Introduction

In this case study, we will attempt to develop a model that to predict whether the United States House of Representatives will remain in Republican control or transition into Democratic control. We will develop another model to predict the outcomes of all the North Carolina Congressional elections, which will be the 13 federal Representatives to Congress. We will construct these models with various datasets from previous presidential and midterm elections and from polling data from the previous and the current year.

Methods

We needed to produce two different models for predicting House election outcomes per congressional district and for predicting the total number of seats overall in the US for Democrats and Republicans. We used the census data for demographics of each county and congressional voting district and merged it with the voter registration data from the past 12 years of elections. These two models would be used to predict turnout for each year. We would then use the predictions from the previous year to predict the voter turnout for the preceding election year in order to check the accuracy of our prediction. Using this method, we could then proceed to predict the outcome and the political split of seats for the 2018 election. We chose between three different models, one using the glmer regression, another with a random forest, and finally one using XGboost for our predictions. In order to make our predictions as accurate as possible, we constructed datasets from other election databases. For example, in one dataset, we have the financial details of the campaigns for all candidates who ran in election starting from 2000 to present day with the incumbency seat, the winner of each election, and the total proportion of votes for each candidate.

Description of Datasets

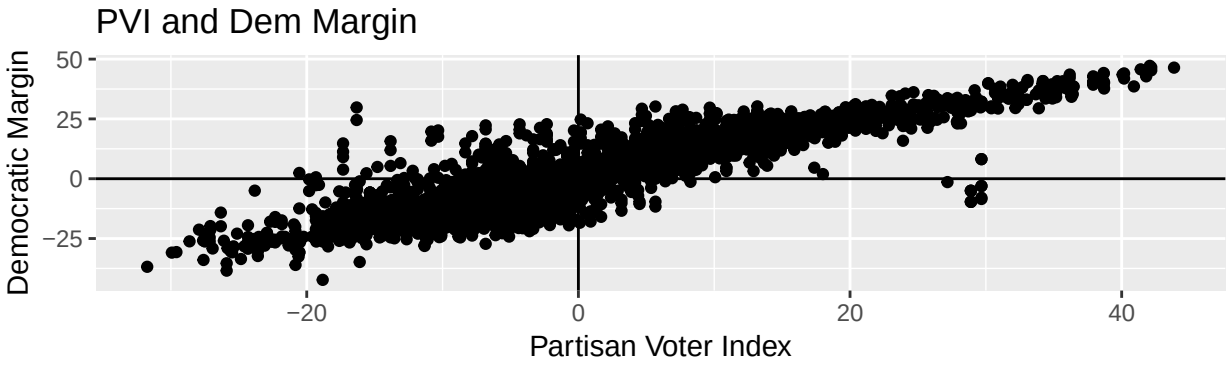
- reg_stats and voter_stats: To begin we had to read in multiple years of data sets that provided us with information on the number of voters who were registered in each county within North Carolina (reg_stats). The data set included the number of registered voters (total_voters) given certain characteristics such as their county (county_desc), the date of the election (election_date), their race (race_code), their ethnicity (ethnic_code), their sex (sex_code), and age range (age). In addition, we read in multiple years of data_sets looked at the number of voters who actually voted in each county within North Carolina (voter_stats). The data set informed us on the number of voters (total_voters) given defining characteristics such as their county (county_desc), their age range (age), their party (party_cd), their race (race_code), their ethnicity (ethnic_code), their sex (sex_code), and the party they voted for (voted_party_cd).
- Daily Kos Elections 2008,2012,2016 Presidential Election Results: This dataset describes all congressional districts throughout the US. We will use the incumbent candidate and their party association along with the vote shares of the past presidential elections from 2008, 2012, and 2016 to predict which party will most likely win in each NC congressional district, as well as who would win the House overall. These three presidential elections include Trump vs. Clinton, Obama vs. Romney, and Obama vs. McCain. This is a simple dataset, but it is extremely useful for our model prediction because it gives all the vote share percentages for every congressional district that was identified in these years. We can use this information to more accurately predict candidate winners for this coming election for North Carolina because we will be able to see if each congressional district in North Carolina tends to vote for the same party as the current and previous presidential candidates. Additionally, we will be able to see the distributions of these elections. Also, we assume we are only looking at Democratic and Republican party candidates in this dataset, since no third or other party candidate was a major candidate in any of these presidential elections.
- Histr_Congr_final: This dataset is a collection of House election from 2000 to 2016. Each year has the distribution of republican seats held and democratic seats held in that year. There exists a fund

