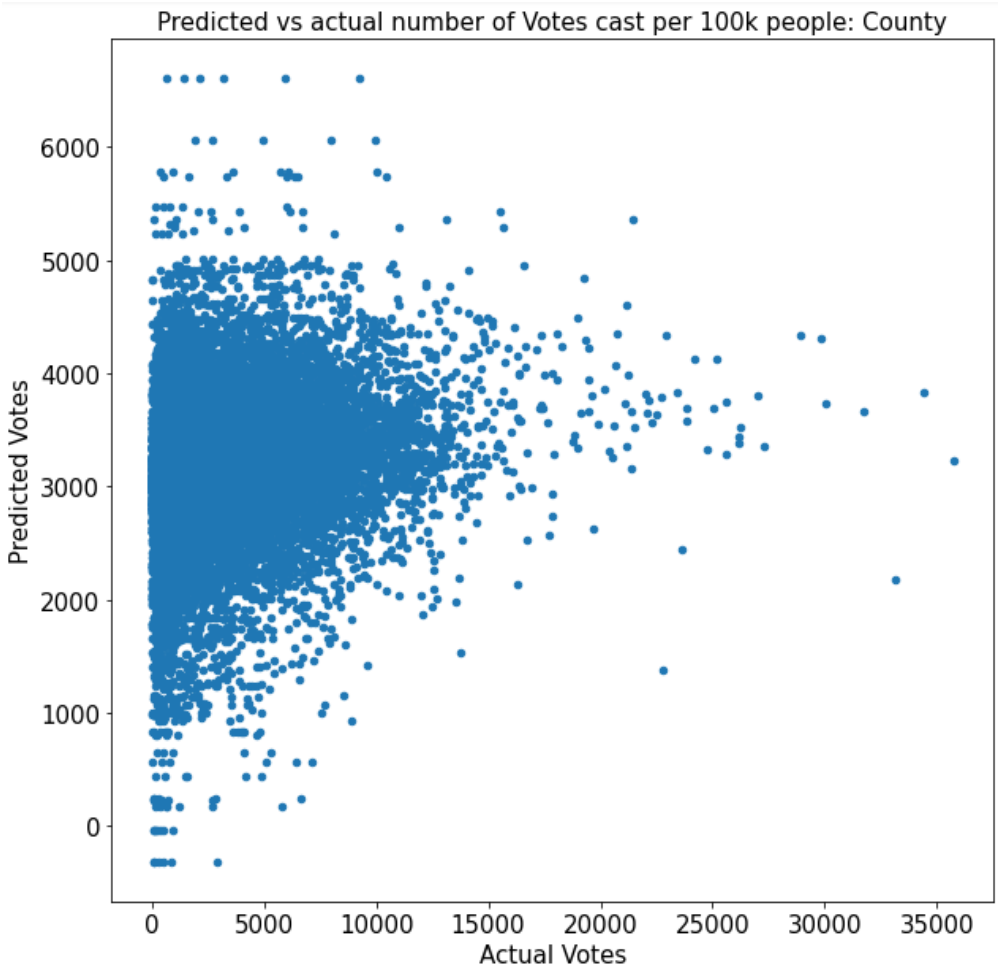


actual votes that they received. Economic factors show a correlation that could potentially be used for predicting votes. This is backed up with a R^2 value of 0.771, which isn't the best however, it is still a good start. When running this same linear regression model for republican candidates, we saw a lower correlation between economic factors and predicting votes. The model also provided us with a lower R^2 value of 0.612.



With the graph above representing predicted votes vs actual votes for republican candidates based on economic factors you will see a sort of exponential relationship. Compared to the democratic candidates this model is much worse at predicting lower amounts of votes with it plateauing at just over 7.

Lastly we wanted to look at how all conditions could predict the number of votes within a primary election.



We hoped that giving the model as much information as possible might help it develop better predictions of the number of votes, but that was not the case. This model had an R^2 value of 0.04 and a MSE value of 9687731. In this model, the percent biracial and food service sales per $1000 had the highest coefficient values, 6.6 and -7.7, and thus were the most influential inputs in the model.

In conclusion, understanding underlying trends in voting patterns and election outcomes at various spatial scales is an essential part of the campaign cycle and election process. We found that voting trends vary by state, county and region, but there is no specific pattern, at least related

to racial demographics. This was shown to us by the irregular voting patterns of the northeast region. Using demographic and economic conditions independently from one another does not result in valuable predictions, but using them together is not significantly better. One interesting trend is that the demographic models appear to be more effective at predicting democratic outcomes whereas the economic models appear to be more effective at predicting republican. Given the number of different factors involved when trying to predict the outcome of an election, it's not shocking that our models underperformed, but we could be more selective in which factors to include in the model. Remember to go out and vote so others can create more predictions.