Predicting the 2018 Midterm Elections

Group 42

Elise Penn, Manish Vuyyuru, Yajaira Gonzalez, Victor Sheng

Fork it on Github!

https://github.com/WildTangles/ac209a_project.git (https://github.com/WildTangles/ac209a_project.git)

Table of contents

#Overview

The goal of the project is to predict the winners of the House of Representatives in each congressional district for the 2018 election. The House of Representatives is divided into 435 representatives, with each representative elected by people living in a congressional district. Each congressional district represents ~711,000 constituents, except in the case of states which have less than 711,000 residents - these states simply get one representative.

In 2018, all three branches of government were controlled by the Republican party. However, in the midterm election directly following a presidential election, the party of the president typically loses seats in the House of Representatives. This fact, combined with the low popularity of the current president and the record-smashing fundraising of the democrats this year, led many people to predict a “blue wave” in the congress this year.

Every 10 years, each state redraws the borders of its congressional district based on the latest update of the U.S. Census. The borders of the districts are often manipulated to favor whichever party holds more power at the time of redistricting - this is known as “gerrymandering.” Gerrymandering in order to create an advantage for a particular party is typically tolerated by the courts, but if gerrymandering disenfranchises a particular race of people, the courts may order the state to redraw their lines. Thus, redistricting happened in some state for almost every year in our study.

#Motivation

Predictions of elections are not only interesting as a media sensation, but also vital for candidates to strategize their campaigns. By affecting the allocation of resources, these predictions not only report on the elections, but also affect the outcome of elections. (Our model will, fortunately, not affect any elections.)

Election predictions are challenging because the outcome is affected by so many variables, but there are very few observations (i.e., elections) relative in proportion to the number of predictors. They are also challenging because the game changes almost every election. In addition to swings in political favor of the population, the party in power has the ability to literally redraw the board to favor themselves via redistricting. The combination of these two factors makes predicting elections a particularly compelling and challenging problem.

Description of Data and EDA

1. **FEC Data**
   Vote counts from prior elections in each district.
   Years available: 2002-2018
2. **Demographics Data**
   Information from the U.S. Census about the demographics of each district. Due to time constraints, used limited demographics data.
   Years available: 2010-2018
3. **Polling Data**
   Polls conducted during the election season. Due to time constraints, we used only national aggregates.
   Years available: 2002-2018
4. **Geographic Data**
   Used for modifying above data. District borders were used to impute Demographics and FEC data where necessary.
   Years available: 2002-2018

FEC Data

FEC data comprises of several potential predictors that are provided on a per district per state per year basis. The information can be summarized as either candidate-level information (e.g. names of candidates) and district-level information (e.g. total number of votes). Candidate-level information such as campaign financing/scandals would have been interesting to consider and there is evidence that they are likely strong predictors of the outcome of a race (see fivethirtyeight). For district-level information, of particular interest were two observations, of the number of votes and the total number of votes cast for each candidate per district per state per year. From this information, we computed the winning candidate per district per state per year and by joining on the candidate-level information (party affiliations of candidates), we arrived at our response variable (winning party per district per state per year).

Also, from the fraction of votes garnered by each candidate and their political affiliation, we compute several metrics which served as predictors. In particular for a given response in year t, we compute metrics from the results of the election in $t-2$. The following metrics were computed:

Let w_t be the vote fraction earned by winner in year t, and l_t the vote fraction earned by the 2nd place candidate. For a response variable in year t,

metric 1:

$$w_{t-2} - l_{t-2}$$

metric 2:

$$w_{t-2}/l_{t-2}$$

Let d_t, r_t be the vote fractions earned by the democrat and republican. For a response variable in year t,

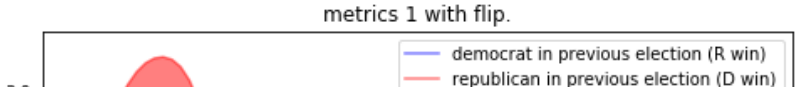
metric3:

$$d_{t-2} - r_{t-2}$$

metric4:

$$d_{t-2}/r_{t-2}$$

.



