Kerem Tuncer and Nikhil Gopal

Applied Data Mining

Wayne Lee

04/19/2021

Political Data Mining: Party Classification of Counties with Demographics

Introduction

Our team consisting of two Political Science-Statistics majors was contacted by Pew Research Center to come up with a machine learning model that would serve two purposes. First, we were requested to predict if a county will vote more Democratic or more Republican a year before an election itself. This information will be used by Pew's election forecasters to improve the accuracy of their forecasting models. Secondly, we had to mine for insights regarding which characteristics of a county make it more likely to vote for either of the two parties. The company will use these insights to see if their surveys about party identification and demographics can be further validated with Census data. Given that Census's sample is more representative of the American population than Pew's samples, we will add further value to their research by decreasing the uncertainty of their findings.

We combined the information from the Census data with the county-by-county election results that we found on MIT Election Lab. Given that the outcome of interest was a binary value (0 if Trump got more votes and 1 if Clinton got more votes), we did this task with a classification algorithm, which refers to predicting which of the aforementioned categories/labels a county will fall in. We used the output from our classification model to try to understand why certain

demographics resulted in our prediction to indicate a Republican or a Democratic majority in any given county.

Dataset

We first downloaded the ACS Demographic and Housing 5-Year Estimates for 2015 dataset from the website of the US Census Bureau. According to the Census Bureau, the 5-year estimates for a given year are more accurate than the 1-year estimates, which was the reason why we decided to use 5-year data. Additionally, the 5-year estimate dataset featured more counties than the single year one, resulting in more observations for us to use in the model. We also went with 2015 data rather than the 2016 one because, technically speaking, the data collection period would not have been done for the 2016 estimates by the time the 2016 election took place.

Cleaning up the Census data took a lot of time because there were over 300 columns. For each indicator, there was a column indicating the total estimate, the percentage, and the margin of error. We used a quick string function to get rid of the columns that had the word estimate and margin because percentages to the total county population were much more useful. Luckily, we were able to cut the dataset down to about 80 columns. Then, we went over each variable one by one to see if it could possibly make a good addition to our final machine learning model. In the end, we were left with 26 columns that we thought were valuable. One of those columns showed the Geographical ID number.

Then, we were able to get county election results from MIT Election Lab, which has county-level data from 2000-2016 (2020 elections are not available yet, unfortunately on a county-by-county basis). Luckily this dataset was much easier to work with. We first did a subset to get only the 2016 results. Then, we realized that each row showed the vote per country per

party. So, once again, we did a subset to only leave Republican and Democratic rows. Then, we had to reshape the dataset to have the Democratic and Republican vote count by county as a column instead of separate rows. In this dataset, the counties were identified with a FIPS id code.

Now, we knew that we had to merge the datasets by county. However, the MIT Election Lab data used a FIPS code, whereas the Census Bureau used a longer code that included the FIPS code at the end of the string. We realized that the FIPS code always appeared after the substring "US" in the Census dataset. So, we extracted that FIPS substring by deleting everything before and including the "US" substring. Later, we merged the two datasets by using the FIPS code as the id variable. We only left the counties that were included in both datasets. In the end, we were left with a total of 3114 observations/counties. To make the dataset easier to work with, we used dplyr select() to remove every column except the features we were thinking of using as explanatory variables and the two columns that showed the Democratic and the Republican vote count, which resulted in a total of 26 columns.

We also want to add that Kerem was responsible for cleaning the MIT Election Lab dataset and the merging process, whereas Nikhil was responsible for Census Bureau's. We thought that this division of labor was fair given the complexity of the census data.

Feature engineering

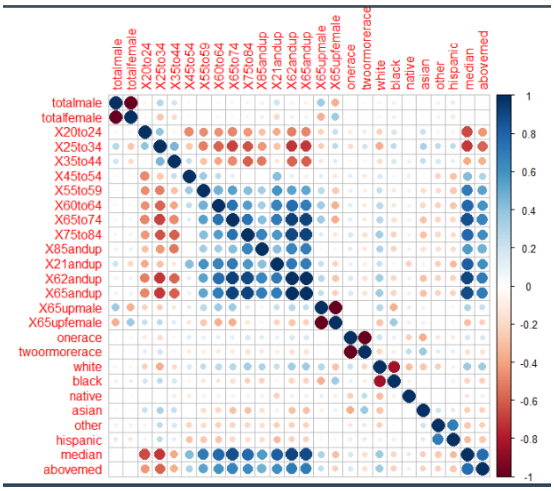
Given that we were treating our research question as a classification problem, we had to have a binary outcome variable. However, our vote counts were numeric. To solve this issue, we made an ifelse() statement to create a new variable called "dem." If a county had more votes for the Democratic Party than for the Republican Party, then the dem value was 1. If a county had fewer votes for the Democratic Party than for the Republican Party, then the dem value was a 0.

