A Statistical Analysis of Campaign Finances in US House of Representatives Elections (2015-2016)

Submitted to:

Dr. Gideon Simpson, PhD
Department of Mathematics
Drexel University
Philadelphia, Pennsylvania

Prepared by:

Jasper MacNaughton
Sanjana Venkat
Paul Ciaccia III

March 21, 2019

Source code maintained at:
https://github.com/jaspermacnaughton/House-Elections-2016

Abstract

This report summarizes statistical modeling and analysis done on a dataset retrieved from the Federal Election Commission of the United States of America. The dataset consists of various campaign finance-related and categorical data of the 2016 election cycle of the U.S. House of Representatives elections. From this dataset we make several observations on the nature of the data, as well as how campaign financial measures related to categorical variables. Lastly, we perform predictive modeling of election results as a function of the variables.

1.0 Introduction

Congressional elections are funded through a combination of contributions and loans. Contributions can be broken down into four categories: individual, party, candidate and other. The majority of campaign financing comes from individual contributions, which come from individual donors. Party contributions come from any political party and candidate contributions come directly from the candidates themselves. The “other contributions” category mostly refers to contributions from PACs (Political Action Committee), as well as contributions from other candidates. On the other hand, loans can be broken into those received from the candidate and those received from banks. Any loan that is from an individual is treated as an individual contribution, rather than a loan. These various campaign contributions differ vastly per candidate, which calls us to question whether campaign financing plays a definitive role in the outcome of elections.

The aim of this investigation was to analyze the ways in which finances have an influential role in election outcomes. As the money involved in election campaigns in the United States continues to rise in magnitude, there is increased debate and awareness regarding how wealth should be regulated in politics, and the extent of the role that it plays in influencing the way citizens vote. The dataset downloaded for this report describes the various variables that compose campaign finances, including, but not limited to: individual contributions, party committee contributions, candidate contributions, loans, refunds, and debts owed. From this dataset, clean up was done to isolate variables that we chose to focus on, and from there statistical modeling was done on the trimmed data.

The data set came from the Federal Election Commission and gives a thorough breakdown of all campaign related finances per candidate across presidential, senate, and

house elections in the 2015-2016 election cycle. We chose to examine the house elections because they gave us the best sample size (being that all 435 seats are up for election every two years). From there, we had to decide which variables in the dataset were relevant to our research. We chose to pull information regarding campaign contributions, loans, expenditures, incumbency status, and election outcome. We ignored variables that dealt with the location of elections, disbursement of funds, receipts, and refund information. From the data selected, we found that party, self and committee contributions had the most zeros. The data was still relevant to our research and we kept it but we do note that many candidates did not receive funding in one, if not all, of these categories.

We performed our analysis in R, which can be broken into three broad categories: data clean-up, descriptive elements, and predictive elements. The methods section provides further details on each of these categories. In the data-cleanup, multiple steps were taken to ensure that the data types of each variable were appropriate for analyses, and that the data being used was rational. Then, descriptive elements were used to visualize the distributions of the various variables and generate hypotheses on those that could prove to be most useful in the predictive models. To build on that, predictive elements were generated regarding election outcomes. We performed 10-fold cross-validation to estimate the accuracy of various classification model types, including logistic, LDA (linear discriminant analysis), QDA (quadratic discriminant analysis), KNN (K-nearest neighbors), and SVM (support vector machine). From this, we determined that SVM with a Radial Basis Kernel function was the best method to classify election outcomes.

2.0 Methods

The preliminary steps in this project consisted of clean-up of the raw dataset obtained from the FEC (Federal Election Commission). The downloaded dataset contained 51 variables and 1814 candidates. First, the data fields involving financial numbers were converted to numeric type by removing any characters such as dollar signs and commas. We set negative and non-existent values to 0. For the scope of our analysis, we removed presidential and senate candidates, as we focused on the House of Representatives elections.

This clean-up process generated a workable data frame from which we selected variables of interest to be used in our analyses. The following variables were chosen and renamed for our final dataset:

Category	Variable Name in Code
Office Status	Office
State Affiliation	Status
Party Affiliation	Party
Incumbency Status	Seat_Status
Individual Itemized Contributions	Ind_Itemized_Con
Individual Unitemized Contributions	Ind_Unitemized_Con
Total Individual Contributions	Ind_Total_Con
Party Committee Contributions	Party_Con
Other Committee Contributions	Com_Con
Candidate Contributions	Self_Con
Total Contributions	Total_Con
Total Loans	Total_Loans
Operating Expenditures	Operating_Exp

Net Contribution	Net_Con
Outcome Election	Winner

Figure 2.1: Variables and Code Names Used in Final Dataset

The final dataset consisted of 15 variables and 1656 candidates. The number of votes was removed as a variable because the votes were only recorded for winners of contested House races. After cleaning-up the raw, downloaded data frame and generating a more compact one consisting of elements of interest, the next steps taken involved performing sanity checks on the data to ensure that the number of “0s” and blank data cells matched what was found in the raw dataset. It was found through this process that the data matched, therefore validating that the clean-up process was done properly.

With a thoroughly checked and compacted dataframe, descriptive elements were then produced to generate visualizations of the data. First, histograms were generated to give an idea of how the variables were distributed. The binwidths were adjusted for better visualization, and zeros were removed as the presence of a zero column took away from the purpose of the plots. Scatter plots of various financial variables were also generated to compare distributions of the numerical variables. These scatter plots helped to notice trends in different subsets of the data. Through color coding, clusters of interest appeared prominently. Next, we created box plots as a method of visualization because they clearly present statistical values such as the mean, range, and quartiles. For the plots, a log scale was chosen in certain cases for more appropriate visualization. Additionally, for the plots that weren’t placed on a log scale, some were scaled manually which led to outliers being cut out of the frame of view.

