**Milestone 3**

**Project: Predicting Midterm Elections**

Goals: Predict results of all federal House of Representative elections in November per district for each state. Predict the probability that a particular party will win the midterm elections.

**1) A description of the data: What type of data are you dealing with? What methods have you used to explore the data (initial explorations, data cleaning and reconciliation, etc)?**

**Data Used:**

For this phase of the project, in order to build our baseline model, we are using 2 datasets, **FEC data** and **Polling data**. FEC data contains official election results for the House of Representatives per year at a district level. Polling data contains results on national polls 2 weeks prior to election day. Below is a table of the predictors we are using for milestone 3 along with their description and types. Each observation in our dataset contains information of how a district voted in the previous year and the current year’s national polling margin.

|  |  |  |  |
| --- | --- | --- | --- |
| **FEC Data** | | | |
| **Predictors** | **Variable Definition** | **Values** | **Type** |
| dem\_win | Is current winner democrat? | 1 = Democrat  0 = Republican | Categorical |
| dem\_win\_prev | Was previous winner democrat? | 1 = Democrat  0 = Republican | Categorical |
| margin\_signed\_minus\_prev | By what margin was the last election won? | (0,1] = Democrat won by some margin between 0-1  [-1,0) = Republican won by some margin between 0-1 | Floating Point |
| State | Which of the 50 states is the district from? | One hot encoded | Categorical |
| **Polling Data** | | | |
| national\_poll | By how many percentage points are the democrats leading in national polls 2 weeks before the election? | + for democrat lead  - for republican lead | Floating Point |
| national\_poll\_delta\_divide |  |  |  |

**Data Extension**

We would like to extend our data to include the state as a predictor. We hope that state information will capture in itself the difference in demographics across states and that it will improve our prediction.

|  |  |  |  |
| --- | --- | --- | --- |
| **Other** | | | |
| **Predictors** | **Variable Definition** | **Values** | **Type** |
| State | Which of the 50 states is the district from? | One hot encoder of size 49 | Categorical |

Furthermore, we plan to explore using more historical data, perhaps going back as far as 2006, so that we can cover two presidencies (Bush in addition to Obama), so that we have data from House elections during years with a Republican president in addition to a Democratic one.

Our efforts to extract and process demographic data from the US Census API are still ongoing, but we hope to try at least a few demographic predictors.

**Data Cleaning and Reconciliation**

Cleaning FEC data

We used the FEC data compiled by the MIT Election Data and Science Lab. Some of the party names were filled with NaNs, so we filled parties for these candidates by looking up their profiles on Wikipedia. We also calculated the winner of each election using the number of votes. However, for two states (Louisiana and Georgia), the election goes to a runoff if no candidate achieves 50%. For elections where a runoff is happening, our current model could be training on the wrong winner. We will find additional data to confirm these elections for the final model.

Redistricting

In several instances, a state went through redistricting. During this process, the borders and number of districts in each state may change. For this preliminary report, we simply dropped the districts that were not present in all years. For the final model, we’re using the district shapefiles to obtain an area-weighted average of the predictors based on the overlap between the new districts and the old.

**Baseline Model**

As a baseline model we used logistic regression on 80% of the data and validated the model results in the remaining 20%. This initial model achieved 94% accuracy on the training data and about 92% accuracy on the validation data (Figure 1). This provides us with 2 main results:

1. We can achieve above 90% accuracy on predicting House of Representative district winner by only using previous year results.
2. It is most common that districts usually vote for the same party in consecutive elections.

The overall breakdown of the House does not change very much in the years 2012-2016 (Figure 2). This may cause problems because we are training with three very similar years, so there are many out-of-sample scenarios (i.e., the Democrats winning the House) which our model is not prepared to predict. In order to fix this, we can add data from additional years.

