**Part 1: Exploratory Analysis**

**Basic Statistical Analysis and Data Cleaning**

**Data Merging and Cleaning:**

We were left with a number of datasets after Project 1:

1. Election Results (From New York Times (NYT) and Federal Elections Commission (FEC))
   1. FEC 2004
   2. FEC 2006
   3. FEC 2008
   4. FEC 2010
   5. FEC 2012
   6. New York Times 2014
2. Candidate – Industry Connections
3. S&P Financial Data

Each of the above data sets needed to be combined before we could begin to address some of our analytical questions, begin generating descriptive statistics, and test hypotheses. For our group, this was a very large undertaking, with almost every group member doing a piece of it. We broke this into pieces, so first the election results files were stacked, and then the different types of file were merged.

For the election results, the FEC Data came in human-readable spreadsheets from the FEC and the NYT data was scraped. The FEC data needed to be converted to a machine-usable form then stacked, and finally collapsed. The NYT data needed to be reshaped. Once the FEC and NYT data sets were in similar shapes, we had to coerce the variables into the same formats so that when we stacked them to get the full year range we were seeking, the variables would be continuous and appear to have come from the same original file. This was particularly difficult due to the idiosyncrasies of the data, such as the format that each data originator used to store edge cases, such as races where a single candidate ran unopposed.

Although we originally collected two different sets of data from Open Secrets, we decided that we only needed the information from one of them, therefore we were able to just clean up the one data set with candidate – industry connections.

The next step was to merge the election results (candidate performance during the election) with candidates’ funding sources. This was also a considerable effort, beginning with yet more variable cleaning and cajoling to line up between the two files, and ending with a fuzzy merge between the files by name. In order to avoid merge mistakes, this merge was performed using a tiered strategy. For the first merge, the strictest merge rules were applied; among the use of other identifying characteristics of candidates, for the first merge they needed to match by all of the election information (state-year-district) and by full name (first and last). For the second tier, the same criteria were used excluding a first name match. For the third merge, the same criteria as the first merge were used, excluding a last name match. And for the fourth merge, clerical errors in candidate names were corrected so that they would merge. At every step in the merging process, the results were hand-checked to make sure no candidates were merged incorrectly.

Tim did the S&P financial data.

We ended up with two final datasets, to be used in separate analyses. The first dataset, PoldataSPIndustries, consists of, for each candidate/year/industry level observation from every election cycle from 2004-2014, the candidate’s political party (party); campaign contribution amount (amount) and percentage of total contributions (industrypercent) that come from the industry; total campaign contributions (candtotal); incumbent status (incumbent); number of votes received (votes) and percentage of votes received (percent; number of votes divided by total votes cast in the race); election winner status (winner); a variable illustrating how the industry’s contribution to the candidate compares to the amounts contributed by other industries (indrank), the total amount of funding all of the candidates in the race received (racetotal), and the percentage of the total race funding that the industry gave to the candidate (racefundperc).

The data we originally scraped from OpenSecrets.org sorted campaign contributions into 95 different industries; in order to compare this data to stock market performance, we sorted these industries into the 10 sectors of the S&P 500[[1]](#footnote-1), based off of descriptions of the OpenSecrets industries found on OpenSecrets.org[[2]](#footnote-2). Industries which did not fit into an S&P sector were sorted into 3 additional categories; not for profit, not publicly traded, and other. After sorting the OpenSecrets industries into S&P sectors, we collapsed the dataset on S&P sector, adding up the contribution amounts from the OpenSecrets industries contained in each S&P sector.

The second dataset, PoldataSPIndustriesStockData, in addition to all of the data in PoldataSPIndustries, contains data on stock market performance for each of the sectors in the S&P 500, for each election cycle from 2004-2012 (yrpercentchange). We calculated performance for each S&P sector by averaging the change in the adjusted closing price (variable adjclse) of the S&P stocks in each sector over each election cycle. The 2014 election cycle had to be excluded from any analysis of the stock data, because we didn’t think a metric based on the 9 months of data from the 2014 cycle that were available at the beginning of the project would be comparable to the metrics based on 24 months of data in the other election cycles. Since we still wanted to analyze the full political dataset, we decided the best approach would be to keep that dataset, and create a new one to look at the stock data.

**Summary Statistics:**

We had about the same number of observations for every year in both datasets; as discussed above, the dataset with stock data doesn’t have any observations for the 2014 election cycle. There were significantly more Republican candidates than Democrats in both datasets, and both contained a small but not insignificant number of Independent candidates. At first we were surprised that our datasets contained so many more winners than losers, so many more incumbents than challengers, and so many more Republicans than Democrats. However, we realized that this was simply a result of having one observation per industry that donated to each candidate; Republicans tended to have more industries donating to them than Democrats, and winners and incumbents tended to have more supporting industries than losers and challengers, respectively. We verified this by shrinking the dataset down to unique year/state/race/candidate observations, and observing that the discrepancies vanished.

The PoldataSPIndustries dataset contains 35,082 year/state/race/candidate/industry level observations. In this dataset, the average amount contributed by one industry to one candidate was $122,100, but the standard deviation was nearly $300,000; some candidates received enormous amounts of funding from some industries, while others received very little. The same pattern was evident in the total amount of contributions each candidate received; the average amount was $865,900, but the standard deviation was more than $1.3M. Our analysis of the votes variable revealed that the average candidate received 194,000 votes, and that we had a number of NA observations for the number of votes. These NA’s came from uncontested elections; for uncontested elections, the FEC didn’t report the number of votes the candidate received. The average amount of funding per race was $1.7M, and the standard deviation was $2.6M, again showing that some races had much greater amounts of funding than others.

The PoldataSPIndustriesStockData dataset contains 20,248 year/state/race/candidate/industry level observations. It has fewer observations because we didn’t have stock market data for the 2014 election cycle, and because three of the industries into which we sorted the OpenSecrets industries; not for profit, not publicly traded, and other; are not represented in the stock market. The average contribution amount per industry was $70,380, and the standard deviation was about $180,000. Total contributions per candidate averaged $859,300, with standard deviation of $1.3M. The average number of votes was 206,700, and the analysis revealed that we still had a significant number of NA observations in the votes variable. Total race funds averaged nearly $1.7M, with standard deviation $2.5M. With regard to the new variables from the stock market data, the average adjusted closing value was 2773, and the average year percent change was 28%.

When we took a closer look at the candtotal and votes variables, we discovered that we had a lot of values very close to zero, and just a few values at the high end of the distributions. We decided that for our analysis, we didn’t want to look at the marginal candidates who only received a few votes or dollars. We also thought we should exclude some candidates at the top of the distribution, as they likely were special cases, and as such would exhibit different behavior from the middle-of-the-pack candidates we really wanted to look at. We also wanted to exclude any candidates whose ran in uncontested elections, for two reasons, because contributions data from these elections would also be systematically different from the middle-of-the-pack candidates. So, we removed any observations which had values greater than 1 interquartile range above the 75th percentile value for the vote percent or candtotal variables, as well as any observations which had values less than 1 IQR below the 25th percentile values of those variables. We also removed any observations for which we did not have voting results. For the dataset containing stock market data, we removed outliers based on the yrpercentchange variable. We did so because we believed that, in conditions of great economic turmoil, when an industry either rose or dropped a great deal, we would not be able to connect changes in the market value of the industry with the industry’s political contributions, since there were much greater forces at work causing the industry to move on the stock market. So, we removed any observations for which the value of the yrpercentchange variable was greater than/less than 1 interquartile range above/below the 75th/25th percentile values. After removing outliers and missing values, we were left with two datasets, containing 29,226 and 15,825 observations, respectively, summary statistics of which are listed in the Appendix.