Literature Review/Related Work

Model Development

We learned a lot from FiveThirtyEight's comprehensive description of their model for the 2018 midterm elections [3]. Primarily, we learned which predictors we should emphasize when collecting data for our model. These were the top four predictors according to FiveThirtyEight in order of importance. The following list is a direct quote from [3]:

> - **The incumbent's margin of victory in his or her previous election**, adjusted for the national political environment and whom the candidate was running against in the prior election.
> - **The generic congressional ballot**
> - **Fundraising**, based on the share of individual contributions for the incumbent and the challenger as of the most recent filing period.
> - **FiveThirtyEight partisan lean**, which is based on how a district voted in the past two presidential elections and (in a new twist) state legislative elections. In our partisan lean formula, 50 percent of the weight is given to the 2016 presidential election, 25 percent to the 2012 presidential election and 25 percent to state legislative elections.

Demographic data is not discussed in this ranking of predictors because FiveThirtyEight employs a system called CANTOR to handle their demographics. CANTOR classifies districts based on their demographics. When states redistrict, FiveThirtyEight uses CANTOR and a kNN to impute affected districts with the data from the districts closest to them in demographics.

Our model largely represents our best attempt to replicate this model where we can, and find other solutions where we cannot. Because of the limited time frame, the amount of data we were able to find free online & clean in the time frame given was extremely limited.

For the incumbent's margin of victory in the previous election, we used FEC data. Unlike FiveThirtyEight, we did not directly adjust this number for the political environment and opposing candidate. We found the generic congressional ballot averaged over the entire nation, but were unable to find it at the district level. We found fundraising data, but were unable to find data for 2018 in time to process it, so we did not use this variable at all. Our equivalent to partisan lean was which party won in the previous year. And finally, instead of CANTOR, we used demographics directly as a predictor.

With these different predictors, we ended up with a very different model than FiveThirtyEight, and a very different order of importance of our predictors. See the results for more details.

Because we didn't have sophisticated enough demographics data (i.e., we didn't have demographics for very many years, and did not have important predictors like race), we were not able to impute using kNN as FiveThirtyEight does with CANTOR. Instead, we decided to use a "population overlap" model to estimate changes when redistricting occurs.

Redistricting

Modeling Approach

The aim of the project is to predict the district winners for the House of Representative in the 2018 election. Given that the parameter we are trying to estimate is a categorical value; whether the winner of a district is a democrat or a republican. This is a classification problem. Our modeling approach is to use a set of classification algorithms trained on different datasets and predictors to find the combination that most accurately predicts the district winners for 2018.

###Baseline

Our baseline approach is to run a set of models with a limited number of predictors and assess the accuracy achieved. Using a small set of predictors reduces model complexity and decreases potential for overfitting. Simpler models also allow us to better understand the relationship between the predictors and the response variable.

###Extensions

Our modeling extension is to incorporate more features into our baseline dataset and run the predefined set of models on each combination. For this we use a combination of FEC Data, National Polling Data, Demographics Data and our approach for Re-districting. The goal of the extension model is to explore if by adding more predictors to the baseline dataset we are able to capture more information about the districts that do undergo a party change in the 2018 election compared to the previous election while at the same time keeping a good classification accuracy on the prediction for the districts whose winner is indeed from the same party as the previous election.

Models Used

The table below shows a summary of the models chosen and the reason for their use.

Model	Assumptions	Reason for use
Logistic CV	Linear relationship between the log odds of the outcome and each predictor	Somewhat interpretable
LDA	Assumes that observations within each class are normally distributed with their own mean but same covariance matrix between the two classes.	Chosen as an additional linear model for comparison purposes.
QDA	Assumes that each class is normally distributed with their own mean and covariance matrix.	Chosen for its flexibility to find non linear decision boundaries. We suspect that nonlinear approaches will do better than linear.
KNN	Flexible model, allows for non-linear boundaries, predictions based on closer observations	Chosen due to its flexibility and the assumption that similar districts (districts closer to each other in the feature space) will behave similarly
RandomForest	No assumptions, highly interpretable but low predictive power	Chosen due to interpretability of the results
Boosting	After the first iteration, predicts the residuals from the previous iteration. It trains on the error of the previous iteration	Chosen with the hope that it will help with rare events in the dataset, such as districts that undergo a party change from the previous election, which is where the majority of the misclassification is occurring.