In the end, this feature we engineered enabled us to try and run the algorithms that we had in mind. As the feature that we engineered was the dependent variable, it surely was useful in the context of our project. It is possible that we could have treated this problem as a regression where the outcome would have been the vote count. However, we specifically wanted to focus our research question on identifying what makes a county “Red” or “Blue.”

We finished this part by deleting the Democratic and Republican vote count columns for convenience. In the end, we had dimensions of 3114 observations (each representing a country) and 25 columns as a percentage of total population: male, 20-24-year-old, 25-34 year old, 35-44 year old, 45-54 year old, 55-59 year old, 60-64 year old, 65-74 year old, 75-84 year old, 85 year old and up, 18 years old and up, 21 years old and up, 62 years old and up, 65 years old and up, 65 years old and up (males), 65 years old and up (females), one race, two or more races, white, black, native, Asian, Hispanic, other, and also median age in years. Luckily, all of the columns were registered by R in the classes that we were expecting them to be in. However, we changed the class of the “dem” column to a factor because some functions require the outcome to be a factor for classification tasks.

Exploratory Analysis

Before building our model, we decided to explore our dataset, which we began by looking at the correlation matrix between our features. From our correlation matrix and its plot, we identified that certain variables had an absolute correlation of above 0.70, which was the cutoff we decided to use. This got our variable size down to 19.



Then, we looked at the summaries for some of the variables that we had. All in all, there were slightly more males on average among all counties. The white percentage was about 84%, the black percentage was about 9%, and the Asian percentage was about 1.3%. The median age was approximately 41 years old. The 65 and up male percentage was below 50%, which means that there were more women on average than men at that age range.

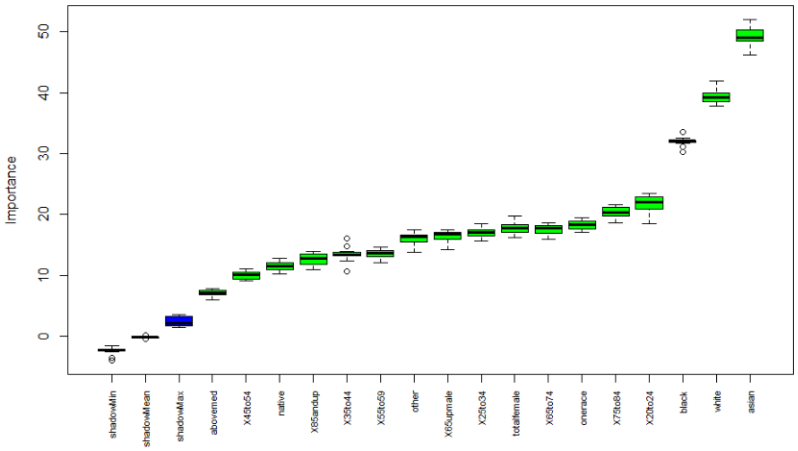
We also had to take a look at the distribution of our two categories. There were 2624 Republican counties and 490 Democratic. This was in line with our expectations because it is common knowledge that the distribution of Democratic voters is highly concentrated in specific dense areas of the US. Roughly speaking, the distribution of our classes was 86% to 14%. Even though there is an imbalance between our two classes, this is not at an extreme level; therefore, we have decided not to rebalance the dataset. Additionally, caret's classification metrics also provide useful information that accounts for the class imbalance.

Furthermore, we were not able to find any NA values in our dataset. This is no surprise given that the MIT Election Lab and Census Bureau are reputable sources that have a wide audience, including people without any statistical background. Hence, it is likely that they have cleaned the data before publishing it publicly. We also checked the ranges and the data types of

each column, and they all seem to be plausible. However, we had to alternate the class of our outcome variable because some functions require the response to be a factor when doing classification.