Finally, it was important for the frequent itemset analysis to bin some of our numerical data into categories, so that we could see look at possible associations between the numerical variables and the other variables in our dataset. So, after outliers and missing values were removed, the total contribution amount and percentage of votes each candidate received in each election cycle was binned into categorical variables (candtotallevel and votepercentlevel, respectively) with four levels (very low, mid-low, mid-high, and high). These levels were calculated by dividing the total range of the variables into four equal segments, and sorting each observation into a segment. For candidate total funding, these bins were [$100, $431,000], ($431,000, $862,000], ($862,000, $1.29M], and ($1.29M, 1.73M], and for vote percent, these bins were [12.3%, 32.4%], (32.4%, 52.6%], (52.6%, 72.7%], and (72.7%, 92.9%].

Summary statistics before outlier removal are presented in Table 1; summary statistics after outlier removal are presented in Table 2.

**Table 1 - Summary Statistics Before Outlier Removal**

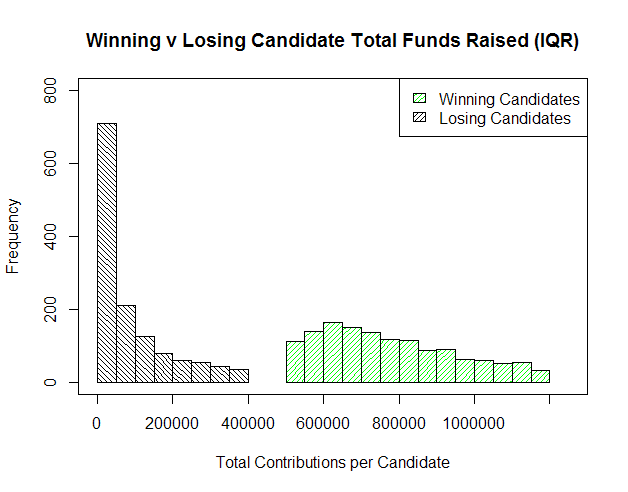
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **PoldataSPIndustries with outliers** | | | | | | | | |
|  | **Year** | **Count** | **Party** | **Count** | **Winner** | **Count** |  |  |
|  | 2004 | 5634 | Dem | 15539 | 0 | 14385 |  |  |
|  | 2006 | 5795 | Rep | 18284 | 1 | 20697 |  |  |
|  | 2008 | 5580 | Ind | 1259 |  |  |  |  |
|  | 2010 | 6184 |  |  | **Incumbent** | **Count** |  |  |
|  | 2012 | 5885 |  |  | 0 | 16022 |  |  |
|  | 2014 | 6004 |  |  | 1 | 19060 |  |  |
|  |  |  |  |  |  |  |  |  |
| **Variable** | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** | **Std. Dev.** | **NA's** |
| Amount | 10 | 10000 | 35950 | 122100 | 115600 | 8829000 | 296218.9 |  |
| Industrypercent | 0.001293 | 0.033 | 0.08094 | 0.1538 | 0.1984 | 1 | 0.1839325 |  |
| Candtotal | 10 | 198600 | 566300 | 865900 | 961700 | 21830000 | 1358704 |  |
| Votes | 5 | 88960 | 125400 | 194000 | 174300 | 7865000 | NA | 431 |
| Percent | 0 | 0.3915 | 0.55 | 0.5289 | 0.6602 | 1 | NA | 431 |
| Totalracefunds | 72620 | 584500 | 907700 | 1714000 | 1743000 | 32870000 | 2588232 |  |
| Racefundperc | 0.00001 | 0.1867 | 0.6948 | 0.5985 | 0.9842 | 1 | 0.3843594 |  |
|  |  |  |  |  |  |  |  |  |
| **PoldataSPIndustriesStockData with outliers** | | | | | | | | |
|  | **Year** | **Count** | **Party** | **Count** | **Winner** | **Count** |  |  |
|  | 2004 | 3937 | Dem | 8629 | 0 | 7717 |  |  |
|  | 2006 | 4021 | Rep | 11011 | 1 | 12531 |  |  |
|  | 2008 | 3883 | Ind | 608 | **Incumbent** | **Count** |  |  |
|  | 2010 | 4294 |  |  | 0 | 8762 |  |  |
|  | 2012 | 4113 |  |  | 1 | 11486 |  |  |
|  |  |  |  |  |  |  |  |  |
| **Variable** | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** | **Std. Dev.** | **NA's** |
| Amount | 49 | 8500 | 25500 | 70380 | 66280 | 6525000 | 178624.5 |  |
| Industrypercent | 0.001293 | 0.02676 | 0.05079 | 0.08256 | 0.1019 | 1 | 0.09603277 |  |
| Candtotal | 130 | 246300 | 570300 | 859300 | 949200 | 21830000 | 1324489 | 126 |
| Votes | 5 | 96600 | 135800 | 206700 | 182600 | 7865000 | NA | 126 |
| Percent | 0 | 0.4038 | 0.5629 | 0.5414 | 0.6676 | 1 | NA |  |
| Totalracefunds | 72620 | 568700 | 887100 | 1651000 | 1689000 | 32870000 | 2513277 |  |
| Racefundperc | 0.0000266 | 0.2452 | 0.7355 | 0.6194 | 0.9854 | 1 | 0.3750487 |  |
| Adjclose | 70.09 | 1487 | 2225 | 2773 | 3571 | 8184 | 1901.18 |  |
| Yrpercentchange | -0.6101 | 0.1545 | 0.3038 | 0.2818 | 0.5235 | 1.332 | 0.3822045 |  |

**Table 2 - Summary Statistics After Outlier Removal**

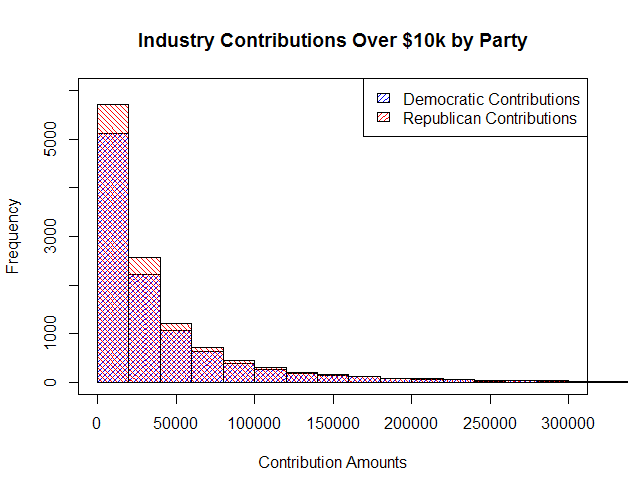
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **PoldataSPIndustries no outliers** | | | | | | | |
|  | **Year** | **Count** | **Party** | **Count** | **Winner** | **Count** |  |
|  | 2004 | 4752 | Dem | 13239 | 0 | 11776 |  |
|  | 2006 | 4688 | Rep | 15794 | 1 | 17450 |  |
|  | 2008 | 4705 | Ind | 193 |  |  |  |
|  | 2010 | 5241 |  |  | **Incumbent** | **Count** |  |
|  | 2012 | 5042 |  |  | 0 | 13608 |  |
|  | 2014 | 4798 |  |  | 1 | 15618 |  |
|  |  |  |  |  |  |  |  |
| **Variable** | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** | **Std. Dev.** |
| Amount | 49 | 8100 | 29000 | 75040 | 88750 | 1379000 | 120266.7 |
| Industrypercent | 0.001293 | 0.03269 | 0.07939 | 0.1538 | 0.1947 | 1 | 0.1865573 |
| Candtotal | 100 | 140300 | 509900 | 546000 | 813300 | 1720000 | 428264.7 |
| Votes | 3713 | 85110 | 120100 | 123400 | 162400 | 259500 | 53989 |
| Percent | 0.0008 | 0.385 | 0.5551 | 0.5295 | 0.6658 | 1 | 0.1999814 |
| Totalracefunds | 72620 | 554900 | 826400 | 1151000 | 1378000 | 22530000 | 1271807 |
| Racefundperc | 0.000012 | 0.1557 | 0.696 | 0.5889 | 0.9845 | 1 | 0.3921799 |
|  |  |  |  |  |  |  |  |
| **PoldataSPIndustriesStockData no outliers** | | | | | | | |
|  | **Year** | **Count** | **Party** | **Count** | **Winner** | **Count** |  |
|  | 2004 | 3486 | Dem | 6580 | 0 | 5962 |  |
|  | 2006 | 2201 | Rep | 8808 | 1 | 9863 |  |
|  | 2008 | 2728 | Ind | 437 | **Incumbent** | **Count** |  |
|  | 2010 | 3820 |  |  | 0 | 6886 |  |
|  | 2012 | 3590 |  |  | 1 | 8939 |  |
|  |  |  |  |  |  |  |  |
| **Variable** | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** | **Std. Dev.** |
| Amount | 49 | 7200 | 21250 | 42230 | 49170 | 1049000 | 64267.69 |
| Industrypercent | 0.001293 | 0.02578 | 0.04742 | 0.07874 | 0.09467 | 1 | 0.09401986 |
| Candtotal | 200 | 206400 | 520200 | 558300 | 810900 | 1720000 | 415769.5 |
| Votes | 3713 | 94380 | 130800 | 133000 | 174000 | 259500 | 53770.39 |
| Percent | 0.0009 | 0.4004 | 0.5717 | 0.5441 | 0.6726 | 1 | 0.1959033 |
| Totalracefunds | 72620 | 539800 | 811300 | 1098000 | 1311000 | 22530000 | 1105324 |
| Racefundperc | 0.0000336 | 0.2192 | 0.7553 | 0.6177 | 0.987 | 1 | 0.380596 |
| Adjclose | 70.09 | 1292 | 2244 | 2801 | 3889 | 8184 | 1981.615 |
| Yrpercentchange | -0.2081 | 0.1607 | 0.3038 | 0.302 | 0.5235 | 0.8584 | 0.2569981 |