Baseline #1

The baseline model refers to the use of a subset of predictors from the FEC data. The predictors for the baseline model are election results from the previous year and the state the district belongs to. Below is a summary of the accuracy of each model.

Baseline Model	R2	Test Accuracy	% Non-Party Flips Correctly Predicted	% Party-Flips Correctly Predicted
Logistic Classifier CV	**0.92**	**88 %**	**100 %**	**0 %**
KNN-5	0.90	85 %	97 %	0 %
Random Forest	0.92	88 %	100 %	0 %
Boosting	0.92	88 %	100 %	0 %
QDA	**0.92**	**53 %**	**58 %**	**20 %**

In the table above we see that even though the logistic regression model achieves ~88% accuracy on the test data, all the predictions that are correct are those for which the winner of the district is from the same party as the winner from the previous election. We call this approach the baseline model as for all observations in the test data the predicted party is always the same as the winner of the previous election. The QDA models shows something different, with a lower R2 score and lower test accuracy, QDA is able to predict 20% of the districts in the test data that whose party changed from the previous election winner. However we gain this at the expense of reducing classification accuracy on the districts whose winner is from the same party at the previous election.

Baseline #2: Extended

The Baseline Extended Model attempts to capture more information about the districts whose winner party is different using a different subset of data from the FEC. We attempted to choose predictors which keep a good classification accuracy on the districts whose winner is from the same party as the winner from the previous election. This model uses the margin by which the winner of the previous election won over the loser party and the state the election took place in.

The purpose of this exercise is to explore the limits of our baseline dataset.

Below is a table summarizing the results.

Baseline Extended Model	R2	Test Accuracy	% Non-Party Flips Correctly Predicted	% Party-Flips Correctly Predicted
Logistic Classifier CV	**0.48**	**53%**	**49 %**	**90 %**
KNN-5	0.91	86 %	97 %	4 %
Random Forest	0.98	86 %	97 %	4 %
Boosting	0.93	87 %	97 %	6 %
QDA	0.48	49 %	53 %	18 %

With the baseline extended model we see that we are able to increase good predictions on the districts that flip by 90%, however this is at the expense of misclassifying more districts whose winner party do not change from the previous election. QDA results are very similar to the previous model, adding more predictors did not improve the QDA model. However, the rest of the models did see an improvement on the classification accuracy of the districts that do flip.

Baseline Extended Takeaway

The majority of the districts vote for the same party that won on the previous election and only a small number of districts see a party switch. Because of loss function maximizes total predictions, and we have so few districts which switch states, it is challenging to create a model which predicts which districts will flip accurately.

Given that the goal of the project is to predict the winner party for each district in 2018 we still need to do well on the districts that do not switch parties but we would like to capture at least some of the districts that do switch. Using only FEC data, the best model we can achieve is the boosting model, which achieves 87% overall test accuracy on 2018 and is able to predict correctly 97% of the districts that do not switch parties between elections and captures only 6% of the districts that do switch parties.

Based on these two baseline models, we conclude that more data sources are needed in order to improve the accuracy of the model on states which flip.

#Model Extensions

In efforts to improve our prediction on the test set while being able to capture information about the districts whose winner party changes from the last election we explored a set of extensions that uses a combination of different datasets on the model described above.