multiplier (democratic funds/republican funds) and the party of the President that was in office in the specified year. Other variables include the effective rate, rate of unemployment, nominal and real GDP, GDP growth rate, and the distribution of vote shares by democratic and republican parties and the proportions of those votes in separate race columns. These numbers were found from opensecrets.org.

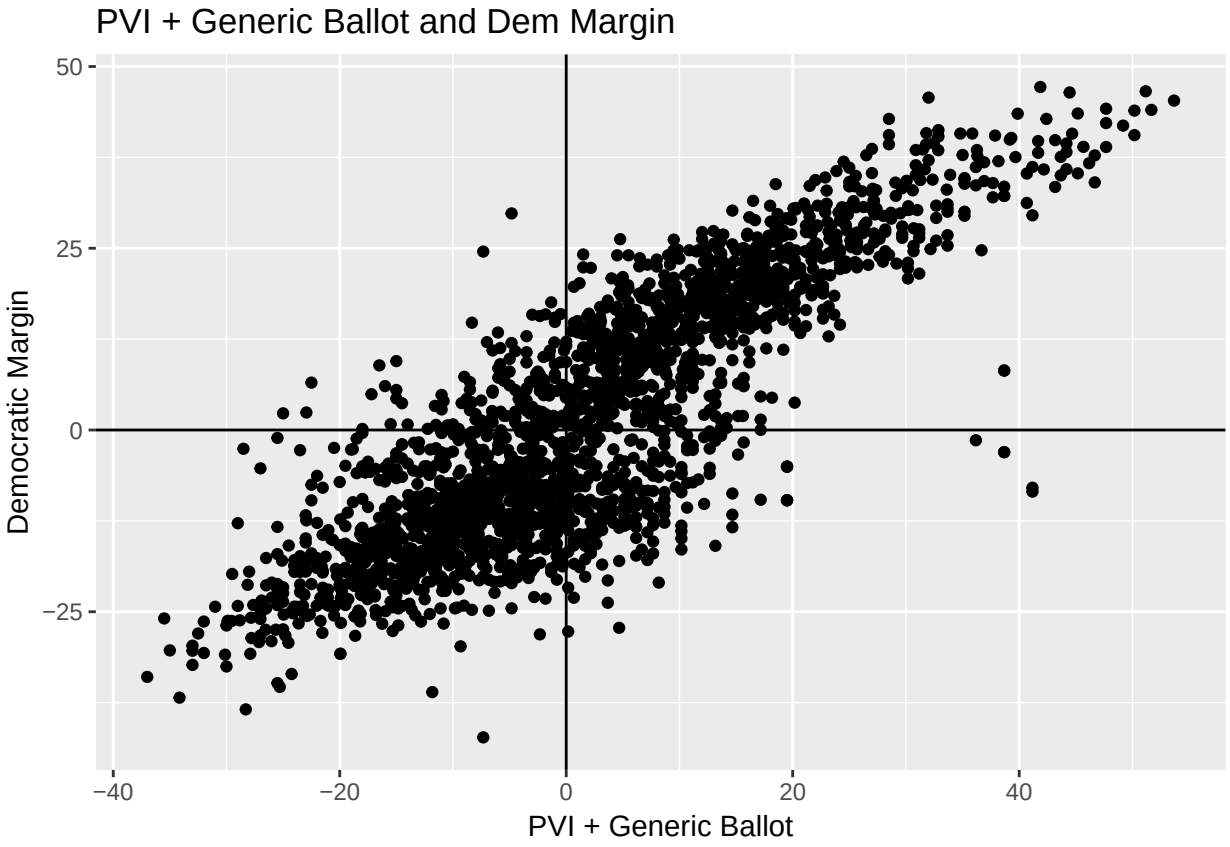
- trump_approval, obama_approval, bush_approval: All three of these datasets were found from https://gallup.com. The datasets are in the same format. There is a result that is based on a three-day rolling average. From the dataset, we are given the percentage of Americans who approve, disapprove, or have no opinion on the job of the president. There results were based on telephone interviews with about 1,500 adults nationwide. Along with the approval percentages of Americans. There are also approval percentages from the Democratic and Republican parties based on the three-day rolling average. We consider approval ratings as an important variable to accurately predict House candidate winners because if more Americans believe that the President is doing a good job, then they will more likely vote in support of the party and vice versa.
- national2000, national2004, national2008, national2012, national2016: The partisan voting index is the measurement of how each congressional district performs at the Presidential level compared to the nation as a whole. This is vital for House predictions throughout the whole US because with knowledge of how likely a district is going to be Democratic or Republican it will be much easier to predict how many seats throughout the country will be democratic and how many seats will be republican. We will be able to filter our the third party votes because we are given in the dataset the third party and other vote distribution. With more accurate proportions for democratic and republican vote shares during each presidential election, we will be able to focus and rely on democrats and republicans rather than on the small percentages of the third party and other party vote shares.
- table-03: This dataset gives the detailed years of school completed by people that are 25 years and over and grouped by sex, age groups, race and hispanic origin from 2017. The education ranges from first grade and high school diplomas to PhDs and Master degrees. From the previous case study, we have seen that voter turnout for younger adults and young voters are much lower than those of older adults. Since voters who are older tend to vote more in midterms and presidential elections, it is important to see which parties they tend to vote for and which congressional districts lean more to the democratic or republican candidates usually. If we can match up this education data with that of the census, then we can get a population estimate for all the people and more accurate predictions when we find an association with party and education status of Americans.