**Histograms and Correlations**

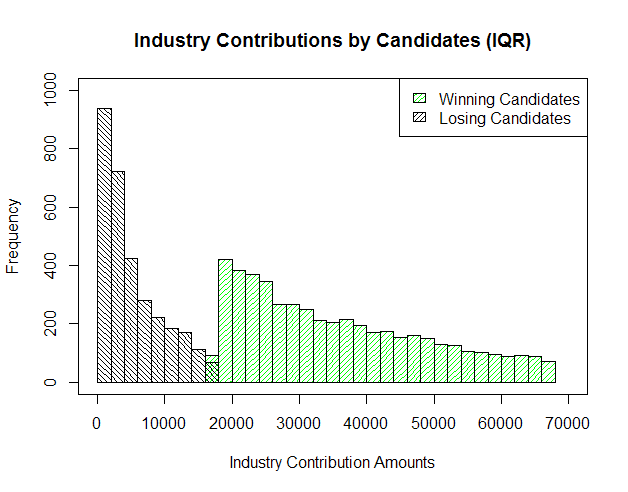
**Histograms**



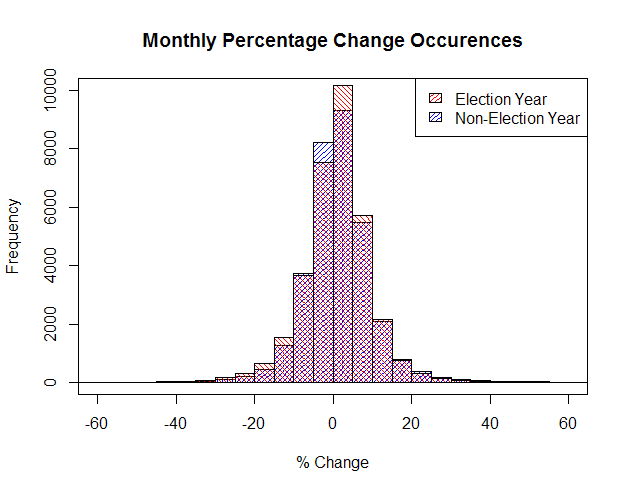
The above histogram provides the total amount of funds raised by each candidate for each Congressional race. Each race is distinguished by year, state, and district. There is a clear dichotomy of funds raised by winning and losing candidates. Not only does this show that the winning candidate always raises more funds than the losing candidate of a particular congressional race, but from the data used, every single winning candidate raised more funds than every single losing candidate of every single race. The data used was the interquartile range of total contributions, which provides a better visual representation without losing the utility of the histogram.



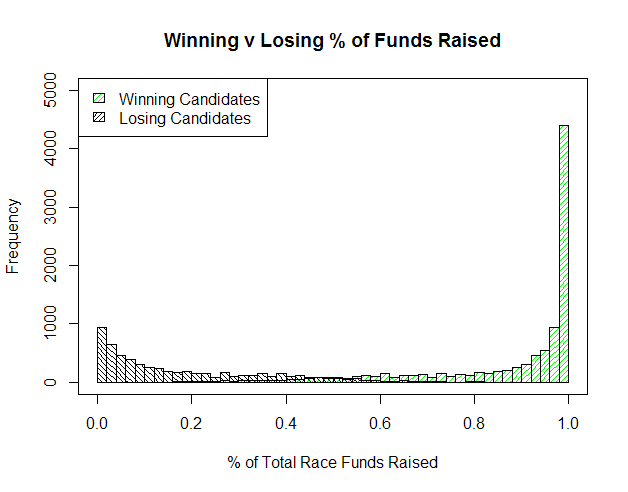
In an effort to show the difference in moneys contributed to the Democratic and Republican parties, the histogram provides visual evidence that Republicans have received the majority of all contributions. The individual contribution amounts are the collective sums of all funds raised from a particular industry to a specific candidate. At nearly every visible level, Republicans received more funds than their Democratic counterparts. Amounts below $10,000 and above $300,000 are considered outliers for the purposes of this histogram and are not represented to provide a better visual representation while maintaining the integrity of the analysis.



When breaking down each candidate’s total funds raised by industry, the visual evidence appears to show either one of two things, or possibly both: candidates who eventually win their congressional race are better fund raisers, in general, or the more funds an industry contributes to a particular candidate, the higher the likelihood that their chosen candidate will win. In the latter case, this would lead us to believe that there is an underlying purpose of supporting one candidate over another which would be of some benefit to the particular industry itself. The data used was the interquartile range of total contributions, which provides a better visual representation without losing the utility of the histogram.



For each month from the beginning of 2004 to September 2015, the percentage change of all adjusted closing prices for stocks listed on the S&P 500 index were calculated and recorded. All percentages occurring in odd-numbered years were classified as “non-election years” while even-numbered years were classified as “election years.” The histogram shows a relatively normal distribution with the balance slightly in favor of positive values, and slightly better performance in election years compared to non-election years.



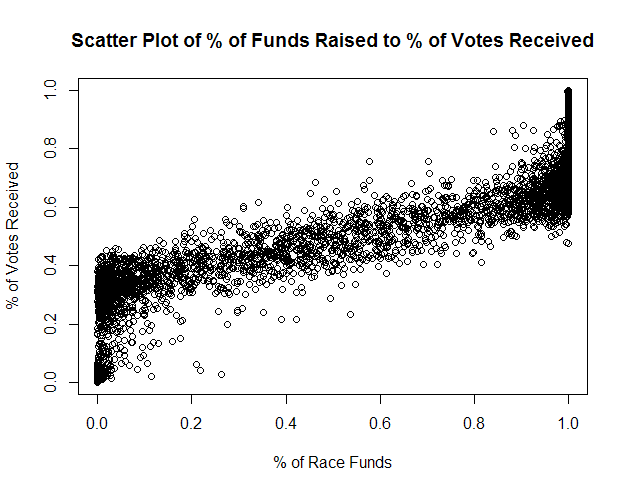
This histogram illustrates the widely-accepted belief that money wins elections. For each election race (distinguished by year, state, and district), each candidate was assigned a percentage of the funds they raised relative to the entire amount of contributions to all candidates involved in the race. It is evident that the higher percentage of funds a candidate raises compared to their competition, the more likely a victory will result. Interestingly, there are some candidates who were able to win their election with less than 20% of total funds raised in a race, and conversely, candidates with more than 80% of all funds raised who lost the election.