After performing basic summary statistics, Principal Component Analysis was performed on our data set. PCA was chosen to analyze how the variation in the data could be represented by the principle components. First, numeric columns were pulled from the data set for the

analysis. After performing PCA, a biplot of the data was generated using autoplot to visualize the results.

The last phase of the statistical analysis involved predictive modeling. Two logistic model plots were produced for visualization, one with a single predictor and one with two predictors. Then, to determine which predictive model produced the most accurate results, cross-validation on the following models was performed: single predictor logistic model, double predictor logistic model, full logistic model, LDA, QDA, KNN (k=3, k=5), linear basis SVM, and radial basis SVM. The cross-validation method used was 10-fold.

3.0 Results

As seen in Figure 3.1, the first descriptive visualization done was histograms to visualize the distributions of the numeric variables. The variables were plotted on a logarithmic scale due to the wide range of the numbers. The log scale x-axis implies that the distributions of the variables are exponential, and that there is a long right tail that causes the wide range. This can most likely be attributed to the fact that there is a very large upper limit to the amount of money that was raised and spent in the House elections.

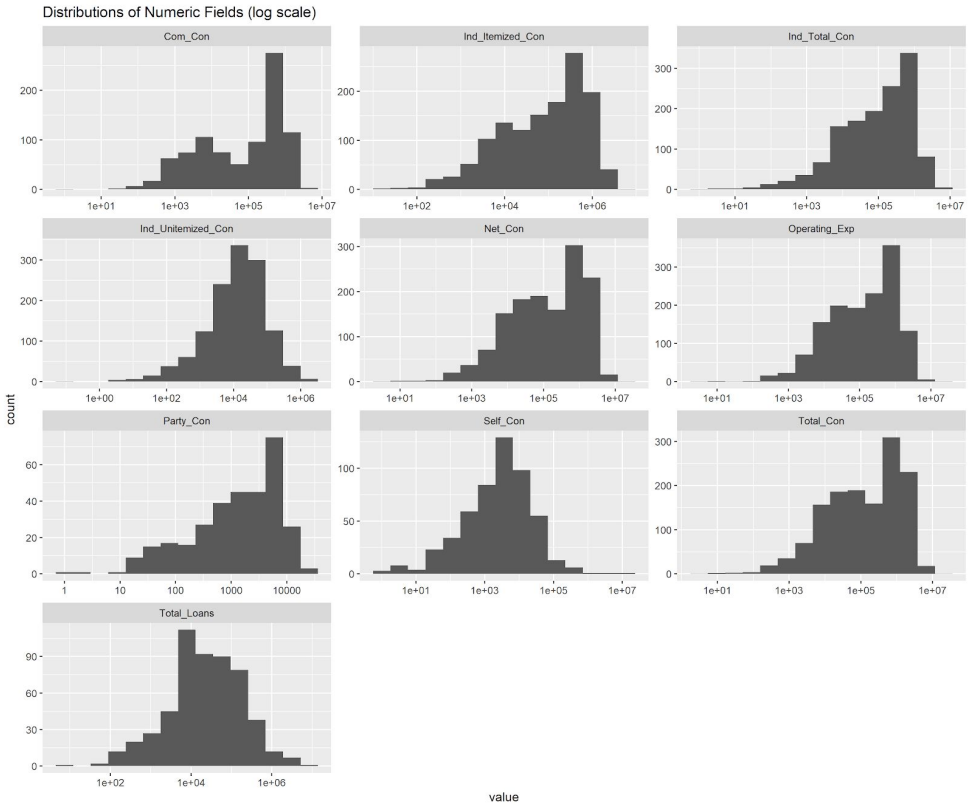

Figure 3.1: Distributions of Numeric Variables in Logarithmic Scale

The next visualization was of pairwise correlations between the numerical variables in the dataframe. As can be seen in Figure 3.2, a few pairs of correlations stand out with very strong correlation coefficients, but many of them are logical and expected. For example, it is expected that total individual contributions and individual itemized contributions would have a

strong positive correlation, since total individual contributions includes itemized and unitemized contributions. However, one pair of variables with a strong correlation coefficient that stood out was operating expenditures & net contributions, with a correlation coefficient of .899. In fact, operating expenditures had a strong positive correlation with many other variables including individual itemized contributions (.844), total individual contributions (.847), and net contributions (.899). Thus, this visualization served as a catalyst for us to further explore these strongly correlated pairs of variables, as can be seen later in this report. We also note the strong correlations between some of the predictor variables, potentially presenting the issue of autocorrelation when it comes to predictive modeling.

