Redistricting Ideology

Drew Keller
Alejandro Navarrete
Victor Perez Martin
Evelyn Siu
Cole von Glahn
The University of Chicago Computational Analysis and Public Policy

June 2, 2022

Abstract

Partisan redrawing of voting maps in the United States is a central strategy for collecting and maintaining political power. Advances in algorithmic tools for this purpose foreshadow a deepening and hardening of these ideological bastions. Preparing responses to this potential future requires an understanding of the relationship between voter and representative ideology. Linking precinct demographics to candidate ideologies through vote share provides a proxy for the ideological spectrum at the precinct level. Feeding this information through predictive models predicated on potetial redistricted maps expresses the expected ideological score of a potential district, and therefore the anticipated ideological position of the most appropriate representative. Analyzing the delta between these predictions and future outcomes provides the public with a window into the extent to which their beliefs and interests are being accurately represented by their Congressperson.

1 Datasets

Our work leverages data from the following four datasets. Our models will be tested on Michigan's voting maps and political information. It will be trained on the political information from Pennsylvania, Wisconsin, and Minnesota which share many demographic, regional, and political features with Michigan.

After the cleaning process, we ended up with a sample of 1,042 precincts, out of the 4,765 precincts identified in the Census, with the respective ideology scores of the candidates that ran for Congress in the 2018 election. Ideology scores are centered around zero, a zero score indicates a moderate ideology, a score above 0 a more conservative leaning ideology, and a score below 0 a more liberal. These were transformed to categoricals according to the following:

Very Liberal if s less than -1

Strong Liberal if s between [-1, -0.5)
Lean Liberal if s between [-0.5, 0)
Lean Conservative if s between [0, -0.5)
Strong Conservative if s between [0.5, 1)
Very Conservative if s greater than or equal to 1

1.1 Candidate Ideology Scoring

Stanford's Database on Ideology, Money, and Elections (DIME) includes Candidate ideology scores. These scores are assigned based on the ideological positions of the Candidate's donors, as well as those causes/candidates to which the Candidate donates. Scores are provided for Candidates who reach the General Election. We performed fuzzy matching on candidates' names to combine election results and ideology scores, we decided to keep only the precincts where all the candidates have a minimum matching score of 70 out of 100.

1.2 Precinct-Level Voting Results

2018 US House Election Results at the precinct level data comes from Redistricting Data Hub. For each precinct, we have the names of the candidates that ran in the 2018 election, their political party and the total number of votes they received.

1.3 Voting Maps

Michigan's public redistricting process provides access to a number of proposed political maps. We use the current map as a control, and expand to the proposed maps providing varied precinct-to-district configurations for analysis.

1.4 Demographic Data

From the Census and American Community Survey (ACS) we got population characteristics at the block level like ethnicity, education, and income. In order to merge the election results with the census information and absent a common precinct identifier in both datasets, we combined the county, minor civil division, ward and precinct codes in the election results data to build an id to mimic the precinct id in the Census. As redistricting is often focused on voters' demographic qualities, the intersection of demography and ideology is a window to the strategic decisions being made by candidates and party bodies as they construct political geography.

2 Model and Experiment Details

We created three models to represent different analysis, categorization and interpretability levels. Each answers the question: what is the precinct's ideological mix and how accurately are they being represented?

2.1 Trees and Forests

As our baseline model, we chose to test out a decision tree. To improve upon the decision tree, we used the ensemble method random forest. Because the two algorithms are classification methods, we created a categorical label by which to classify our data. The y-labels used for these classifications indicate a winner (Democrat, Republican), and whether the election was close or not (Close, Split). The election was considered close if the percentage of votes each party received was within 50 percent of each other (inside range of 25-75). If the percentage vote breakdown was outside of this range (outside range of 25-75), the election was considered "split". These two labels tell us who is predicted to win the election, and whether the outcome is likely to be highly partisan (split) or not.

We varied tree depth as the hyperparameter. Below is a list of the accuracy scores for trees from depth 1 to splitting on all available attributes. For the random forest, we ran a cross-validation test to get the mean of several aggregated forests. The mean of the accuracy scores was 0.815, with a standard deviation of 0.040.

Decision trees of depth 1 and 2 had the highest accuracy score, at 77 percent, splitting on population density. The tree below only shows up to 2 decimals, but the average population density is 8.14 E004, so it likely does not accurately show the density it is splitting on. Instinctively, this makes sense, as we know rural areas are more likely to be Republican, and densely populated cities are more likely to be Democratic. Adding in proportion Black did not affect the accuracy score. Naively, Republican, Close makes up 66 percent of the entire sample, so a 77 percent accuracy is a mild improvement. In terms of feature importance, for the decision tree, prop more than 100k had the highest importance weight, at 54.5 percent, followed by prop some college, weighted at 15 percent.