**Correlations:**



The three variables used to test for correlation were the percentage of votes received by a candidate in an election, the percentage of funds an individual candidate raised from the total amount raised by all candidates, and the percent change of the value of an industry from the beginning of a congressional term to the end of the term. The percentage of votes received by a candidate is the most important factor in determining an election winner. The percentage of funds raised relative to the entire race gives us a measure of support from industries. The industry value percentage change will allow us to observe potential effects the contributions made to candidates may have.

*%VotesReceived / %ofRaceFunds*



As seen in the histogram “Winning v Losing % of Funds Raised,” there is a very high correlation between the amount of funds raised by a candidate and the amount of votes they receive. A correlation coefficient of over 0.873 indicates a very strong linear relationship between the two variables. In most cases shown in the scatter plot, raising more than 50% of a particular race’s funds resulted in receiving more than 50 % of the total votes.

*%VotesReceived / IndValue%Change*

With a correlation coefficient of -0.020, there is virtually no correlation between the two variables. This result is unremarkable, as there is no intuitive relationship between the percentage of votes a candidate receives relative to the percent change in value of an industry that supports the candidate financially. There may be some type of relationship between the change in value of an industry with the winning or losing candidates that have particular stances on specific issues, but that is beyond the scope of our data.

*%ofRaceFunds / IndValue%Change*

The lowest apparent correlation is between the percentages of funds collected by a candidate and the change in value of an industry between votes. The calculated correlation coefficient is -0.019. Again, there would be no reason to believe that the percentage of funds a candidate collects would have any type of correlation with the change in value of an industry that contributes to the candidate without subsetting the data into more detailed variables.

**Cluster Analysis:**

**Hierarchical Clustering**

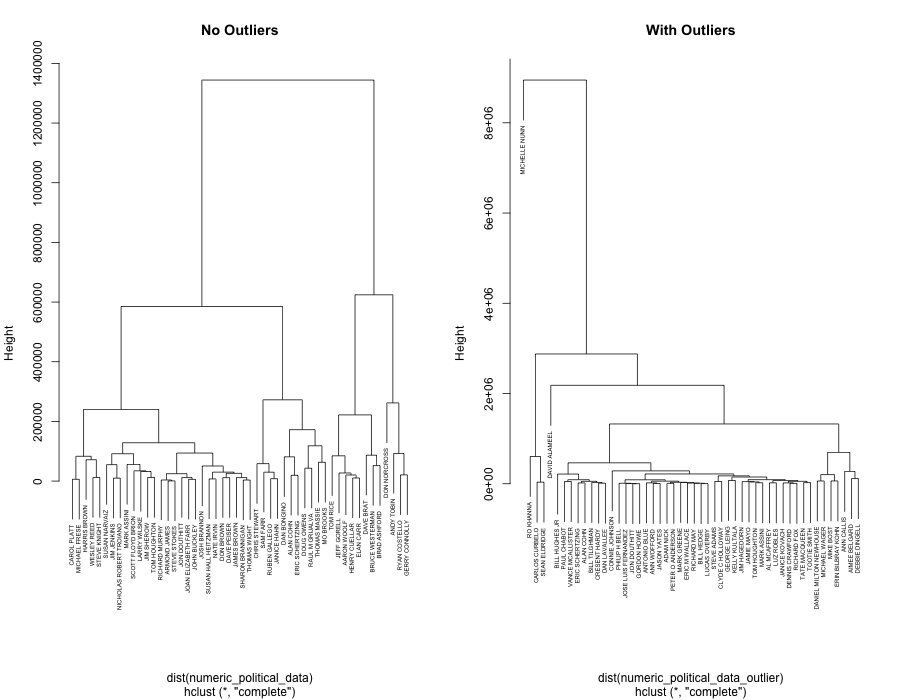
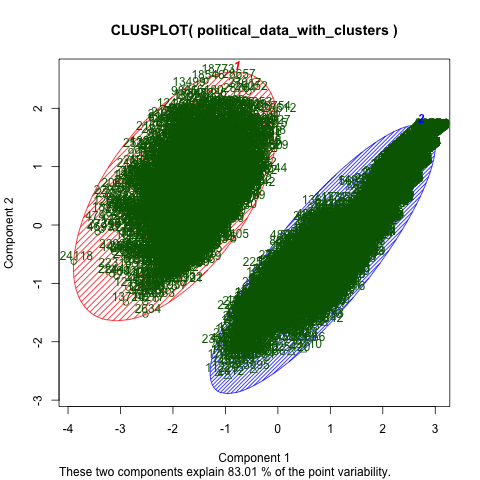
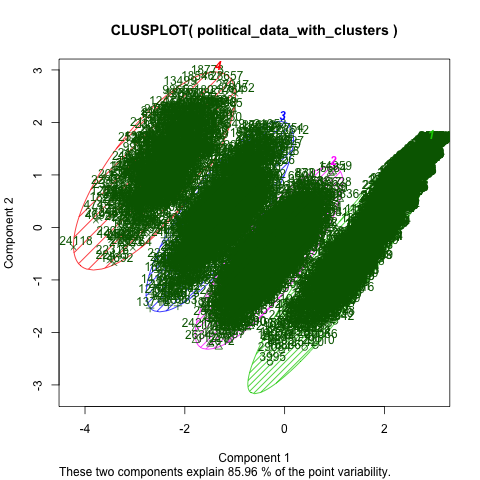


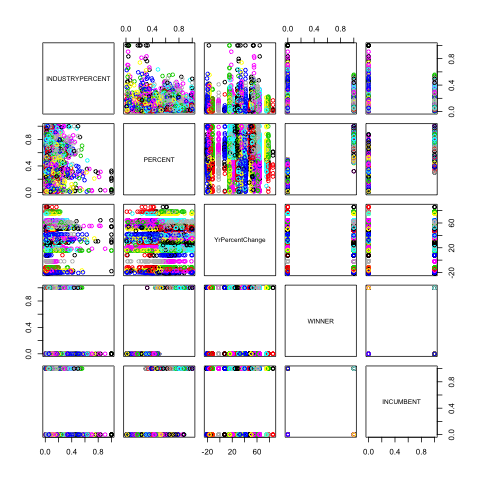
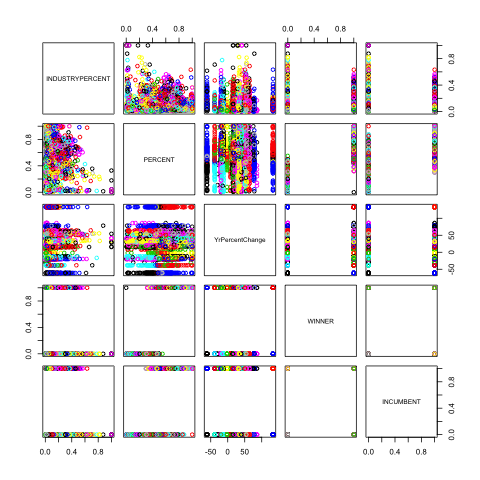
Figure 1 Cluster based on 2014 Congressional Candidates

When attempting the hierarchical clusters, the clusters were based and plotted by year. The “closeness” of candidates is determined by total money contributed by organizations, the number of organizations backed them, and the number of votes that they received. I wanted to also see the differences that the plots would have with or without outliers. From each year, in order to make a readable plot, 50 candidates were sampled from an outlier free dataset and 50 were sampled from a dataset with outliers removed. This was done because a Senatorial candidate tends to be determined from a significantly greater pool of votes than a House Candidate. It’s interesting to see the initial closeness between candidates in 2014 (with outliers removed) is close to zero. This could be because these candidates have close to the same number of votes, amount of funding, or the same number of organizations support them. In the plot including outliers, the structure holds in similar ways with the exception of some stand-alone candidates like Michelle Nunn and David Alameel. Both of these were democratic candidates in deep red states, so less funding or less votes could have resulted in both candidates being placed further away from the clusters. Please note, that for the years 2004,2006,2008, 2010, and 2012, there are corresponding plots in the submission folder.