Below is a summary of the different extensions we looked at:

Extension	Brief Summary
Extension1: **Full FEC +** **National Polls**	**FEC** Uses a combination of candidate level information (party affiliation, votes earned) and district-level information (total votes cast) to construct several candidate margins to capture how close the previous election was. **Polling** Contains information about the national polls performed 2 weeks prior the election. National polls survey how people plan to vote for the upcoming election.
Extension 2: **Full FEC +** **Polling +** **Demographics Data:**	**Demographics** Socio-economic characteristics of the population aggregated at congressional-district level from the Census Bureau's American Community Survey (ACS). Details include: gender, age, veteran status, native origin, unemployment status, educational attainment, household income, etc.
Extension 3: **Full FEC +** **Polling +** **Demographics +** **Redistricting**	**Redistricting** Redistricting: data sources typically provided information on a per district per year basis. in some cases, we impute data across years (e.g. when census data is unavailable) or calculating margins across years (e.g. how close the previous election was). Accounted for this fact that districts across different years can represent different geographical areas by projecting data from per-district space to population-overlap space.

Changes in Project Goals

The original baseline model which used logistic classification with a small set of predictors learned that most congressional districts do not change parties in consecutive elections and so for every district prediction it predicted the winner to be from the same party as the previous election. This gave us a good classification accuracy as the majority of the districts vote in the same way as they did in the previous election, however it was not able to predict a single district whose winner was from a different party as in the previous election.

To improve the classification accuracy on districts whose party changes from the previous election we tried upsampling the number of observations where this event happens (which is very small) or using AdaBoost to assign greater weight to districts were incorrectly classified. Upsampling did not improve the classification accuracy, however AdaBoost was able to capture some of the districts that voted for a different party as they did in the previous election. For this reason we added AdaBoost to our list of models to try.

Variable Selection

Of the 24 possible predictors in the baseline model, we handpicked 10 of them as likely good predictors of the election. However, we wanted to make sure that none of the predictors were highly collinear. Collinearity would cause our models to find suboptimal solutions because one of the assumptions in linear regression is that the variables are independent of one another.

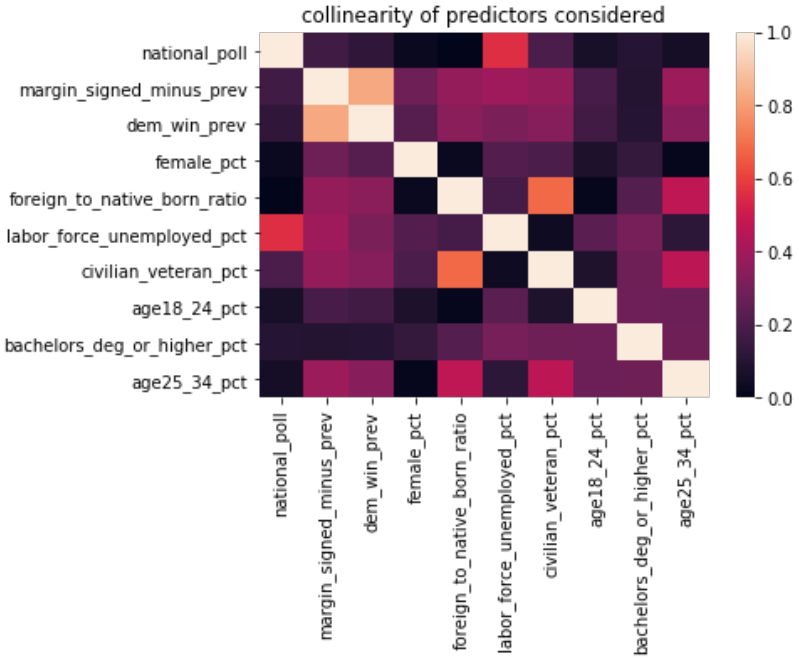


Figure 3: Collinearity of predictors considered. Color bar represents the correlation coefficient between the two predictors where 1.0 represents perfect correlation and 0.0 represents no correlation.

Unsurprisingly, we saw that `margin_signed_minus_prev` (% of votes received by democrat - % of votes received by republican candidate in the previous year) and dem_win_prev (whether a democrat won last year) were highly correlated. Of the two, we chose to use only `margin_signed_minus_prev` because it provided additional information about how decisive the victory was, rather than simply who won. Immediately after we removed `dem_win_prev` , we saw increases in prediction of flipped races (from 0% to 20%+) without much change in the overall accuracy.