Feature Selection

After finishing our exploration of the dataset, we moved on to finding features to include in our model. For this purpose, we used two separate techniques: stepwise regression and the Boruta library (which provides an automated feature selection method). You can find the plot of the Boruta output below:



According to its developers, Boruta "finds relevant features by comparing original attributes' importance with importance achievable at random, estimated using their permuted copies (shadows)." As evident from the plot above, variables about the race breakdown of a county were more important than those about the age distribution. The top three features included "Asian," "black," "white," "20 to 24 year old," and "75 to 84 year old." From our personal experience as political science students, it makes sense to see that these two age groups might play an important role in deciding if a county will be Red or Blue.

Next, we ran stepwise regression (both forward and backward) with all of our variables. According to the AIC values, the optimal logistic regression model one was with the following variables: female, 20 to 24, 25 to 34, 55 to 59, 65 to 74, 75 to 84, 65 and up (male), one race, white, black, native, other, and median age. The AIC value for this model was 1632.

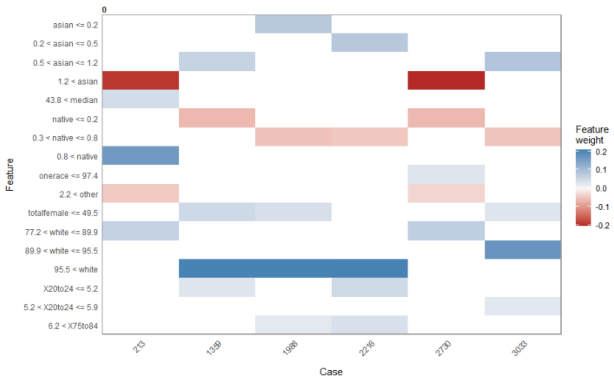
Model Building and Validation

With the aid of StepAIC and Boruta outputs, we used caret to try several models with three different classification algorithms: Logistic Regression, Naïve Bayes Classifier, Decision Tree, and KNN. We tried all four algorithms with three different sets of explanatory variables: all variables, the variables produced by the stepAIC output, and the top five Boruta variables. In all cases, our classification accuracy was close to 0.90, and the kappa was around 0.58. For our project, the Kappa value is very valuable given the imbalance of our dataset. Although there is no fixed way to interpret, Landis and Koch (1977) put 0.61-0.8 Kappa as a substantial agreement. Fortunately, the best duo of accuracy and Kappa value we got was with a K-nearest neighbors algorithm where the k was equal to 7, and we used all variables in our dataset. The highest combination of accuracy and kappa we were able to get was 0.9141189 accuracy and 0.6298046 Kappa. To further investigate our algorithm's performance, we also looked at the p-value of our accuracy being greater than the no-information rate. The p-value was 5.294e-05, which is much below the 0.05 threshold. Therefore, our model was much more useful (with statistical significance) than predicting by constantly choosing the larger category. This statistic is an important indicator of the uncertainty of the model because the imbalance of the dataset would have caused our model to have a good classification accuracy regardless of the model's quality.

We would also like to mention that the metrics that we have discussed above were achieved under repeated cross-validation. We have used caret' repeated k-fold cross-validation feature, for which we set k = 10. This function split our data into 10 groups and repeated it several times. Ultimately, the metrics that it presented to us were their mean from several cross-validation attempts. The repeated 10-fold cross-validation method was the best way to validate the robustness of our model on unseen observations, and it did well. Hence, our model does well regarding the bias-variance tradeoff, and our model is likely predicting based on an observable pattern rather than random chance. Once MIT Election Lab releases county results for the 2020 Election, we would also validate our model on them in the future.

Mining for Insights

Given the number of features and the complexity of the KNN algorithm, we tried to explain why our classification model was classifying an algorithm in a certain direction. To accomplish this task, we got help from the LIME package, which tries to explain individual predictions but cannot be used for getting insights regarding predictions for all observations in any given dataset. To get it started, we gave LIME our KNN model and then requested to have explanations for 6 randomly selected predictions (which we got with the sample() function) by using five features. The following plot summarizes the result.



