Group 34 Milestone 3

Data Description

For our project, we are utilizing a combination of economic, social, and sports data in order to attempt to develop a prediction scheme that could provide insights into elections beyond simply incumbency. One of the driving motivations behind our data choices was a desire to connect public sentiment to the elections and determine social trends and their party correlations. General mood seems likely to have a strong relation to election results, and it is well known that factors such as incumbency and sitting presidential party have significant effects on elections. While we will be looking to include these factors in our final model, for this initial exploration and the corresponding model we were looking to isolate the impacts of these other economic and social factors.

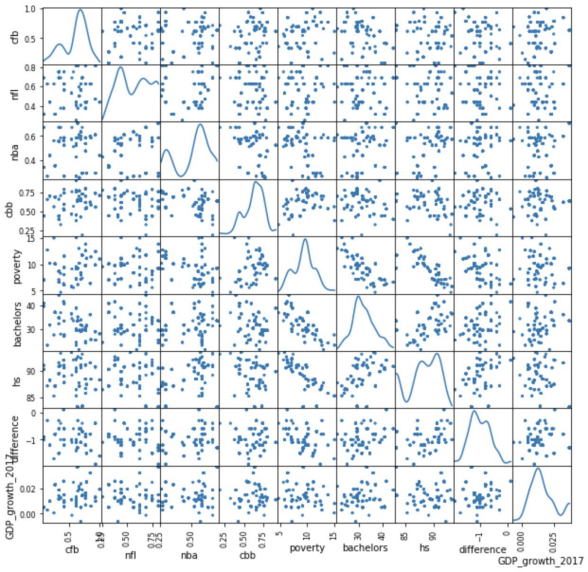
With regards to our collection methodology, we used government data for the economic and election information, and used google for the sports information. For the sports team information, we took the best major team in the state for each category, and for states without a major team we used different approaches for each sport. For college football, we gave states without a team the average of all other states, while for the NBA we used nearest geographical team and for NFL we used the primary TV market for the state. For the economic data, this was provided on a statewide basis and so could then be easily combined with our existing state by state sports data. For the social data, this was provided on a district by district basis and so for this model we chose to consolidate this into statewide averages. For the election data, we had to determine the winner for each district as our data came only with the vote counts, and so we took

these counts and created a new dataframe with the result for each district and used this as a basis for all further explorations. From this, we merged our other data frames onto this election dataframe using the states as the merge columns and then used this for our visualizations and the splits for our model.

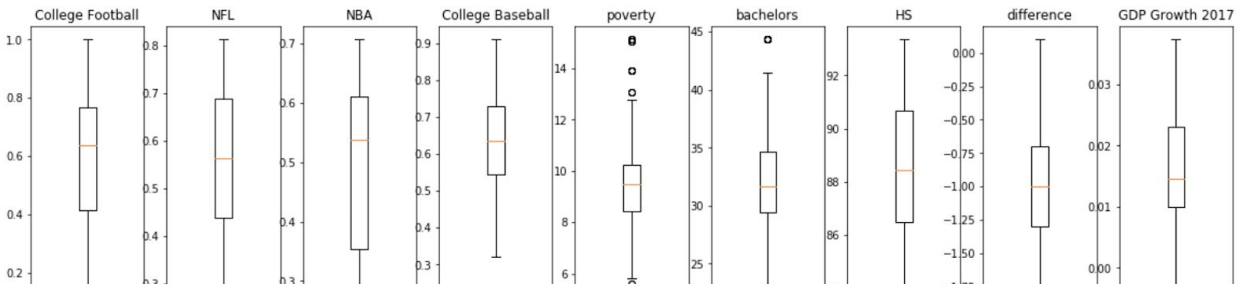
EDA

For our EDA, we looked at creating visualizations demonstrating the party split present in our indicators. In doing so, we discovered that often there was not as clear of a relationship between the features and parties; however, we were still able to find some interesting relations.

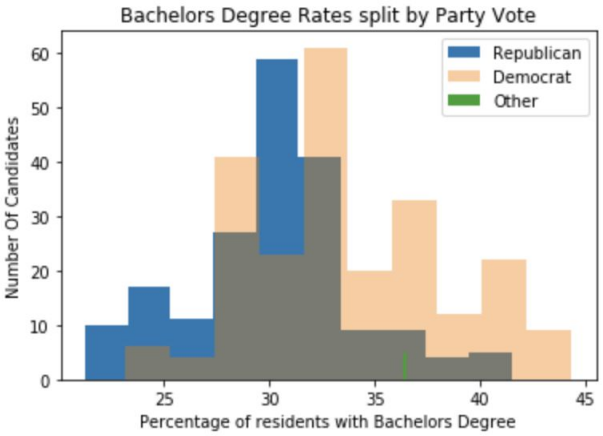
A scatter matrix of the features reveals a clear relationship between the number people living under the poverty line and education. The poverty rate decreases with the number of high school and college graduates, which is logical due to the fact that increasing educational attainment is correlated with better economic outcomes.

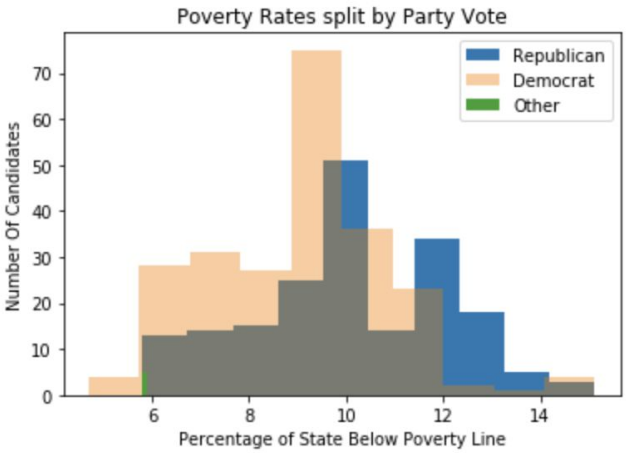


The outliers in the dataset occur in the boxplots for college baseball results , poverty, and the number of bachelor degrees (Massachusetts as discussed) . The unusually low win rate for college baseball (0.185) is for Maine which voted Democrat. The high poverty rates occur in Mississippi and Louisiana which voted republican and New Mexico which voted democrat, so voting decisions are split among districts with high poverty rates.



The histograms of each feature give more color to the distribution of each feature according the elected party. Most of the states with the highest poverty rates voted republican, while the states with lower to average poverty rates voted democrat. The average number of college graduates seems higher in states that voted democrat.





Revised Question:

Our revised question is as follows: How do local economic, athletic, and other off-meta factors improve the ability to predict incumbent or party success in midterm elections?

How do the features that display a significant correlation with elections outcomes improve predictions in comparison to a model that is just trained on incumbency data?