The only other notable correlation we between `foreign_to_native_born_ratio` and `civilian_veteran_pct` . We thought this symbolized the urban/rural divide rather than any causation, but the correlation coefficient was high enough that we removed `civilian_veteran_pct` to reduce collinearity.

Interestingly, `labor_force_unemployed_pct` was moderately correlated with `national_poll` . Since we were only looking at the last 3 years, the economy happened to be recovering from the 2008 recession (thus, decreasing unemployment everywhere) at the same time as the republicans were gaining support. Whether this is a causal relationship is up for debate. In the end, we removed `labor_force_unemployed_pct` not only because of this correlation, but also because it cannot be imputed to prior years (because unemployment changes quickly between years) in the same way as other demographics.

#Results

Evaluation of Datasets

A model which uses only FEC and polling data for eight years is shown in Figure 4 below. With only FEC and polling data for 8 years, we were able to attain a high level of overall accuracy, but this was primarily achieved by predicting that the previous year's winner would win again. As a result, each model predicted almost all districts which did not flip correctly, but none of the states which flipped. Although we are using more years of data, this is still very similar to the baseline model. As discussed above with the baseline models, this indicates that more information is needed to improve the model.

Extension 1: Full FEC Data: 2004-2018

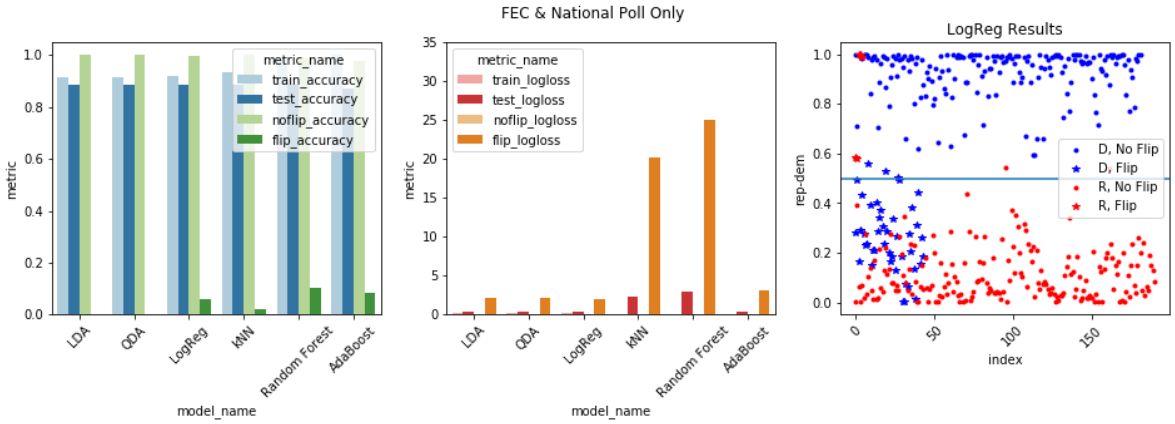


Figure 4: This model uses only the FEC and polling data. States are not dropped if they were redistricted the prior year. The first panel shows accuracy for the training set, the testing set, and for flipped and nonflipped districts in the testing set. The second panel shows logloss. The third panel shows the percentages predicted by the Logistic Regression model. Above 0.5 on the y axis indicates a democrat was predicted, below 0.5 indicates a republican was predicted. Colors indicate the true value. Predictors used: `national_poll` , `margin_signed_minus_prev`

Figure 5 below shows a model which uses both FEC and Demographics data, but only for four elections. With only four years of data, we were able to improve the prediction of districts which flipped from 0% to 40% (in the case of Logistic regression). 40% is still not a high rate of prediction. However, more important than whether the model made a correct prediction is whether the model was able to give reasonable probabilities for each result. This is interpreted by log loss, which was fairly low.