A random forest with a depth of 5 had the highest accuracy score, at 80.8 percent. This was an improvement over the decision tree. For the forest, total pop was weighted the highest, at a score of 12.2 percent, followed by prop white, weighted at 0.9 percent.

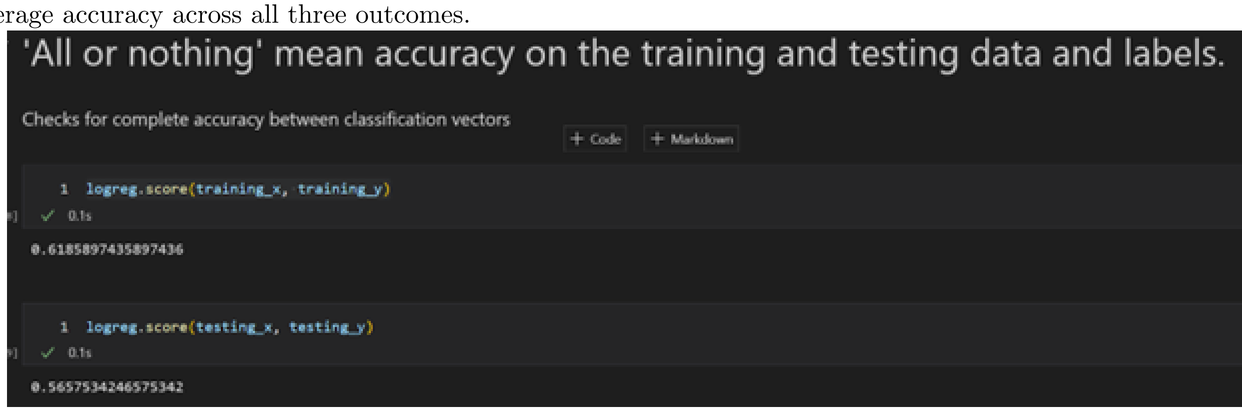
2.2 Logistic Regression

A logistic regression serves well in the case of using compositional data to predict an outcome. Even in the case of a multiclass classification, it is very much intuitive and interpretable to use a model that applies weights to features to generate probabilities.

In our model, we combined one multi-class dependent variable, a weighted ideology score representing the precinct, and two binary dependent variables, party of winning candidate and vote split (how the ticket was split between the top two candidates). Using the multi-output classifier wrapper available in SKL, we could run three multiclass predictions at once to predict a vector of the three variables of interest. This design would allow us to experiment with predicting all three outcomes independently and see if those outcomes can thus be predicted on the precinct demographics on their own. From there, we can look toward predicting the winning party, vote split, and weighted ideology using the weights generated by each classifier in a multi-output scheme.

With this data-set, to avoid multicollinearity in our compositional data, we removed one of the variables from each grouping of proportions. In this model, we removed the variables measuring the proportions of white, single, post-secondary graduates with income over 100K. This alters the baseline and interpretability of our model. These dropped variables would then be the baseline from which to interpret the model. Given that we used a model that does include a bias/intercept, this would mean that the weights given to all the remaining attributes would show shifts from the baseline. In other words, how much those attributes shift the probabilities of these outcomes away from what would be expected of a precinct composed only by those having the attributes from the omitted group.

Since this classifier is an ensemble of three classifiers, one accuracy metric would be an all or nothing measure. This would see if all values in the predicted outcome vectors match with the true outcome vectors. If not, then that prediction would be marked incorrect. Another metric would be to calculate an average accuracy across all three outcomes.



Average outcome-wise accuracy

Checks for accuracy per outcome in classification vector and returns the average percentage correct.