EDA

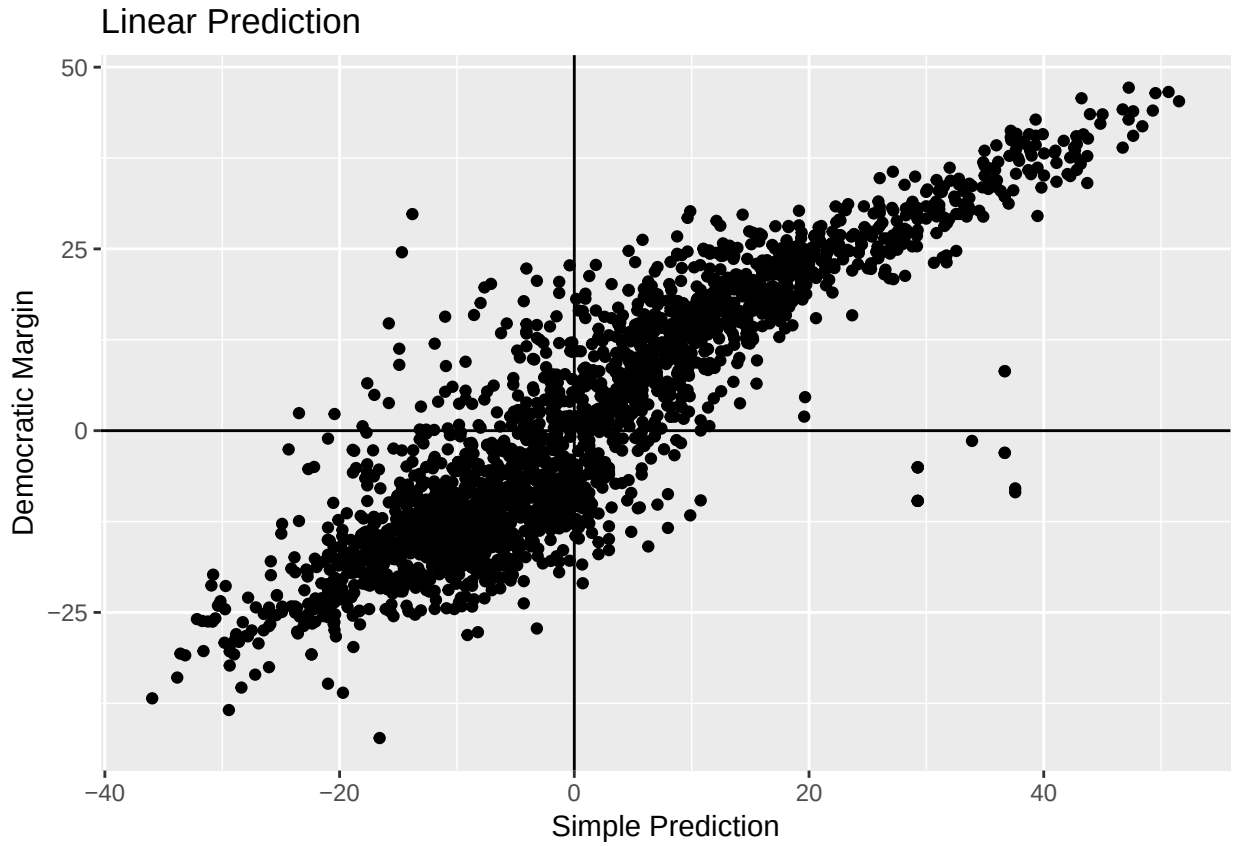
We decided to dive further into utilizing our Daily Kos Election datasets, namely after reading in the data sets, we did our own analysis on the PVI, the Partisan Voting Index of each county from each State in the past. To do this, we merged together all the Kos datasets, and then for each year that we had data, we calculated the baseline partisan split at the national level based on the expected percentage of democratic votes. Then using inspiration from the Cook political report, we decided to calculate the PVI for every two years starting from 2008 to 2018, inclusive. For all the years except 2012, we first found two values. First, for each county the democratic percentage of each county in excess of the baseline for that year from the previous election, and second, the democratic percentage of each county in excess of the baseline for that year from two elections ago. These two values were then averaged. So for example, the PVI of 2018 would be the average of the democratic percentage for each county in excess of the 2016 baseline in the 2016 election and the democratic percentage for each county in excess of the 2014 baseline in the 2014 election. As for 2012, since there was inconsistent data regarding district lines for 2008, the PVI was based solely off the the democratic percentage for each county in excess of the baseline in the 2010 baseline in the 2010 election. These PVI's were examined and were in tune with our previous findings, and are utilized moving forward in our models.