Unfortunately, we observed that LIME used even a single feature several times to explain the classification. Hence, it was very difficult for us to come up with insights that were not local but instead could be generalized for the whole of the dataset. Instead of giving up, we decided to use a different algorithm than our KNN one with the highest prediction. Instead, we went back to the logistic regression model with the top five features from the Boruta output. After doing some trial and error of trying out several combinations of the top 5-6 features, we were finally able to end with a logistic regression model that had a statistically significant p-value regarding accuracy> No Information Rate and a slightly lower accuracy & kappa to our KNN model (under repeated 10-fold cross-validation). We felt that this model was good as it significantly boosted the interpretability aspect without damaging the accuracy much. To be specific, the accuracy was 0.90, and the kappa was 0.56. Here, you can see the summary of the GLM output.

```
Call:
glm(formula = dem ~ X20to24 + X75to84 + white + black + asian,
    family = "binomial", data = reduced_final)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.6060  -0.4244  -0.2648  -0.1985   2.9839

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.498256   0.610368   2.455   0.0141 *
X20to24      0.089898   0.022637   3.971 7.15e-05 ***
X75to84     -0.151748   0.057036  -2.661   0.0078 **
white       -0.051266   0.006344  -8.081 6.43e-16 ***
black        0.014033   0.006237   2.250   0.0245 *
asian        0.475458   0.039512  12.033  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2710.8  on 3113  degrees of freedom
Residual deviance: 1734.6  on 3108  degrees of freedom
AIC: 1746.6

Number of Fisher Scoring iterations: 6
```

The odds of a county being majority Democratic were increased by a multiplicative factor of exp(0.089898) for each percent increase in the population of the county that was between 20 and 24 years old, keeping other variables constant. Therefore, there is a positive association between a county being Blue and the number of people in the age range of 20-24 as a proportion of the county population.

The odds of a county being majority Democratic were decreased by a multiplicative factor of exp(-0.151748) for each percent increase in the population of the county that was between 75 and 84 years old, keeping other variables constant. Therefore, there is a negative association between a county being Blue and the number of people in the age range of 75-84 as a proportion of the county population.

The odds of a county being majority Democratic were decreased by a multiplicative factor of exp(-0.051266) for each percent increase in the population of the county that was white, keeping other variables constant. Therefore, there is a negative association between a county being Blue and the number of white people as a proportion of the county population.

The odds of a county being majority Democratic were increased by a multiplicative factor of exp(0.014033) for each percent increase in the population of the county that was black, keeping other variables constant. Therefore, there is a positive association between a county being Blue and the number of black people as a proportion of the county population.

The odds of a county being majority Democratic were increased by a multiplicative factor of exp(0.475458) for each percent increase in the population of the county that was Asian, keeping other variables constant. Therefore, there is a positive association between a county being Blue and the number of Asian people as a proportion of the county population.

Literature Review

To validate if the findings above were due to chance or not, we decide to examine external sources to identify their validity. According to an NPR article (2020), whites have always voted more for Republican Presidential candidates since the 1960s. The last US President to win the white vote as a Democrat was Lyndon Johnson. Hence, it makes sense that a county

with a lot of white individuals tends to vote more for a Republican President. Likewise, a US-wide survey by the Pew Research Center (2020) also shows that whites tend to vote more Republican, whereas black and Asian people tend to vote more Democrat. Therefore, it makes sense that a positive association exists at the county level between the Black/Asian population as a percentage of county population and being Democrat. It also makes sense that a negative association exists at the county level between the White population as a percentage of county population and being Democrat.

We also checked to see if our findings of the age breakdown were validated by external sources. Luckily, Pew Research Center had also done surveys regarding party identification and age demographics. People who were born in the silent era (1928-1945) tended to identify as Republican more often than not. The opposite was true for Millennials (who would have been 20-24 years old during the 2016 Elections) because they substantially identified as more Democrat.

The information from NPR and Pew Research Center strongly supports the insights that we have obtained from our logistical regression model. We should also consider that the Census Bureau provides an almost complete sample of the population, meaning that this is an extremely representative sample. Hence, we have no reason to believe that our findings are due to luck.

Discussion about Iterations and Data Snooping

Almost everything went according to the plan in our project. We tried everything we thought about initially. The only time that we had to have a change in plans was we were not really able to find any generalizable insights using LIME on our KNN model, as we had discussed during office hours. Unfortunately, we had to use a different and more interpretable

model (a logistic regression one) to come up with insights from our dataset. We also tried a few different combinations of variables – in accordance with the Boruta and StepAIC output – to see what we can do.