Ex: If overlap is 2/3 between one record's true vector of values and predicted vector, returns percentage of 66.67%

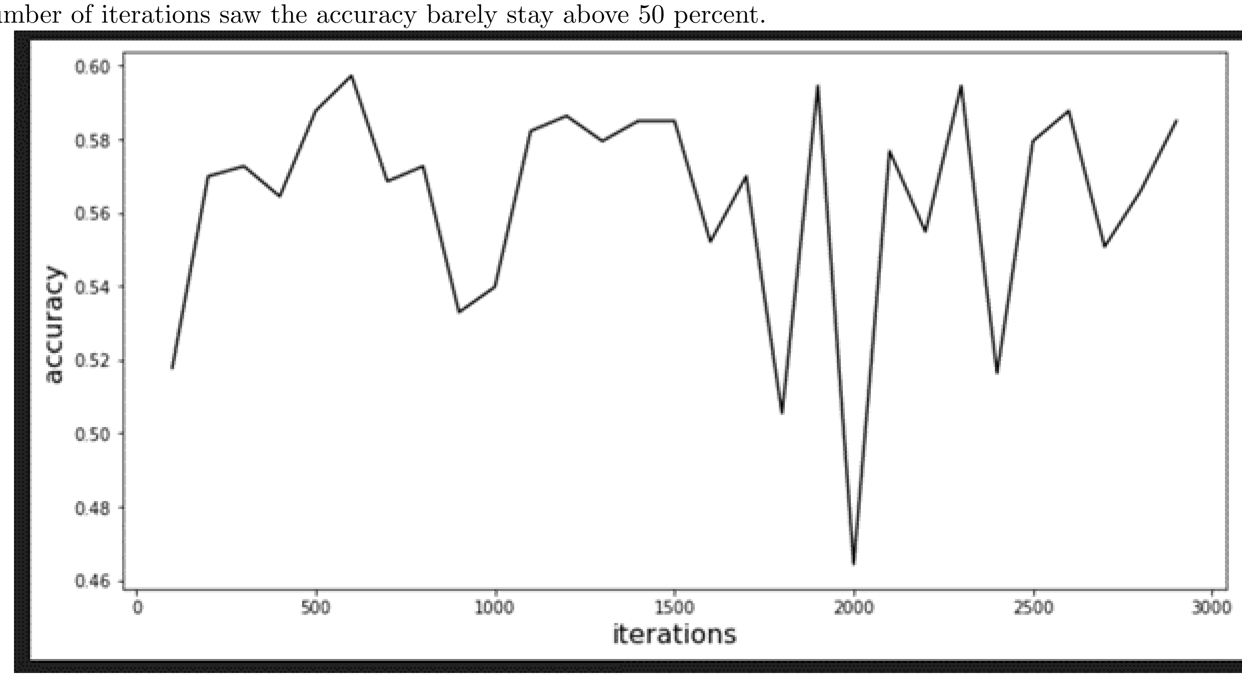
```
train_acc = np.count_nonzero(logreg.predict(training_x) == training_y) / (3*len(training_y))
print(train_acc)

test_y_hat = logreg.predict(testing_x)
test_acc =  np.count_nonzero(logreg.predict(testing_x) == testing_y) / (3*len(testing_y))
print(test_acc)
```

✓ 0.1s

```
0.8536324786324786
0.8273972602739726
```

To investigate the relatively low values of the mean accuracies, I experimented with two alternatives from the baseline. First, I manipulated the number of iterations to see if the learning algorithm required more iterations to converge the error to near zero. This was not the case since the increasing number of iterations saw the accuracy barely stay above 50 percent.



The next approach was to use a chaining ensemble method to possibly increase the accuracy where two of the three results may be more probable in conjunction with each other.

This approach actually led to a decrease in mean accuracies for all of our records in the dataset, from about 60 percent to 18 percent. However, the outcome-wise accuracy remained the same, around 83 percent.

To conclude, it would be that the logistic regression on this dataset would be better than random classification for all three values and highly accurate on each independent outcome variable. However, to create a conjunction of multiple, multiclass classification it would not perform as well as the classification tree.