The FEC & Demographics model using only four years (Figure 5) was by far our best, both in terms of raw prediction and in terms of applying acceptable probabilities to the predictions. This is the only model which correctly predicted that the democrats would win the house in terms of the raw prediction (to get a true estimate of which party it predicted, one would need to run this model multiple times and obtain the distribution of outcomes).

Extension 2: FEC and Demographics Data: 2010-2018

Evaluation of Models

We used two metrics to evaluate our models: Accuracy and log loss. For a probabilistic model such as the one we built, accuracy only tells part of the story. We acknowledge that because our information is incomplete, the best model we can build is a model that doesn't always get the right answer, but which has a high degree of uncertainty on its prediction when it does make an incorrect prediction. Log loss is a metric which is designed to score a model based on the probability of its predictions.

We scored these metrics on four subsets of the data: the full training set, the full testing set, the districts which didn't flip in the testing set, and the districts which flipped in the testing set. Predictably, all metrics performed best on the training set, equal or slightly less well on the training set, extremely well on districts which didn't flip, and poorly on districts which did flip. This result was uniform across models.

However, LDA, QDA, and LogReg performed consistently better than kNN, Boosting, or Random Forest. These models, which allow for more complicated lines, either may overfit the training data, or they are too "good" at finding the best predictor (the winner of the previous year), at the expense of the districts which flipped.

Interestingly, although the flip accuracy varied quite a lot between the different datasets for LDA, QDA, and LogReg, the flip logloss did not change at all. This may indicate that most of our models were quite similar, but that small variations in information allowed them to place flipped districts slightly more on one side of the 50% line than the other.

Future Work

There are several directions that future work can explore.

On the front of data, the current set of features could be expanded to include finer, candidate-level information such as campaign financing that are reported in similar works to be strong predictors (see fivethirtyeight) and other features could certainly be explored (how common the candidate's name is). There are several other directions that this could take, such as using polling at levels below national (e.g. state-level polling) which are also reported in similar works as strong predictors.

On the front of the modelling, it would be more appropriate to consider architectures that more faithfully replicate representations in the real world. This could again take several directions. Notice below, a plot of the geographical location of districts that flip between 2010-2016. 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 expect to see no districts flipping at all. It appears that flips are typically relatively clustered geographically. An example of how this could be accomplished would be to use a KNN model that identifies nearest neighbors by geographical distance.

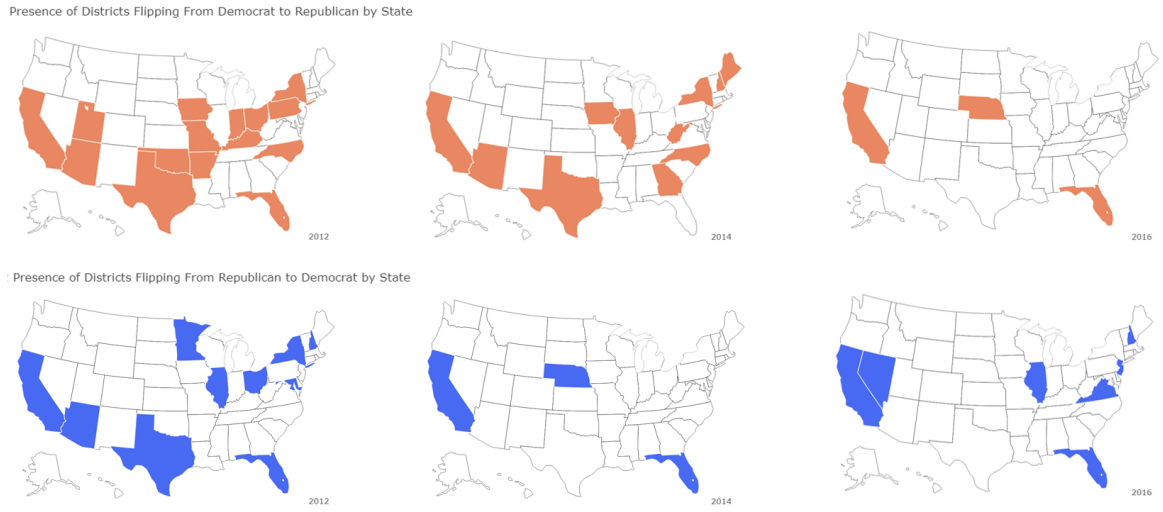


Figure 7: Geographical distribution of party flips across elections years.