We found that by just using the Partisan Voter Index, we were able to correctly predict 88.8408304% of the elections correctly, with the majority of incrrect predicitions occurring in districts where PVI was close to 0. We attempted to add in thethe effects of the national generic ballot to see if that improved the predictions.



Just adding together the generic ballot margins and the PVI actually gave us a worse prediction, only correctly predicting 83.9532872% of elections. However, once we built a simple linear model using generic ballot and PVI, our results improved greatly.



While the actual accuracy only rises to 88.884083%, the error in each predicition is reduced greatly.

Modeling and Prediction Strategy

The first step we took when predicting seat numbers for the 2018 Congressional House election was to merge the datasets by year and party affiliation. To predict the number of seats for the democrats and the number of seats for the republicans we split the dataset and made two-thirds of it the training set and the other one-third into the testing set. For our first training model, we made a Random Forest model which was built on the following variables at the time of the election: the number of Republican seats, number of Democratic seats, the Effective Federal Funds rate, whether or not the election includes a vote for President, who currently controls the house, the GDP growth over the past year expressed as a percent, the unemployment rate, the amount Democrats have fundraised divided by the number the Republicans have used, the most recent electoral votes for President by each party and the political party of the President.

We built a model that subsisted of these more "macro" variables because we observed a few trends that suggested the outcomes of the election might be better observed from a national viewpoint than from a district-by-district one (we also did district-by-district analysis). For example, after Presidential election years, there was almost always a strong overcorrection – note for example, in the 2010 House of Representatives elections, the Democrats lost 63 seats two years after the election of Barack Obama. Poor economic conditions could also cause a voter to look for a change of guard in their elected representation. However, in this case such an assumption might hurt our prediction as, despite record low numbers in unemployment, many analysts predict a "Blue Wave" of democratic voters this election.

Overall though, the model reported a high Mean Squared Error value of 717.