**K-means Clustering**

The above K-means clusters were created using the non-outlier datasets that did not include stock data. The cluster was created looking at the total amount of money a candidate received, the number of votes they received, and the number of organizations supporting the the candidate. One cluster was created so that two centers would be created and the other was created with the idea that four centers would be created. The 2-cluster plot is interesting because it shows that there are stark differences in terms of class between two types of candidates. Based on the hypothesis tests, it is possible these two groups might be incumbent vs challenger or winner vs loser. The plot with 4 clusters was based on the 4 possible pairing of winner vs loser and incumbent vs. challenger . The assumption that could be made is that clusters 4 and 1 (see labels on plot) could be incumbents who won and challengers that lost and and 2 and 3 could be the remaining two cases. Unlike the 2-cluster kmeans, the one with four clusters did not show as much space between division, possibly indicating that looking at four different cases isn’t as telling as looking at only two cases.

**DBSCAN Clustering**



DBscans plot were created looking at the dataset including the stock data with outliers and the dataset without outliers. The attributes analyzed were: the industry percentage of contributions to candidate, the percentage of votes a candidate received, the percent change of stock price for an industry and whether the candidate won or is the incumbent. All plots comparing winners and incumbents create very polar structures along the borders. This is probably due to the fact that both winners and incumbents are binary (either 1 or 0). It’s interesting to see the plots between voting “percent” and “yrpercentchange”, an attributes based on an industry’s stock price change during a relevant year. In both dbscan plots, there appears to be a linear like structure occurring. Industry percent vs voting percent seems to be an incoherent scatter plot, which probably indicates no significant relationship between the two attributes appears to exist.

**Association Rules / Frequent Itemset Mining Analysis**

We ran association rule mining on the political subset of our data, consisting of political contributions to candidates by industry, as well as election results. Apriori rule mining was run to find rules with a minimum confidence level of .2, under three different support levels (.4, .2, .05). Eclat rule mining also was utilized with the same support level but, as it yielded nearly identical results to the Apriori algorithm, the analysis focuses on the rules generated by the Apriori algorithm. A selection of rules deemed interesting is found at the end of this section, and files containing all of the rules generated by the Apriori algorithm are available in the submission folder.

A variety of characteristics occurred frequently with the incumbent variable, a binary variable which represents whether or not a candidate was an incumbent. Incumbent and election winner occurred together in 50.9% of the observations in the dataset; incumbent candidates won 95.3% of reelection opportunities, and 85.3% of all election winners were incumbents (rules 1, 2). This is in line with the historical average proportion of incumbents who win reelection[[3]](#footnote-3), providing evidence that our dataset is representative of the real world. Another way in which the rule mining analysis mirrored popular opinions about American politics is in the relationship it revealed between a candidate’s fundraising and election success. 72.4% of candidates who raised a very low amount of funding lost their election bids, and 74.9% of election losers raised a very low amount of funding, while 77.9% of those who raised a high amount of funding won their election bids (rules 3, 4, 5).

Challenger (non-incumbent) and election loser also frequently occurred together, in 37.8% of the observations. 81.1% of challengers lost, and 93.8% of losing candidates were challengers (rules 6, 7). The incumbent variable also had a strong relationship with candidate funding levels; in general, incumbents received more funding than challengers. Challenger was frequently associated with very low levels of campaign contributions, occurring together in 31.4% of observations; 67.5% of challenger candidates received a very low amount of funding, and 75.5% of all candidates who received a very low amount of funding were challengers (rules 8,9). On the other end of the funding distribution, 76.5% of candidates who received a high amount of funding were incumbents, and 70.1% of candidates who received a mid-high amount of funding were incumbents (rules 10, 11).

Apriori association rule mining revealed some interesting frequent itemsets featuring the political party variable. Over the time period examined (2004-2014), belonging to the Republican party was frequently associated with being elected. 62.6% of Republican candidates won their elections, and 56.6% of election winners were Republicans (rules 14,15). Interestingly, 57% of Democratic candidates also won election over the time period (rule 16). There weren’t enough independent candidates in the dataset to show up in the association rules even at support level .05, but one takeaway from the Democratic and Republican results is that Independent candidates generally do not do very well in American congressional politics; out of the 193 Independent candidates in the dataset, only 23 were elected.

By ranking every industry that contributed to each candidate by contribution amount, we hoped to determine if the breakdown of a candidate’s contributions affects the candidate’s performance in elections. However, even with a minimum support level of .05, there were only two combinations of the industry rank variable, an industry, and another variable that occurred frequently enough to get picked up by the association rule generation. The “industry” not for profit occurred with an indrank rank of 1 (meaning nonprofits were the candidate’s primary source of funding) and a very low level of funding in 5.4% of the observations; 49.3% of candidates whose primary industry was not for profit received a very low level of funding (rule 18). Considering that overall, 41.6% of candidates received a very low level of funding, this is not a very significant result; candidates whose main contributor was the not for profit industry were slightly more likely to receive a very low level of funding than the average candidate. Not for profit also occurred with an indrank of 1 and the election loser indicator in 5.8% of cases; 52.8% of candidates whose primary source of funding was nonprofits lost their elections (rule 19). This is a fairly significant result; only 40.3% of candidates in the dataset lost their elections, so candidates whose largest contributor was nonprofits were much more likely than average to lose.

Apriori rule mining revealed many associations which were in line with our expectations; while the rules we found strengthened our conviction that our dataset is representative of the real world, they failed to bring much new information to light. The main takeaway from the frequent itemset mining is that in American congressional politics, life is hard for challengers; incumbents have a large fundraising advantage and win an incredibly high proportion of elections in which they participate. Deeper analysis is necessary to determine if the sources of candidate’s funding actually have an impact on election results, or if it is only the amount of funding that matters.