2.3 Light Gradient Boosting Tree

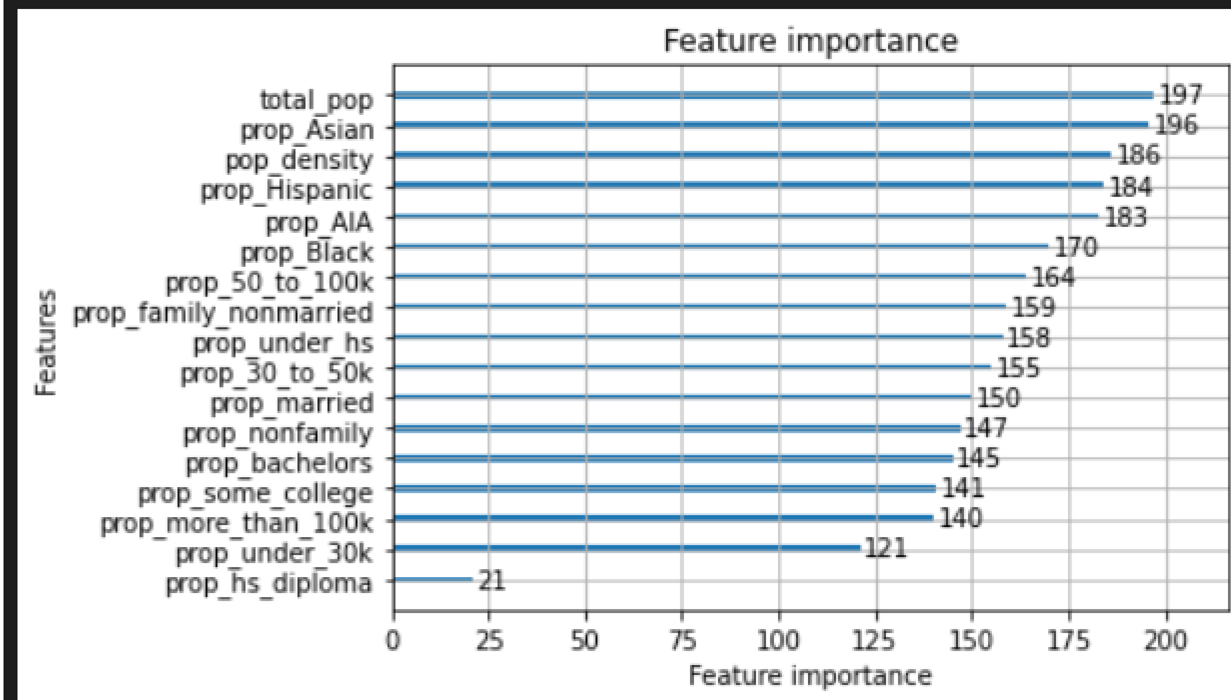
We employed Light's GBM Regressor model to provide a continuous variable alternative to the classification methods described above. Ideology as a contiuum is essential to our analytic paradigm. While it could be estimated by classification models, it could not be fully represented. Light GBM also yields high accuracy on relatively small datasets with many features by downweighting the lowest gradient and most mutually exclusive attributes.

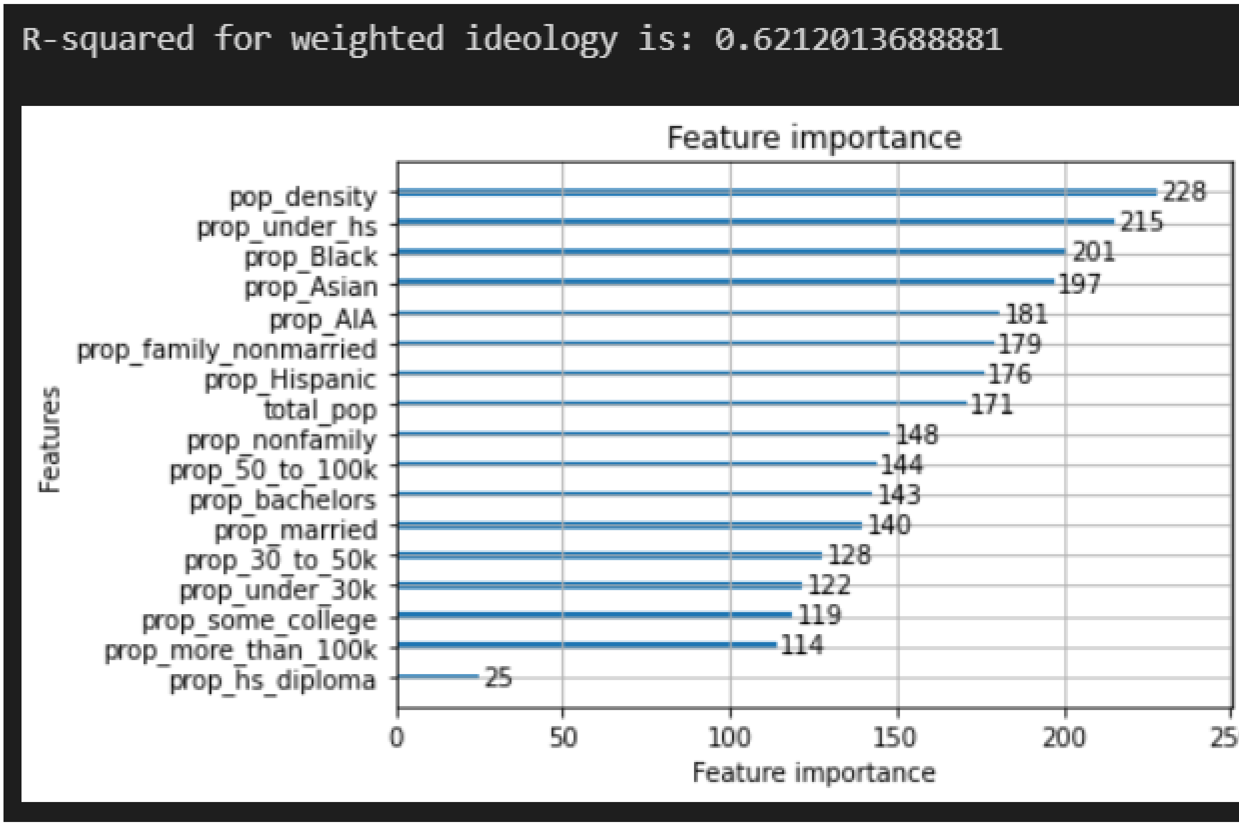
We used the Light GBM regressor to make predictions on voting outcomes as well as precinct ideology. For voting, our outcome variable was the percent of votes for the Democratic candidate. For ideology, our outcome variable was the ideology of all candidates in the field weighted by their vote share within the given precinct.