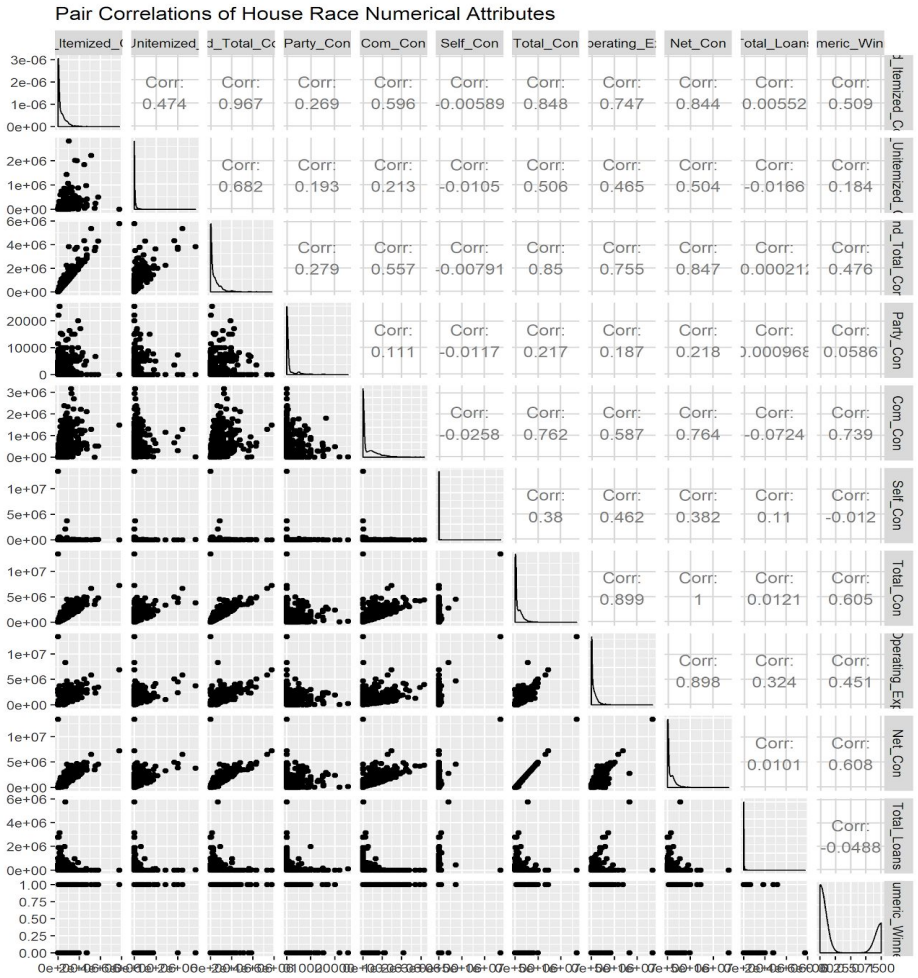

Figure 3.2: Pair Correlations of Numerical Variables in House Elections

Next, specific scatter plots were generated to hone in on the effects of various numerical variables. As seen in Figure 3.3, the first scatter plot produced was individual itemized contributions vs. individual unitemized contributions. It is apparent through the color-coding that the candidates that won their elections are clustered towards the top right of the graph (blue, or "1", signals a winner), while the candidates that lost elections are more widely dispersed. Additionally, since the scatter plot is on a logarithmic scale, the trend itself seems to be exponential. Thus, one conclusion drawn is that higher amounts of individual contributions, both itemized and unitemized, correlated with an increased chance of winning an election. Additionally, Figure 3.4 was generated, which is a scatter plot of committee contributions vs. party contributions (on a logarithmic scale). What can be seen from this scatter plot is that party contributions had very little effect on the outcome of the election, while there is a strong positive correlation between committee contributions and the candidate winning the election.

Figure 3.3: Scatter plot of Individual Itemized Contributions vs. Individual Unitemized Contributions

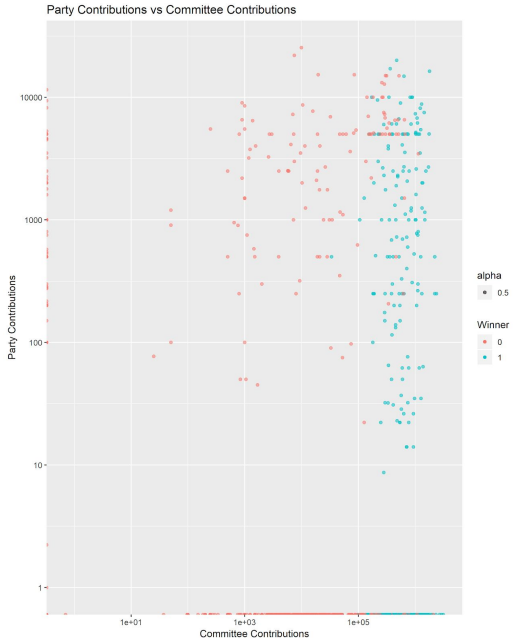

Figure 3.4: Scatter plot of Committee Contributions vs. Party Contributions

Furthermore, scatter plots were generated for the intriguing pair of variables that was highly correlated when the pair correlations were computed: operating expenditures and net contributions. For a deeper analysis, scatter plots were were color-coded by seat status and election. In Figure 3.5, it can be seen that the seats that were open are clustered at the bottom left (purple), while the incumbents are dispersed wider, demonstrating a wider range of net contributions. The challengers also appear to be dispersed throughout with no clear pattern. Thus, it can be seen that for the open seats, the amount of money contributed and expended overall tended to be less, perhaps because the chance of winning an election was higher when the seat is open, so less campaigning was done by the candidate. In Figure 3.6, which is coded by election outcome where “1” signifies a winner, it can be seen that those that lost elections are clustered towards the bottom left, while those that won show an overall larger range. In fact, the patterns between the two figures below are very similar, with the winners mirroring the incumbents and the losers mirroring the open seats.