**Table 3 - Interesting Frequent Itemsets**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Rule** | **Support** | **Confidence** | **Description** |
| 1 | {INCUMBENT=1} => {WINNER=1} | 0.509512078 | 0.953451146 | 95.3% of incumbents won reelection |
| 2 | {WINNER=1} => {INCUMBENT=1} | 0.509512078 | 0.853352436 | 85.3% of winners were incumbents |
| 3 | {CANDTOTALLEVEL=Very Low} => {WINNER=0} | 0.301615 | 0.724322104 | 72.4% of candidates who raised very low amounts lost |
| 4 | {WINNER=0} => {CANDTOTALLEVEL=Very Low} | 0.301615 | 0.748556386 | 74.9% of losing candidates raised very low amounts of money |
| 5 | {CANDTOTALLEVEL=High} => {WINNER=1} | 0.055601177 | 0.779002876 | 77.9% of candidates who raised a high amount won |
| 6 | {INCUMBENT=0} => {WINNER=0} | 0.378053788 | 0.811948854 | 81.1% of challengers lost |
| 7 | {WINNER=0} => {INCUMBENT=0} | 0.378053788 | 0.938264266 | 93.8% of losers were not incumbents |
| 8 | {CANDTOTALLEVEL=Very Low} => {INCUMBENT=0} | 0.314309177 | 0.754806902 | 75.5% of candidates who raised a very low amount were challengers |
| 9 | {INCUMBENT=0} => {CANDTOTALLEVEL=Very Low} | 0.314309177 | 0.675044092 | 67.5% of challengers raised a very low amount of funding |
| 10 | {CANDTOTALLEVEL=High} => {INCUMBENT=1} | 0.05460891 | 0.765100671 | 76.5% of candidates who raised a high amount of funding were incumbents |
| 11 | {CANDTOTALLEVEL=Mid-High} => {INCUMBENT=1} | 0.111510299 | 0.70116179 | 70.1% of candidates who raised a mid-high amount were incumbents |
| 12 | {VOTEPERCENTLEVEL=Very Low} => {INCUMBENT=0} | 0.114863478 | 0.995551601 | 99.6% of candidates who received a very low vote percent were not incumbents |
| 13 | {VOTEPERCENTLEVEL=High} => {INCUMBENT=1} | 0.111510299 | 0.952923977 | 95.3% of candidates who received a high vote percent were incumbents |
| 14 | {PARTY=R} => {WINNER=1} | 0.338328885 | 0.626060529 | 62.6% of Republican candidates won |
| 15 | {WINNER=1} => {PARTY=R} | 0.338328885 | 0.566647564 | 56.6% of election winners were Republicans |
| 16 | {PARTY=D} => {WINNER=1} | 0.257955245 | 0.569453886 | 57% of Democratic candidates won |
| 17 | {WINNER=1} => {PARTY=D} | 0.257955245 | 0.432034384 | 43.2% of election winners were Democrats |
| 18 | {PRIMARY.INDUSTRY=Not for profit,indrank=1} => {CANDTOTALLEVEL=Very Low} | 0.054061452 | 0.493133583 | 49.3% of candidates whose primary industry was not for profit raised a very low amount of funding |
| 19 | {PRIMARY.INDUSTRY=Not for profit,indrank=1} => {WINNER=0} | 0.057893656 | 0.528089888 | 52.8% of candidates whose primary industry was not for profit lost |

**Network Analysis:**

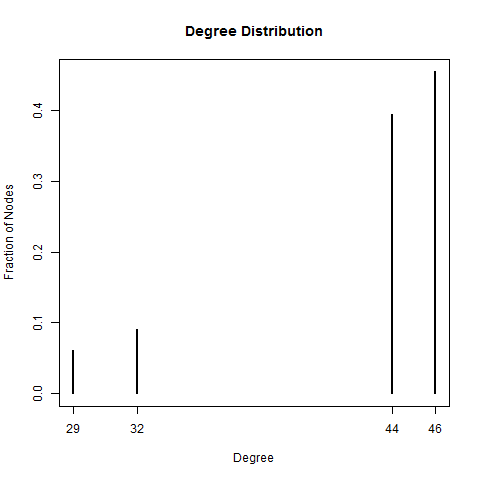
It is clear how the most straight forward implementation of a network to our data, or maybe the most obvious, might be bi-modal. Such a network would have nodes belonging to both candidates and industries. In particular, the visualization of such a network might make clusters of candidates much clearer. However, it is also clear that multimodal networks, even if the network under consideration is just a bimodal network, quickly become difficult to interpret and manage.[[4]](#footnote-4) With this in mind, and in accordance with the assignment, we decided to consider the network of candidates and industries, but converted to the unimodal domain.

In order to make a suitably small matrix, we used a subset of our data to build it. The subset was determined by taking all of the winning candidates for senate seats in 2014, and their two largest sponsors. What we were hoping to see was something interesting about the primary industries that supported those candidates. Perhaps the Republicans and Democrats would be in two clusters, or perhaps different industries focused on different candidates by region of the United States. What we ended up finding was that it looks as though industries hedge their bets. Using the top two industries, as we did here, we connect all of the winning senators a remarkable amount. There is a strong possibility that this is because of the binning strategy applied to the industries. However, an analysis of the results follow below.

The resulting dataset after the subsetting above results contains 33 senators. These senators each have two rows with industries that could connect them directly with other candidates. Keep in mind the following summary statistics for the senators’ dataset:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Industry** | Consumer Staples | Energy | Financials | Not for profit | Not publicly traded |
| **Count** | 1 | 1 | 18 | 30 | 16 |

After converting our network using a strategy by Solomon Messing found in his blog, we applied the example from Blackboard and implemented our network using the R igraph package.[[5]](#footnote-5) Since the original dataset had 33 senators, we end up with a network with 33 nodes. Amazingly, with just the top two industries per candidate, the candidates can connect to one another with a mean degree of 42.9 (we also tried this statistic after removing multiple edges, finding a mean degree of 31.6). Of course, the only way to get more than 32 degrees is to connect to a candidate through both of the top two industries. Only 5 of the 33 candidate failed to connect to every other candidate. Please find a degree distribution for the network with multiple edges, below.

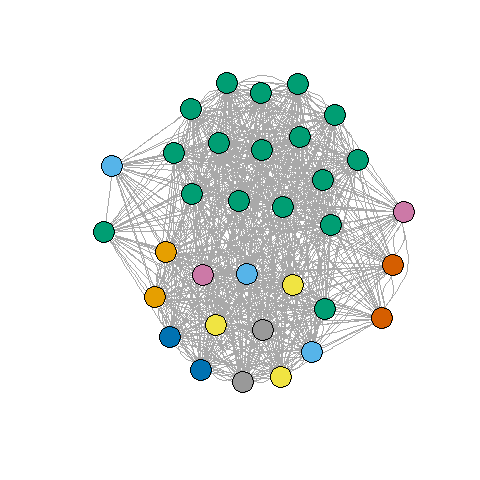


Unsurprisingly, this was also born out in the betweenness statistic, so it didn’t provide any additional value. The 5 that didn’t connect to every other candidate had betweenness scores of 0 since the other candidates don’t need to pass through any nodes to get to everyone. Every other candidate had a betweenness score of 0.2142857 due to the 5 being able to cross through them 1 time to get to any that they can’t get to otherwise.

The mean clustering coefficient was 0.56. For the 5 less connected candidates, in two cases there was a coefficient of 1, and for the other 3 it was close to 1 at 0.88. All of the other were around 0.5. This shows that 2 of the 5 less connected candidates were the only winning candidates for one of their sponsoring industries (of those where that industry was one of the top two sponsors). That would be the only way to get a coefficient equal to 1. The other three, however, share whatever industry it is that they have in common, bringing down the coefficient.

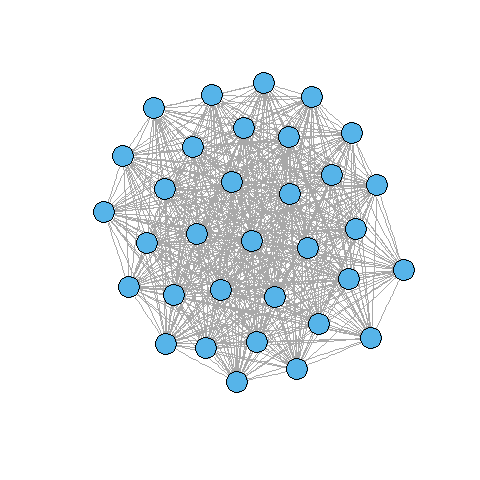
Given the above statistics, we already have a pretty clear idea of the remaining statistics. The graph density is greater than 1 due to the multiple edges at 1.34. The graph diameter is equivalent to 2, which indicates that even the 5 candidates that stand out have a first connection the same network. The number of connected components is equivalent to 1 (If just one industry is picked for each Senator – essentially turning the network into clustering – there are four components). Finally the largest k-core was 39. These are all indicators of just how many of the senators share both top supporters.

Please find the edge-betweenness network, below.