**2)** **Visualizations and captions that summarize the noteworthy findings of the EDA.**

**Data Exploration**

**Figure 1. Proportion of districts that have a change of party in consecutive elections**

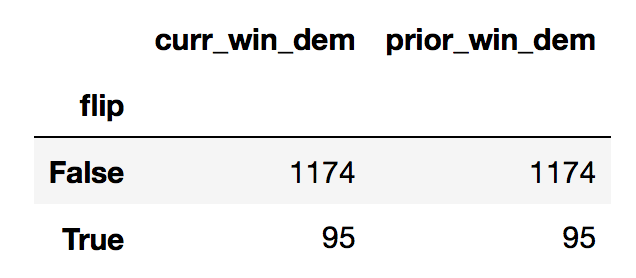
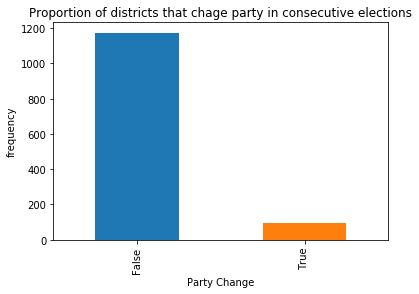
****

Figure 1 shows that the majority of the districts vote for the same party as they did on the previous election (blue bar) the supporting table shows the exact numbers we found on our dataset. With these numbers we see that ~92.5% of the time the district re-elects candidates from the same party and only ~7.5% of the time do they vote for the opposing party.

**Figure 2. House Breakdown 2012-2016**

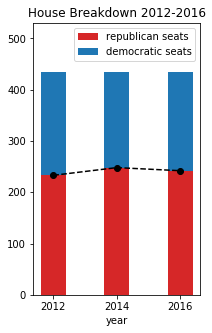
****

Figure 2 shows the breakdown of the House for each year in the preliminary model. The Republicans (red bar) hold a majority for all three years. The number of seats held by each party do not change very much between the 3 election cycles despite 8% of the seats flipping. This suggests that the “flipped seats” cancel each other out.

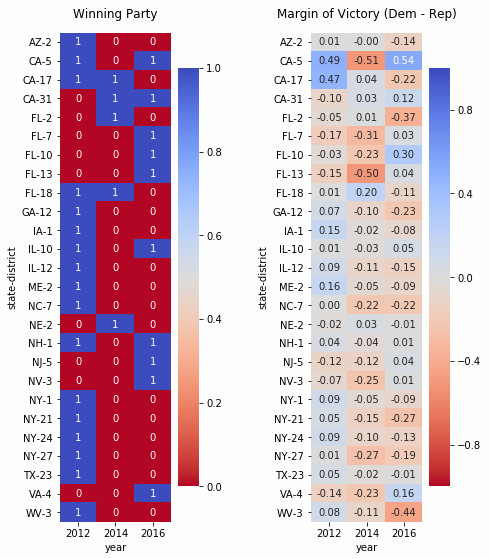
**3) A revised project question based on the insights you gained through EDA.**

Based on the model results we noticed that the model learned that most congressional districts did not change parties in consecutive elections. To improve our model we would like to improve our prediction performance on the observations in which the district’s party actually changes in consecutive elections.

There are several ways we could address this. One option is to change the training data so that half of the observations represent districts which flipped and half represent districts which did not flip. We can either do this by 1) downsampling the districts which flipped by randomly selecting a portion of those districts or 2) upsampling the districts where the party didn’t flip by bootstrapping them to create a number of observations equal to the number of districts which flipped. A third option is to use Adaboost to assign greater weight to districts where we make errors in our predictions.

**Supporting Materials**

**Figure 3. Exploring Margin of Victory for Congressional Districts Experiencing Party Flips**



**Figure 3 Commentary:**

After our baseline model revealed that most congressional districts did not switch parties in each House election, we wanted to explore this through visualization. For the two most recent elections (2014 and 2016), only 26 congressional districts experienced a change in parties.

The figure on the left is a direct depiction of which party won each election, where blue denotes Democrat and red denotes Republican.

The figure on the right shows a heatmap of the margin of victory for each of these elections, where the margin is calculated as Democrat vote % - Republican vote %.