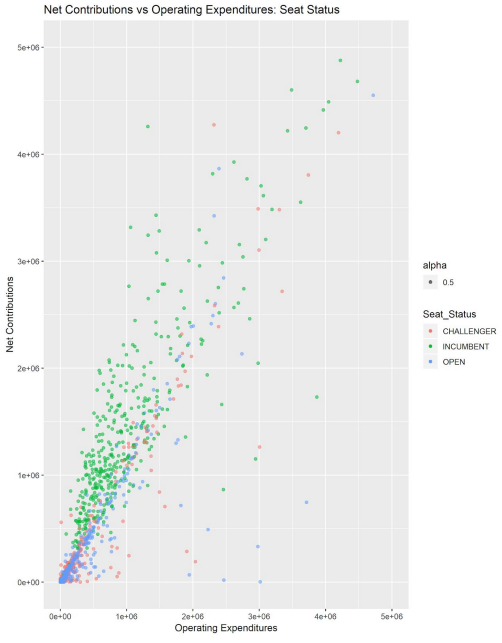

Figure 3.5: Scatter plot of Operating Expenditures vs. Net Contributions (By Seat Status)

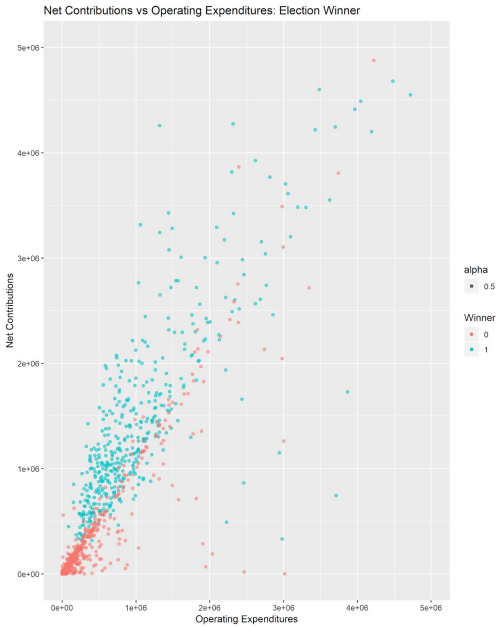

Figure 3.6: Scatter plot of Operating Expenditures vs. Net Contributions (By Winners & Losers)

Another type of visualization done was box plots. box plots are a useful method of visualizing summary statistics, and is a descriptive element that was used to compare total contributions in candidates of various seat status. As can be seen in Figure 3.7, the minimum, mean, and maximum are all significantly higher for candidates that were incumbents. This is an interesting observation, and suggests that contributions are on average significantly higher to incumbent candidates than they are to challengers or those running for open seats. Perhaps this is due to the name recognition, status, and reputation that is associated with being an incumbent candidate.

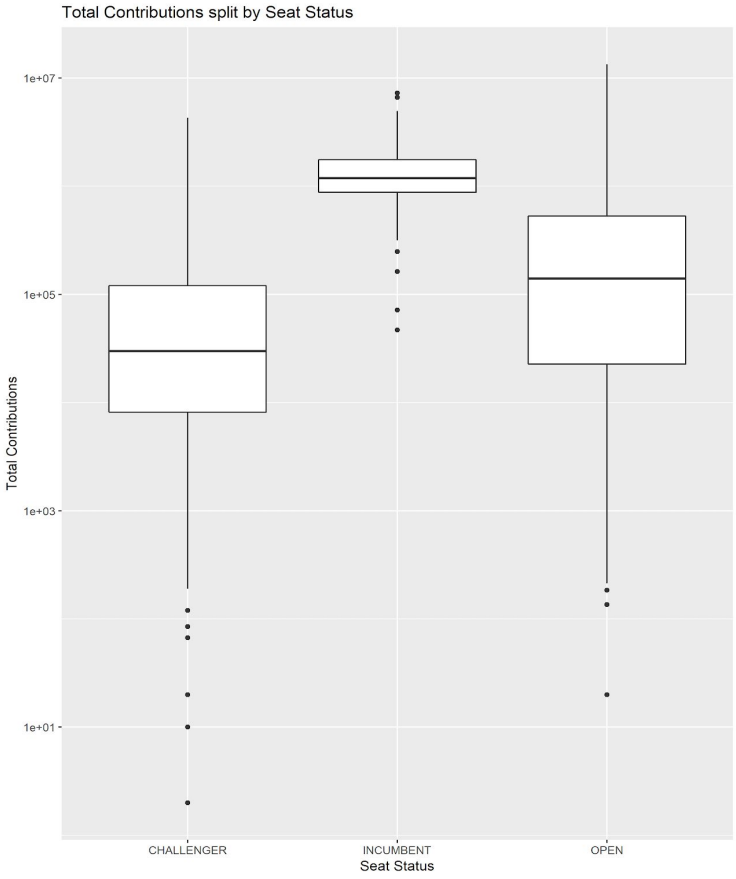

Figure 3.7: Box plot of Total Contributions Split by Seat Status

Lastly, a box plot was generated to visualize the total contributions between the two major parties, and in the context of the election outcome. While Figure 3.8 shows very little

difference in the contributions between the parties in each election outcome, it is clear that the candidates that win the election on average had significantly larger numbers of total contributions. Thus, there is once again evidently a very strong positive correlation between total contributions to campaign and a winning outcome to the election.