Conclusions and Summary:

At the outset of the project, we naively hoped to create a model which would predict the 2018 elections with a high accuracy score. We soon learned that no matter what datasets we threw at the model, all of them were insufficient to create a model with a better *accuracy score* than our baseline model. Instead, we refocused our efforts to create a model which had more realistic *probabilities* on the outcome. This probabilistic model of the 2018 elections allowed us to use the information we had while also acknowledging the limitations of our data.

Successes:

We **improved our prediction of flipped districts** over the baseline model by adding demographics data and using only the most meaningful data from prior elections. Demographics allowed us to further constrain if a district was likely to flip.

We found a way to **address redistricting** without the comprehensive demographics data used by FiveThirtyEight. By estimating the percent of population which moved from one district scheme to the next using the overlap of district boundaries, we found a tolerable way to impute prior election data.

The probabilistic models enabled us to gain a deeper understanding of how our models were working. In the end, we had only a few data points which were predicted with a high degree of accuracy but which were incorrect.

Challenges:

Our final model ran into two fundamental problems: our **imbalanced dataset** and **insufficient data**.

Because only a small number of districts flipped (i.e., voted for a different party than they did last year), the model could be highly successful by predicting that all or almost all of the districts would flip. This is the classic problem of the imbalanced dataset. Although we tried upsampling the prior dataset, we could try changing our performance metric to something specifically designed to combat this problem, such as Cohen's Kappa.

We were also plagued by insufficient data, both in terms of the number of years the data covered, and the number of meaningful predictors we had available. One issue that we ran into repeatedly was that running a new model was a 5-minute problem... but cleaning a new dataset is a 5-hour problem! Fortunately, the solution to this problem with more data does not require any ingenuity, only time. The datasets which we believe would improve our model the most are *district-level polls*, *fundraising* data, and *demographics data*, including race & ethnicity, for the full time period on which we trained our model.

References

[1] FEC results from (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/IG0UN2 (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/IG0UN2)) and (https://www.fec.gov/introduction-campaign-finance/election-and-voting-information/#election-results (https://www.fec.gov/introduction-campaign-finance/election-and-voting-information/#election-results))

[2] 2018 House election results from https://en.wikipedia.org/wiki/2018_United_States_House_of_Representatives_elections (https://en.wikipedia.org/wiki/2018_United_States_House_of_Representatives_elections) (accessed 12072018)

[3] Methodology/Importance of Predictors from https://fivethirtyeight.com/features/2018-house-forecast-methodology/ (https://fivethirtyeight.com/features/2018-house-forecast-methodology/)

[4] Shape Files: *2002-2014:* (http://cdmaps.polisci.ucla.edu/ (http://cdmaps.polisci.ucla.edu/)) and *2016-2018:* (https://www.census.gov/geo/maps-data/data/tiger-line.html (https://www.census.gov/geo/maps-data/data/tiger-line.html))

[5] Inspiration for shapefile algorithm: https://acdisc.gesdisc.eosdis.nasa.gov/data/Aura_OMI_Level3/OMNO2d.003/doc/README.OMNO2.pdf (https://acdisc.gesdisc.eosdis.nasa.gov/data/Aura_OMI_Level3/OMNO2d.003/doc/README.OMNO2.pdf), Section 6

[6] Inspiration for shapefile code: https://nelson.wisc.edu/sage/data-and-models/software.php (https://nelson.wisc.edu/sage/data-and-models/software.php)

[7] Demographics from: https://www.census.gov/programs-surveys/acs/ (https://www.census.gov/programs-surveys/acs/)