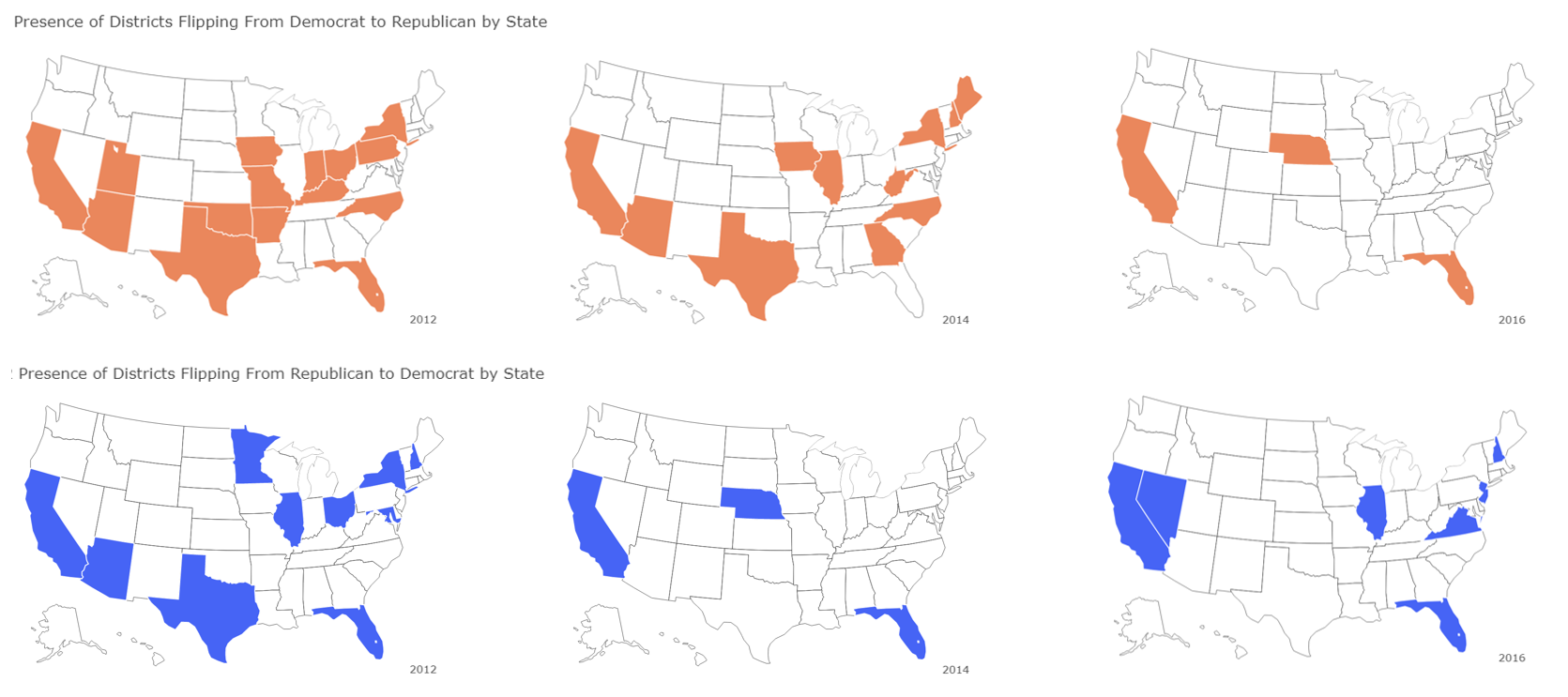
A higher color intensity points to a larger margin of victory, and blue and red still correspond to Democrat and Republican, respectively.

We see that, in elections where the party flips, the margin of victory is often small, indicating a close race.

Note that there's a single outlier (California District 5), which experienced redistricting.

We also produced a heatmap of margin of victory for districts that did not experience party flips in the past two elections, but didn't include it here since there would be 400+ districts to show. On average, the margin of victory for elections in these districts tended to be higher. However, there were still some districts that had relatively low margins of victory despite not flipping parties.

**Figure 4. Exploring States with Congressional Districts Experiencing Party Flips**



**Figure 4 Commentary:**

Given a baseline accuracy of >90% without predicting any flips, being careful to accurately predict the necessary conditions for a flip in a congressional district becomes particularly important. To that end, we explored the states (and their geographical proximity) that demonstrated these flips across the 3 elections that were analyzed. It’s clear that not all states are ‘created equal’ in the context of flipping districts. For example, the larger states (California, Florida, Texas, etc.) are typically very likely to contain a district that flips. Interestingly, it seems that there are also many states where we can expect to see no districts flipping at all. There’s no immediately clear phenomena of flips being clustered into states that are in close geographical proximity. We will incorporate this information and further exploration into the final model when carefully considering the different circumstances that need to be considered for predicting accurately district flipping.