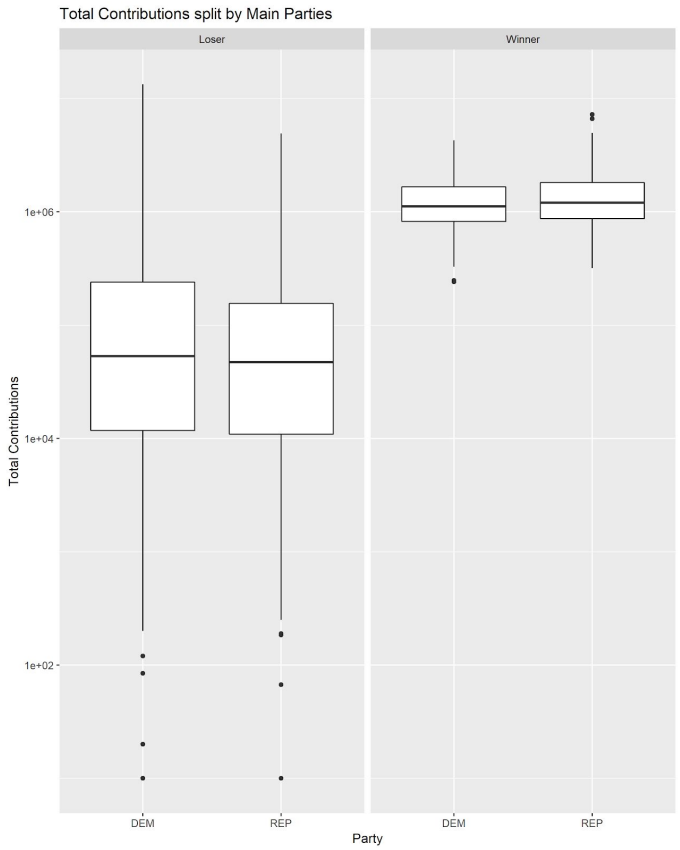

Figure 3.8: Box plot of Total Contributions Split by Main Parties and Election Outcome

The next phase of our project was to perform principal component analysis (PCA) on our data set. Our results were less than desirable. As can be seen from Figure 3.9, the variation was only represented 45.48% by the first principal component and 62.26% by both the first and second principal components. It can be seen in the biplot that most of the data is heavily concentrated, with long tails on one end. Thus, we did not attempt to cluster the data.

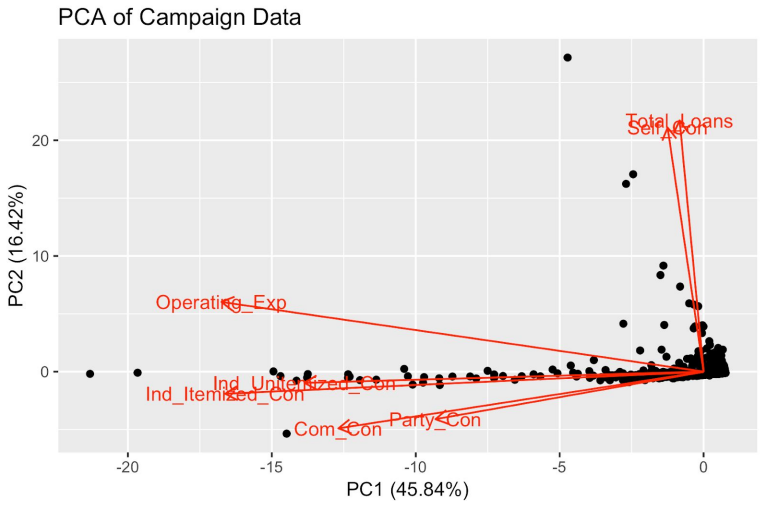

Figure 3.9: Plot of PCA Analysis

Lastly, we performed predictive modeling of election outcome versus several input variables using multiple classification methods. Initially, two simple logistic models were trained on the whole dataset for visualization purposes; one with a single predictor and one with two predictors. The single predictor model predicts the election outcome using total campaign contributions, while the two predictor model predicts the election outcome using net contributions and operating expenditures. Both models output a continuous value on the [0, 1] interval, representing the predicted probability of winning the election. We selected 0.5 to be the decision boundary in both cases.

The models are visualized in the following charts, and the accuracy of the models on the whole dataset are displayed in the following table. In the single predictor model, the red and green data points represent real data points (true winners and losers of the House elections), while the blue data points represent the predictions made by the model. In both graphs, we visualize the decision boundary as the red line. One observation made on the two predictor simple model is that the calculated decision boundary does not appear to be the optimal linear

decision boundary in the following chart (Figure 3.11). We heuristically explain this discrepancy to be the heavy clustering of losing candidates (red dots) near the lower end of both measures, while the winning candidates (blue) have a much higher average in both measures, especially net contributions, thus rotating the decision boundary clockwise. The error rates for the one and two predictor logistic models are displayed in Figure 3.12.