As you can see, and as we guessed based on the network statistics, the network is very intertwined. We did take a look at and tried to analyze the green cluster in particular, and this appears to be associated primarily with candidates that had both “Not for Profit” and “Financials” as their supporting industries. It is not clear why this cluster was able to snag the two dots not in its prime area. It is not clear why the bottom cluster was not also colored the same, as it must represent the combination of “Not for Profit” and “Not Publicly Traded”. The five that are more separated from the rest are clearly our 5 disparate points, with the two loners on the left holding “Consumer Staples” and “Energy” as one of their two top supporters, respectively, and the three on the right sharing “Financials” and “Not Publicly “Traded” in common.

Please find the modularity network, below. For this network, multiple edges were removed since that was a requirement of the greedy algorithm to run on this network.



Not surprisingly, the greedy algorithm determined that all of the nodes were part of the same community. Once multiple edges are removed it is very hard to see any significant pattern in the visualization of the network.

**Part 2: Predictive Analysis**

**Hypothesis Testing:**

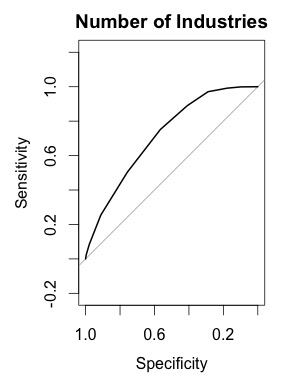
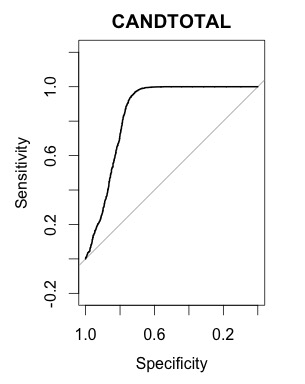
**Student’s T-test**

For the Parametric Statistical Tests, the dataset, “PoldataSPIndustriesStockData no outliers”, was used, the student t-test was run on one of the three hypothesis developed, while the logistic regression model was run on a possible linear relationship. In hindsight, it would have been interesting to run cross validation on the model, but, unfortunately, time was a factor in forgoing it. ROC curves were provided on both of the attributes tested by the logistic regression model.

The first Null hypothesis tested was: there is no difference between the total contributions that an incumbent gets and that a challenger gets. We used our Merged Data set without outliers. We preformed a **student t-test** (not pairwise) to test this null hypothesis. We got a p-value of < 2.2e-16 and given that that this a social science analysis, the threshold should be .05. The p-value crosses this threshold and is well in the rejection region, so the p-value of extremely significant. So we reject the null hypothesis in favor of the alternative which is that there is a difference. Given the mean of the two categories of contributions, it is clear that the incumbent has a higher amount of the contributions compared to the challenger. This is also verifiable when compared to the association rules.

**Logistic Regression**

The Second Null hypothesis that tested involved a **logistic regression model**. The idea behind the model is to predict who the winner will be based on the Total amount of money raised and the Number of supporting industries. From the confusion matrix, we know the following; The Precision of the model is 0.8125881, the recall is 0.8662994, is F-measure is 0.8385846. Below are the ROC plots of both variables. The accuracy against the training data is 0.8184791; however, the number seems low because the prediction should have better matched against the actual results. This may mean that the data might not best tested using a model that assuming normally distributed data.



From the ROC curves, it seems like the logistic model predicts with a surprising degree of accuracy. This could signify that the model maybe be over fitting. This is probably because the model is predicting the values based upon itself. As excepted, the total amount of money raised seems to have a bigger impact on accuracy as opposed to Number of Industries.

**Confusion Matrix:**

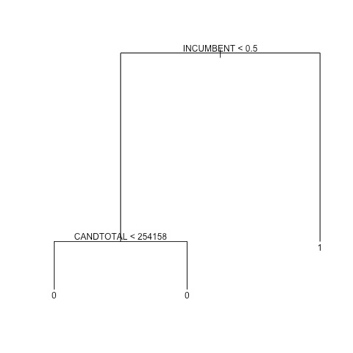
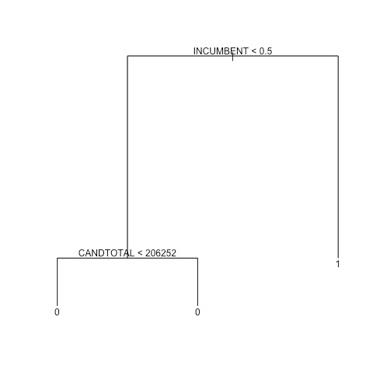
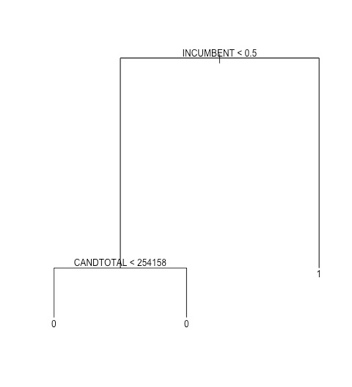
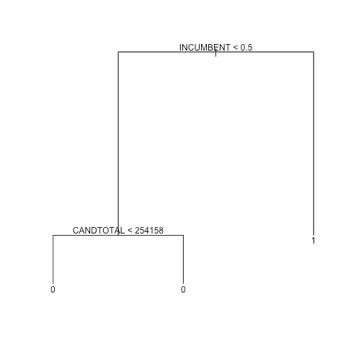
|  |  |  |
| --- | --- | --- |
| Predict: | lose | win |
| Lose | 1273 | 399 |
| win | 267 | 1730 |

**Decision Tree**

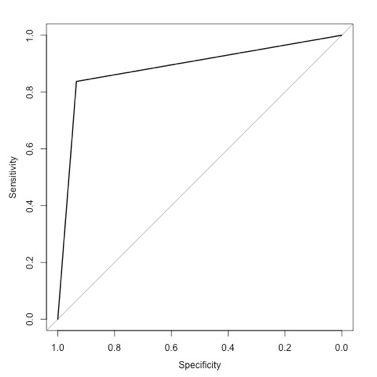
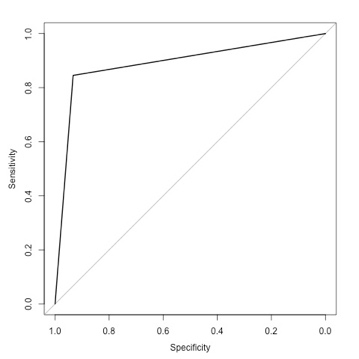
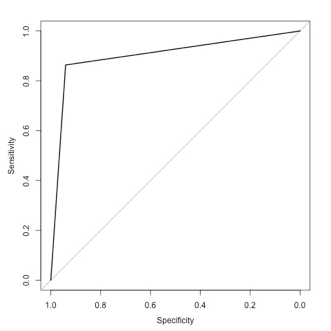
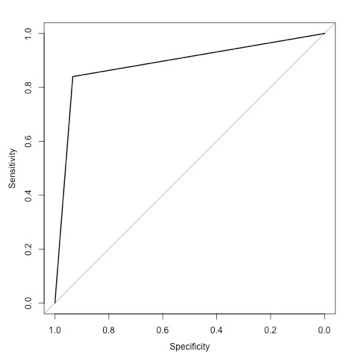
For the data driven Predictive models, we looked into the following **third hypothesis**, the idea was to see if the winner attribute of a candidate, a binary one, could be predicted looking at Candidate’s Total Industry Contributions, the number of supporting industries, and whether the candidate was an incumbent. In each of the three models, cross validation was run 5 times. The training set comprised of 80% of the data set whereas the test dataset comprised of 20% of the original dataset.