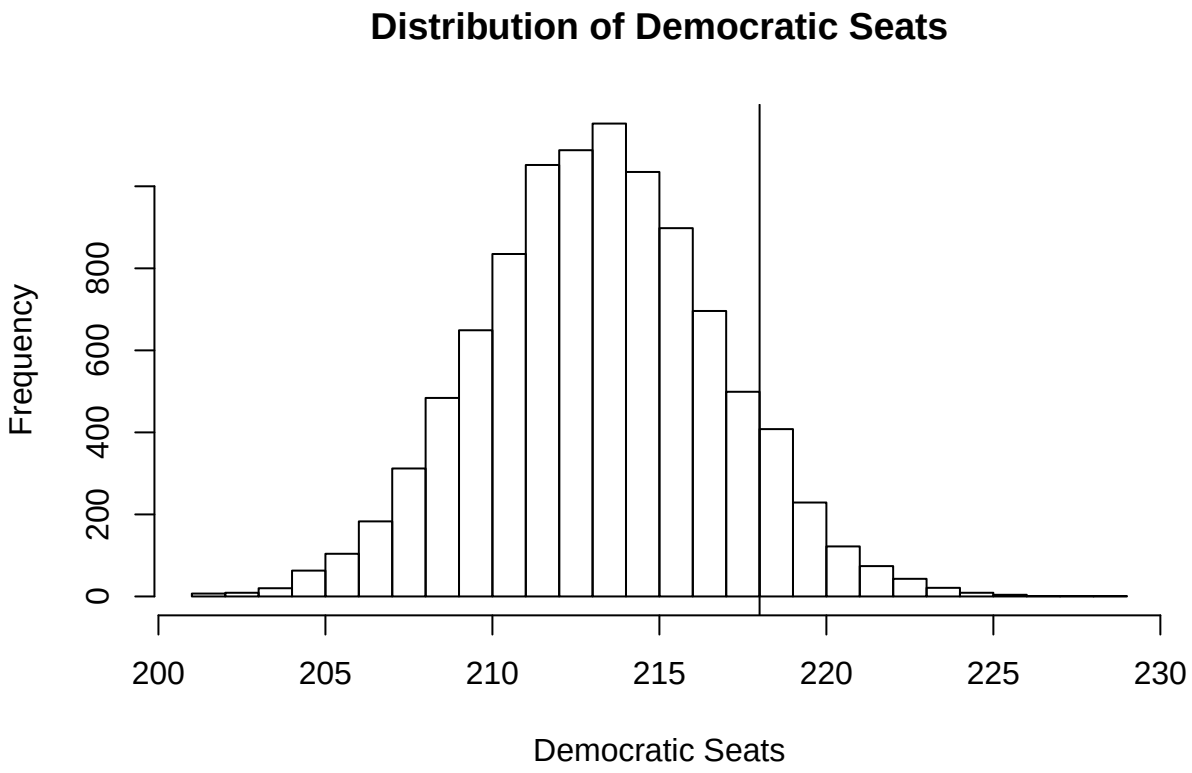
Our second model that we created was with the XGBoost package in R. It was similar to the Random Forest model that we created in terms of splitting the dataset into a testing set and a training set, but the difference

was that we included more variables and data in the model. Also a main difference was that for this model, we looked at each individual congressional district in the United States in 2018 and the previous years. From the output of this model, we predicted that the democratic would hold 204 seats and the republicans would hold 231. That did not seem reasonable for this election since we were originally expecting the democrats to hold more seats than the republicans from the midterm elections. We proceeded by creating a more accurate XGBoost model. The variables that were included in this model were incumbency status, partisan voter index, party of president, the spread in national generic ballot, results of previous elections, the spread in presidential approval, and whether the year was a presidential year or not. For point estimates and confidence intervals, we extracted the NC congressional districts from this model and used them for our estimation of the midterm election just for North Carolina.