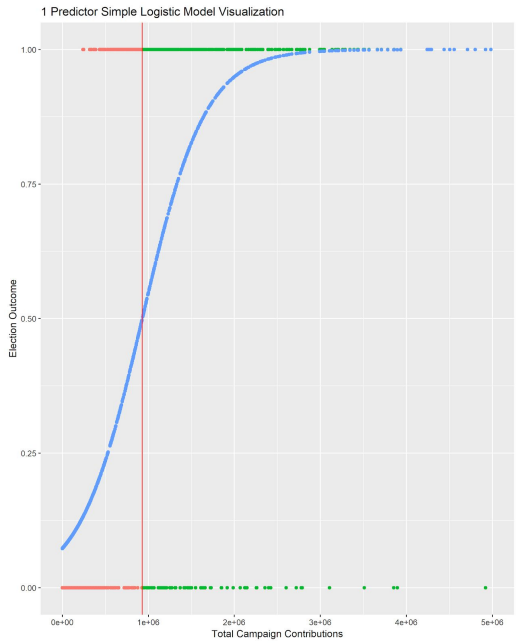

Figure 3.10: Single Predictor Simple Logistic Model Visualization

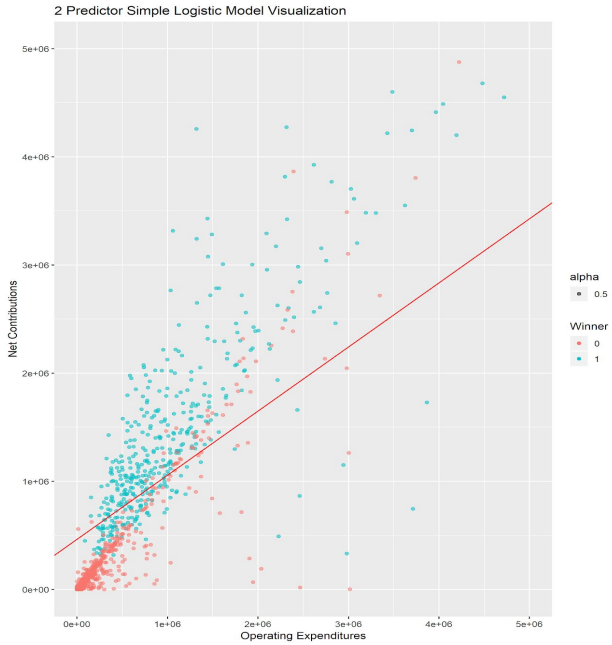

Figure 3.11: Two Predictor Simple Logistic Model Visualization

Model	Error Rate
Single Predictor Example Logistic	14.22 %
Two Predictor Example Logistic	11.42 %

Figure 3.12: Table of Error Rates for Logistic Models

Before moving forward, we briefly paused to examine the coefficients in the two predictor logistic model, detailed in Figure 3.13.

While operating expenditures are positively correlated with election win outcome on its own, we see that the coefficient value of operating expenditures is negative, representing a negative effect on election probability. We attributed this to the presence of autocorrelation between predicted variables (in this case, net contributions and operating expenditures), and decided that while our inclusion of multiple autocorrelated variables may limit our models interpretability, we included them anyways to minimize the misclassification rate in our future models.

Field	Coefficient Value
Intercept	-2.53
Net Contributions	5.45 x 10^-06
Operating Expenditures	-3.22 x 10^-06

Figure 3.13: Coefficients of Two Predictor Logistic Model

Next, we performed 10-fold cross validation with several predictive methods: logistic regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K Nearest Neighbors (KNN) with 3 and 5 nearest neighbors, and Support Vector Machines (SVM) with both Linear and Radial Basis Function (RBF) Kernels. In SVM, the hyperparameter(s) were tuned coarsely over a broad range (cost from 10^-3 to 10^3, and gamma from 10^-2 to 10^0 for RBF Kernel). The error rates for each model can be seen in the table below (Figure 3.14).

Model	Error Rate
Full Logistic	4.76 %
LDA	5.04 %
QDA	5.67 %
KNN k=3	4.55 %
KNN k=5	4.49 %

Linear SVM	5.18 %
RBF SVM	3.57 %

Figure 3.14: Table of Error Rates for Predictive Models

It is clear from the error rates that RBF SVM performs the best on the dataset, as its estimated prediction error rate is only 3.57%, which is much lower than the others. Thus, the decision was made to move forward with creating a final radial basis kernel SVM as the final model. Then, the first piece of making the final model was to extract an appropriate range for both hyperparameters of RBF SVM (cost and gamma). One of the RBF SVM models trained during the 10-fold cross validation was used to give a general range, with which the final model was returned around. The following are visualizations of error with respect to each parameter individually (Figure 3.15), and jointly (Figure 3.16), with the optimal cost/gamma combination displayed in red on the latter chart. In the latter, bottom two cost values were trimmed from the plot, as they produce much higher error values, and essentially make the model a linear decision boundary.

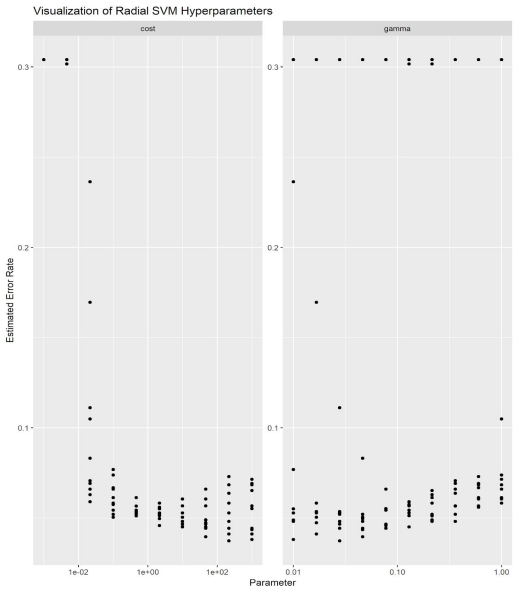

Figure 3.15: Individual Visualization of RBF SVM Hyperparameters

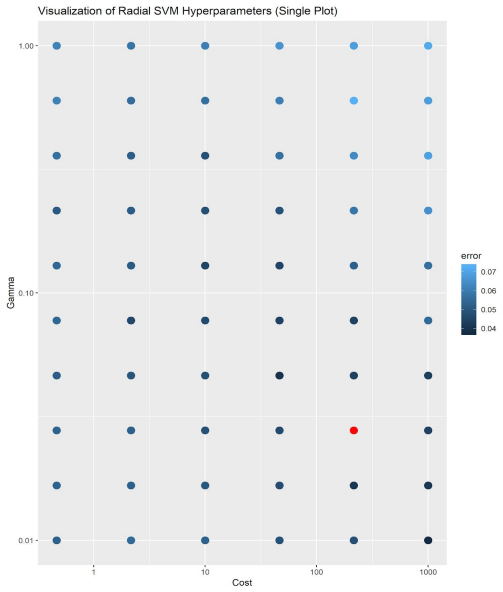

Figure 3.16: Joint Visualization of RBF SVM Hyperparameters

While no precise measure or range for what hyperparameters to move forward with were clearly obtained, the neighborhood of cost between 10 and 10,000 and gamma between 0.001 and 0.1 seemed reasonable. After training SVM models over a base 10 exponential range of that neighborhood with the whole dataset, an optimal model at cost 501.18 and gamma 0.02 was produced. On the whole dataset, this model had an error rate of 1.54%, and contained 127 support vectors. While this is still a significant number, it is a large reduction from the 1,427 initial data points. The model is not particularly biased to predicting either election result, as is shown in Figure 3.17. In Figures 3.18 and 3.19 respectively, election results and predicted election results are plotted.