For the decision Tree, the follow occurred with the tree after in was pruned in terms of confusion matrices, accuracy of model, precision, recall, and the F-measure:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Iteration 1 | Iteration 2 | Iteration 3 | Iteration 4 | Iteration 5 |
| Predict: 0 1  0 314 70  1 16 334 | predict 0 1  0 316 58  1 14 346 | Predict 0 1  0 313 60  1 20 341 | predict 0 1  0 321 59  1 17 337 | predict 0 1  0 332 52  1 21 329 |
| accuracy: 0.882834 | accuracy: 0.901907 | accuracy: 0.891008 | accuracy: 0.896458 | accuracy: 0.900545 |
| precision: 0.826733 | precision: 0.856436 | precision: 0.850374 | precision: 0.85101 | precision: 0.86352 |
| Recall: 0.954286 | Recall: 0.96111111 | Recall: 0.944598 | Recall: 0.9519774 | Recall: 0.94 |
| F-measure: 0.88594 | F-measure: 0.90576 | F-measure: 0.89501 | F-measure: 0.89867 | measure: 0.90013 |

Due to an issue in the for loop to run the iteration, I wasn’t able to get the tree structure. However, from the tree structure, it is obvious that the candidate total does not impact whether a challenger wins or not given the status of being an incumbent.

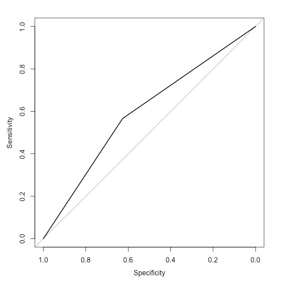
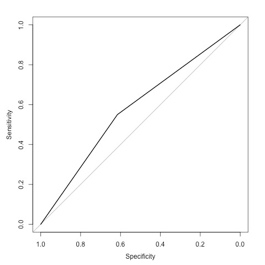
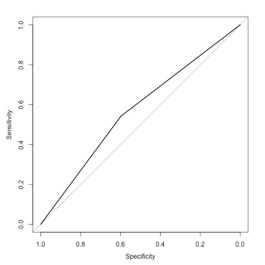
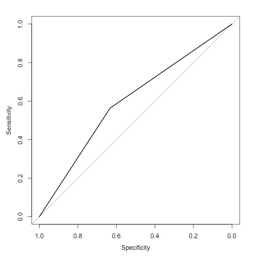
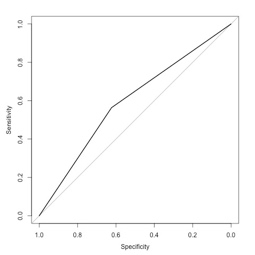


From the ROC curves, it seems like the model does a good job of predicting the candidate that will win. Interesting note is that the decision tree seems to associate winning the election with being an incumbent. The accuracy rates also seem to confirm this idea in addition to the tree structure and confusion matrix because they are close to what the winning rate for an incumbent is. Based on this model, it confirms a deeply believed theory in congressional politics, incumbents win. However, the model may not be the most useful because it’s precision is lower than the other categories.

**Lazy Learner**

For the KNN (lazy learner) algorithm, the k value was benchmarked for values between 1 to 20; the benchmark was based on finding the k value that resulted in the most accuracy. This could have been improved if, for each k, was tested approximately 1000 times, but given the size of the dataset, one time was enough. In the end, the K value that was used was 17.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Iteration 1 | Iteration 2 | Iteration 3 | Iteration 4 | Iteration 5 |
| testTarget  pred 0 1  0 256 16  1 86 376 | testTarget  pred 0 1  0 244 18  1 80 392 | testTarget  pred 0 1  0 247 14  1 96 377 | testTarget  pred 0 1  0 236 13  1 89 396 | testTarget  pred 0 1  0 229 20  1 111 374 |
| accuracy: 0.576294 | accuracy: 0.611716 | accuracy: 0.572207 | accuracy: 0.588556 | accuracy: 0.5722 |
| precision: 0.959184 | precision: 0.956098 | precision: 0.96419 | precision: 0.96822 | precision: 0.94924 |
| Recall: 0.81385281 | Recall: 0.830508 | Recall: 0.797040 | Recall: 0.816495 | Recall: 0.771134 |
| measure: 0.880562 | F-measure: 0.88889 | F-measure: 0.87269 | F-measure: 0.88591 | F-measure: 0.85097 |

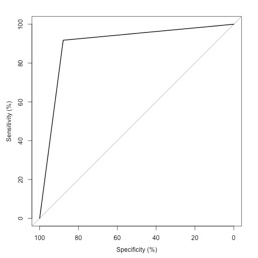
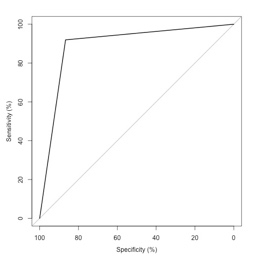
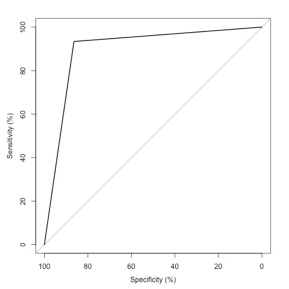
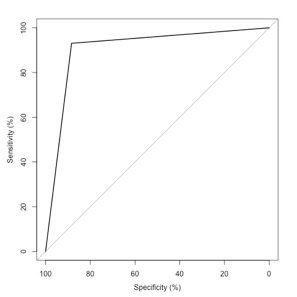
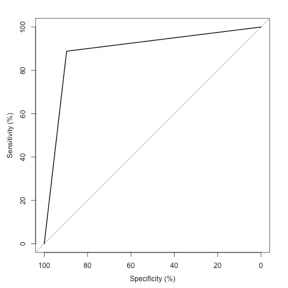


Based on the accuracy, this model doesn’t do a good job of determining accuracy; however, the precision is very good, meaning that the amount of true positives is very good. The downsize is that there is a significant of false positives that are detected. This model may not be the best for overall detection. The issue may be because of the incumbent value.

**Naïve Bayes**

For the Naïve bayes algorithm, the follow occurred in terms of confusion matrices, accuracy of model, precision, recall, and the F-measure:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Iteration 1 | Iteration 2 | Iteration 3 | Iteration 4 | Iteration 5 |
| pred 0 1  0 333 54  1 19 328 | pred 0 1  0 324 60  1 18 332 | pred 0 1  0 327 58  1 17 332 | pred 0 1  0 321 62  1 15 336 | pred 0 1  0 328 49  1 18 339 |
| accuracy: 0.900545 | accuracy: 0.89373 | accuracy: 0.897820 | accuracy: 0.895095 | accuracy: 0.908719 |
| precision: 0.858639 | precision: 0.84694 | precision: 0.85128 | precision: 0.84422 | precision: 0.87371 |
| Recall: 0.94524496 | Recall: 0.9485714 | Recall: 0.9512893 | Recall: 0.95726496 | Recall: 0.9495798 |
| F-measure: 0.89986 | F-measure: 0.89487 | F-measure: 0.8985 | F-measure: 0.8972 | F-measure: 0.91007 |



From the results of the confusion matrix, the accuracy, precision, recall, and F-Statistics, are equal to or better than the other two models. The ROC curves cover a bigger area than than either KNN or Decision trees. The false negative numbers are significantly less as well, which is also evident by precision and recall.

1. https://eresearch.fidelity.com/eresearch/markets\_sectors/sectors/sectors\_in\_market.jhtml [↑](#footnote-ref-1)
2. https://www.opensecrets.org/industries/slist.php [↑](#footnote-ref-2)
3. https://www.opensecrets.org/bigpicture/reelect.php [↑](#footnote-ref-3)
4. Source: Scott Weingard, “Networks Demystified 9: Bimodal Networks”, http://www.scottbot.net/HIAL/?p=41158 [↑](#footnote-ref-4)
5. Source: blog, https://solomonmessing.wordpress.com/2012/09/30/working-with-bipartiteaffiliation-network-data-in-r/ [↑](#footnote-ref-5)