Results

```
## Joining, by = "raceYear"
## Joining, by = "Year"
## Joining, by = "Year"
## Loading required package: lattice
##
## Attaching package: 'caret'
## The following object is masked from 'package:purrr':
##
##     lift
##
## Attaching package: 'xgboost'
## The following object is masked from 'package:dplyr':
##
##     slice
## Loading required package: xml2
##
## Attaching package: 'rvest'
## The following object is masked from 'package:purrr':
##
##     pluck
## The following object is masked from 'package:readr':
##
##     guess_encoding
##
## Attaching package: 'httr'
## The following object is masked from 'package:caret':
##
##     progress
## Joining, by = "State"
## Joining, by = "cd"
## Joining, by = "Year"
```

```
## Joining, by = "State"
## [1] 206
```



The following table delineates the predicted democratic vote percentage for each of the 13 districts within North Carolina. We also include the predicted upper and lower bounds of the vote percentage based on the confidence interval from the results of our model.

cd	predDemVote	upper	lower
NC-01	74.18807	83.88812	64.48802
NC-02	42.43060	52.13065	32.73055
NC-04	72.98545	82.68550	63.28540
NC-05	40.56592	50.26597	30.86587
NC-06	42.48783	52.18788	32.78778
NC-07	43.86128	53.56133	34.16123
NC-08	42.48783	52.18788	32.78778
NC-09	43.58943	53.28948	33.88938
NC-10	35.64699	45.34704	25.94694
NC-11	35.95982	45.65987	26.25977
NC-12	75.36498	85.06502	65.66493
NC-13	44.18491	53.88496	34.48486

Discussion

Overall we feel that our model predicts fewer Democratic house seats than we expected when compared to reliable election predictions like Nate Silver's 538. We postulate several reasons why the numbers seem low below:

The dataset that we used goes back to 2006 for the district-level model and 2000 for our "macro" model and the democrats have done poorly in every single house election since 2006, so there is not much data for democrats holding more seats in elections and this would likely bring our overall prediction down. For the district-level model, we calculated a Partisan Voter Index based on the past two presidential elections and we only had historical presidential election data from the year 2000 onwards.

Furthermore, we were unable to include in the "macro" model enough variables that would indicate the "Blue Wave" that many analysts are predicting. One variable we hoped to include was voter enthusiasm, a measure of how excited a respondent to a survey was to vote in this election compared to years past, but that information was only available for midterm elections and would thus not be robust enough to help in the creation of the model. A recent Gallup poll showed that voter enthusiasm for this election is the highest it has been in over 20 years and that would perhaps tilt our model in favor of the Democrats.

We also wished to include polling data in the creation of our model, but we were unable to find historical polling data that would allow us to build a model and incorporate polling data from this current election season.

Early voting data could also have been incorporated into our model, but we ran into similar problems in that historical early voting data is hard to access. Another issue with early voting data is that rates of early voting have increased over previous years and it may be hard to compare from year to year.

Appendix

```
library(tidyverse)

library(randomForest)



congr_results <- read.csv("Histr_Congr.csv")

congr_results %>% select(-c("Year", "Rep_Seats", "Rep_Seats_prev")) -> congr_results



dem_seats_2018 <- 194

fund_ratio <- 1.507296

info_2018 <- c(dem_seats_2018, fund_ratio, 0)



changeYN <- function(x) {
  if (x == "y") {
    d = 1

  } else{
```

```
    d = 0

  }

  return(d)

}

congr_results$Pres.Year <- sapply(congr_results$Pres.Year, changeYN)



rf1 <- randomForest(congr_results[, -1], congr_results[, 1])

predicts <-
  predict(rf1, info_2018) # dem_seats for 2018 predicted to be 20



xgb_tune2 = train(
  outcome ~ DemStatus + PVI + pres_party + prevDemVote + pres_year + generic_spread,

  data = train,

  method = "xgbTree",

  trControl = cv.ctrl,

  verbose = T
)



sqrt(mean((
  as.numeric(test$DemVotesMajorPercent) - predict(xgb_tune1, test)
) ^ 2))

mean(predict(xgb_tune2, test) == test$outcome)



sum(predict(xgb_tune2, test) == "dem") +




  ## Neural Network



  library(brnn)
```

```
brnnGrid = expand.grid(neurons = 8:11)



brnn_tune1 = train(
  DemVotesMajorPercent ~ .,

  data = train,

  method = "brnn",

  trControl = cv.ctrl,

  tuneGrid = brnnGrid,

  verbose = T
)



brnn_tune2 = train(
  outcome ~ Year + DemStatus + PVI + pres_party + generic_spread + prevDemVote,

  data = train,

  method = "brnn",

  trControl = cv.ctrl,

  verbose = T
)
```