	Predicted Loss	Predicted Win
Election Loss	979	13
Election Win	9	526

Figure 3.17: Confusion Matrix of Final Model

Figure 3.18: Visualization of True Values

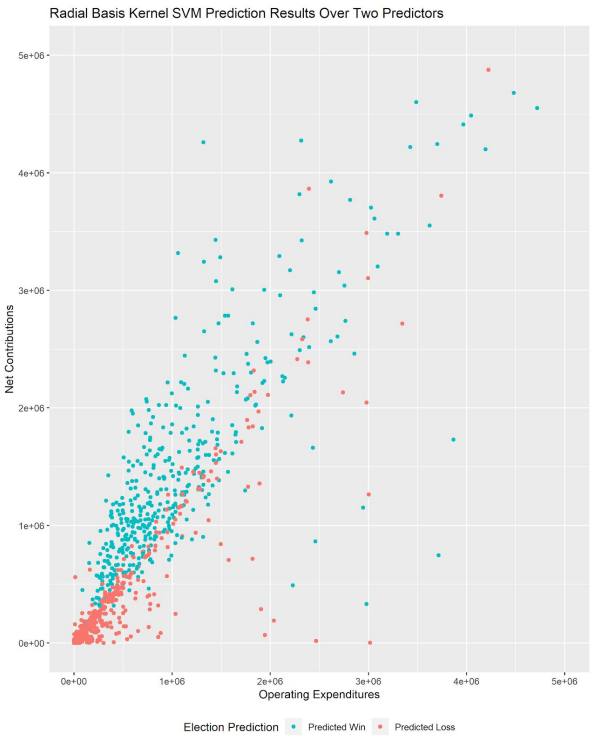

Figure 3.19: Visualization of Final Model Predictions

We next combine the two charts into one, with incorrectly predicted data points plotted in red. We see no particular grouping of the 22 mispredicted points, and so conclude that the only error in the model is due to noise in data, rather than an underlying issue. The visualization of this plot can be seen in Figure 3.20.

Figure 3.20: Visualization of Final Model Mis-Classifications

To gain a deeper understanding of the model, two final visualizations are presented. The first one is simply of the support vectors and their magnitudes on our same two predictor variables (operating expenditures and net contributions). In this visualization, dark blue points represent a strong vote for election loss, while light blue points represent an election win. Unfortunately, no real trend could be discerned from this visualization, which is shown as Figure 3.21.

Finally, a visualization of the predicted probabilities of the model at synthetically generated levels of net contributions was created (all other variables are held constant at their means). Two faceted plots of this data are presented, one by party and one by seat status (Figure 3.22). From these charts several observations can be made. Firstly, as one would probably expect, all of these marginal functions appear quite sigmoidal in nature. Secondly, and also in line with expectations, from the second faceted plot, it can be observed that incumbent candidates are predicted a much higher win percentage rate at similar levels of contributions, across all party classes. Next, it is noted that Democrats appear to have a much lower win rate than than both Republicans and third party candidates. Apart from the straightforward interpretation that Democrats are at a systemic disadvantage, this may be showing that net contributions are not as significant for Democratic candidates as opposed to other fields, or that this is a result from a Republican candidate winning the presidential race, potentially due to the "straight-ticket voting" effect where voters vote for the same party across the ballot.