We do not think that there any significant problems regarding data snooping. At the beginning of our planning process, we understood that we might be unable to identify any significant insights, as this is common in all kinds of scientific research. We presented the dataset in its entirety without dividing it into subgroups by region (which is a common strategy in p-hacking). We have documented our procedure of finding relevant features to use that would not be hard to interpret. The only part of this project where we played with the features and the algorithms was when we were trying to achieve a high prediction accuracy and kappa. However, as previously discussed, the insights that we have mined do not come from the model that had the best accuracy. Additionally, according to our literature review, our findings were supported by data presented in surveys of Pew Research Center, which is very reputable in the social sciences field.

Conclusion

In this project, we looked into the relationship between demographic factors in a county and whether the county voted for a Democrat or a Republican in the 2016 US Presidential Elections. We specifically focused on age and race demographics for 2015 because the data collection for the Census Bureau would not have been finished by the time the 2016 Elections took place. Our analysis revealed that certain demographic factors are, in fact, influential in predicting whether a county was “Blue” or “Red.” We found out that an increase in the percentage of very old people and white people increased the odds that a party would have more

votes for Trump than Clinton in any given county. On the other hand, we found out that an increase in the percentage of young adults (20-24 year old), black people, and Asian people increased the odds that a party would have more votes for Clinton than Trump in any given county. We also found out that our insights were supported by external sources, including NPR and the Pew Research Center.

In addition to these insights, we were able to utilize demographic factors to predict whether a county voted more for Trump or for Clinton using the classification algorithm known as K nearest neighbors, which tries to predict the features of an observation by examining the features of the data points that are similar to it. Our prediction accuracy was statistically significant even considering the inflation caused by the class imbalance. In simple words, our model was able to predict each county's voting tendency with an accuracy that was better than educated guessing, which would have resulted in 84% accuracy (always picking the category with the highest amount), whereas ours was 92%.

Overall, this project provided insights into how demographics affect Presidential voting preferences and how this information can be used by electoral forecasters. We added value to our client Pew Research Center by confirming their findings with data that better represented the US population than the samples used in their surveys. We believe that we added value to their survey findings by decreasing their uncertainty. For the prediction aspect, we hope that the MIT Election Lab will soon release the county-by-county results for the 2020 election so that we can see how our classification model will perform. It should be noted that our models were trained using data specific to Donald Trump and Hillary Clinton, so we cannot be sure if our results are generalizable to all Republicans/Democrats or all elections, or just to the 2016 election. However, even the current model is promising enough to provide value for election forecasters,

who may want to incorporate our model into their methodology to make better predictions (hence, increased reward).

If we were to do this project again, the main thing we would have done differently would be to use more datasets from the US Census Bureau. The 5-Year Estimates include data on other topics such as income and religion. It would be a good idea to see if they would increase the prediction accuracy or provide any other interesting insights on voting behavior.

Critique of Isaac and Begum's Project

The purpose of this project was to gain an understanding of users' online expression of depressive/anxious behavior on social media. A classification model was trained using Facebook comments labeled by NIH researchers and then used to classify a sample of tweets from users as depressive/anxious or not. Afterward, analysis was done to explore whether users' Twitter expression of depressive behavior could be mapped to CDC survey data on anxiety and depression, helping to clarify if mining social media sites like Twitter can be useful tools in measuring the prevalence of mental health conditions.

While this was already addressed in the write-up, we think that the non-overlap of the dates between the CDC and the Twitter data is significant. Seasonal depression affects a lot of people, but the Twitter data excluded lots of data from earlier winter dates that the CDC data did include. This definitely would have affected the results, in our opinion. Aggregate depression scores for data collected in the spring should be lower than that of data collected in the winter.

In terms of positive feedback, we thought that the idea of using Islam et al.'s dataset to build the model was a good choice. Twitter (and all social media) is filled with sarcastic data, and simply searching for keywords would have often resulted in data that did not consider

context or tone or errors such as spelling mistakes. The write-up commented on how the random forest model was complicated to understand, but we think that for a predictive model, this is fine, and it is important to prioritize models that have high predictive accuracy. All in all, this was a very well done project that had valuable insights for a critical topic.

Works Cited

Demby, Gene. "Who Is The White Vote?" *NPR*, NPR, 5 Nov. 2020, www.npr.org/2020/11/05/931836604/who-is-the-white-vote.

Pew Research Center. "1. Trends in Party Affiliation among Demographic Groups." *Pew Research Center - US Politics & Policy*, Pew Research Center, 28 Aug. 2020, www.pewresearch.org/politics/2018/03/20/1-trends-in-party-affiliation-among-demographic-groups/.