Our primary hyperparameter tuning involved the learning rate. We tested learning rates iteratively on both outcome variables, finding slightly different optimal rates for each. We measured success of this model through the R-squared value to show the model's efficacy in explaining variance relative to inputs.

The tables below shows the features sorted by importance as well as the R-squared values for both model outcomes.

```
R-squared for percent of democratic votes is: 0.617666317213521
```





In summary, Light GBM effectively explained the variation in both cases. However, it was stronger with the ideology mapping. Further tuning and larger datasets could improve performance. These results are encouraging, as the primary purposeof employing GBM is to best model the continuous spectrum of voter belief, whereas the classification prediction regarding winner is better handled by classifiers like trees and logistic regressions. That said, a future iteration of this project may benefit from employing GBM classifier models as comparison to the other classifiers.

3 Limitations

We expect our analysis to be limited by the following factors:

- The DIME dataset is missing a small number of candidates
- The DIME dataset is a back-out of ideology, relying on 3rd party contributors as measurement. It has the advantage of representing more candidates who received votes, thereby increasing our ability to understand the granular aspects of precinct voting preferences. However, it may be a less

precise metric than scoring systems that narrow their scope to election winners and label politicians based on their voting records.

- The census datasets do not label their geographic units in the same way as our other datasets. For this round we defaulted to county level measurement. We are continuing to develop our matching strategy at the precinct level.
- In our reading, and as shown in some of our results, demography is not the most specific set of features for predicting ideological belief. This is especially true at the regional/state level where the number of observations is limited.

4 Negative Societal Impacts

There is the potential that this work is used by political operatives to further understand and abuse differences in ideology and demography. Rather than a tool for changing candidate behavior and encouraging a closer match of representative to represented, it may be leveraged to instead alter the playing field in favor of particular candidates and their ideological beliefs.

5 Conclusion

Using the baseline categorizations of the decision tree and logistic regression, our random forests and gradient boosted decision trees provide effective means of determining both the voting outcomes and ideological makeups for precinct specific voter behaviors. These measures give policymakers the opportunity to assess the impact of redistricting on representation accuracy. If a district's aggregated ideological outcomes are split away from their voting outcomes it is a sign that the particaluar needs of that community are not accurately represented at the legislature level. That knowledge can reveal roadmaps for improving the health of representative democracy. More importantly it canprovide communities with an understanding of how likely their needs are to be met, and the extent to which their beliefs are fought for on the national stage.

6 Asset Attribution

- Bonica, Adam. 2019. Database on Ideology, Money in Politics, and Elections: Public version 3.0 [Computer file]. Stanford, CA: Stanford University Libraries. ¡http://data.stanford.edu/ dime¿.
- Bonica, Adam. 2018. "Inferring Roll-Call Scores from Campaign Contributions Using Super- vised Machine Learning" American Journal of Political Science, 62, (4): 830-848. (https:// onlinelibrary.wiley.com/doi/full/10.1111/ajps.12376).

- U.S. Census Bureau. (2012). 2020 American Community Survey. Retrieved from https://www.census.gov/data/datasets.html.

- Voting and Election Science Team, 2019, ”2018 Precinct-Level Election Results”, https://doi.org/10.7910/DVN/UBKYRU, Harvard Dataverse, V59.