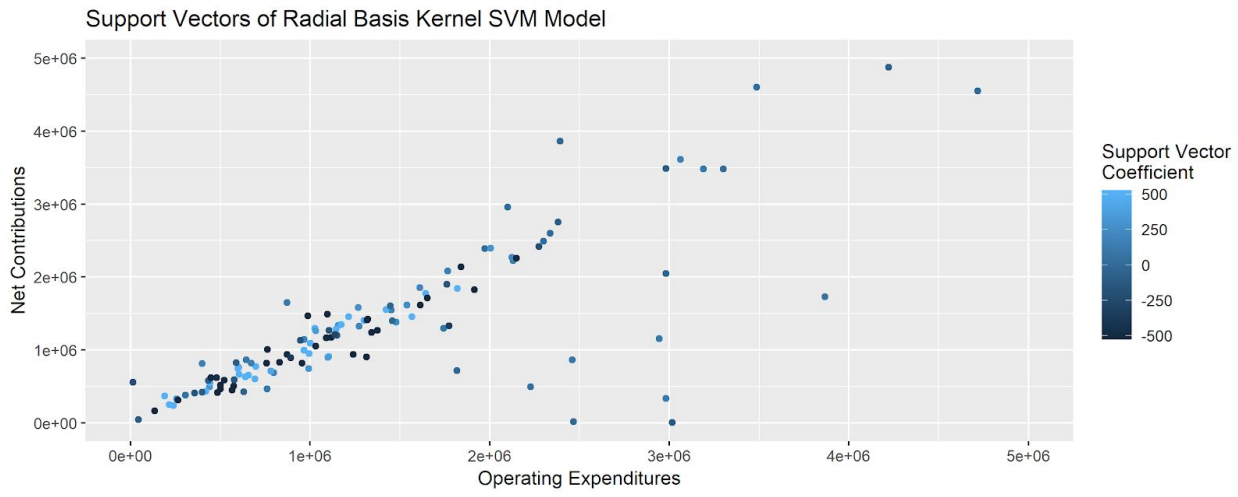

Figure 3.21: Visualization of Final Model Support Vectors

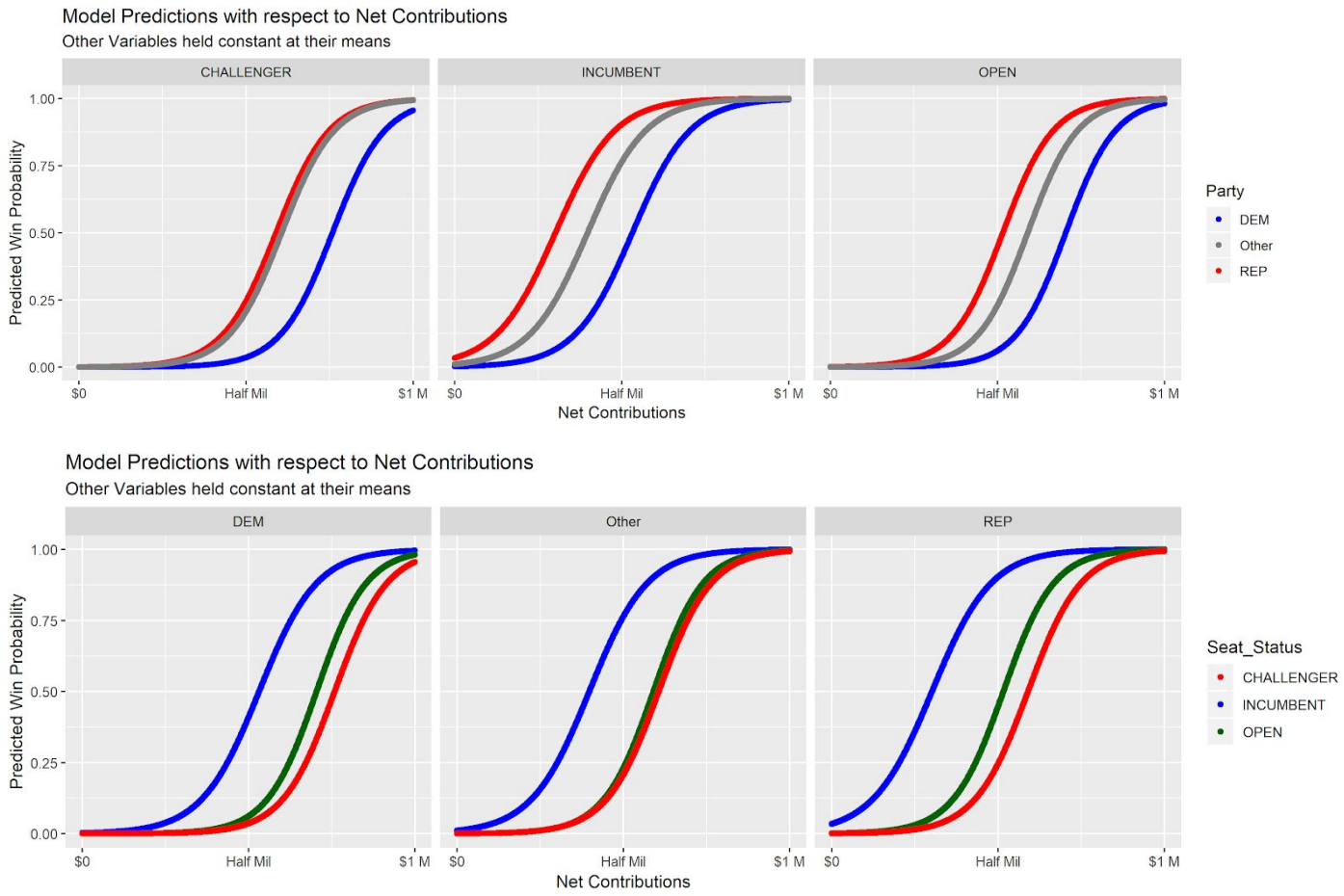

Figure 3.22: Visualization of Win Predicted Probability over Net Contributions with other fields held constant

4.0 Conclusions

Upon the start of this analysis, after data clean-up, the working dataset produced was first examined for trends and correlations using descriptive elements. The first set of plots produced were histograms of the numerical variables, from which we were not able to deduce much about the distributions other than that they were exponential in nature, due to the need for a logarithmic x-scale. It was clear that there was a wide range in each variable, signaling the large disparities between the finances of the various candidates. Next, the pair correlations between the numeric variables were plotted in an attempt to gain an overview of any variables that may be correlated. As Figure 3.2 showed, many of the significant correlations were obvious (for example, individual itemized contributions and total individual contributions), but one that was unique (though intuitive) was the strong positive correlation between operating expenditures and various contribution variables. This led us to the conclusion that the candidates that raised large amounts of money also tended to spend large amounts of money.

The second set of descriptive plots that were produced were scatter plots. From Figure 3.3 it can be concluded that large amounts of individual contributions correlated with a winning election outcome, implying that larger amounts of money raised from individuals had an influence on a candidate's ability to win the election. Additionally, from Figure 3.4, our work shows that while high amounts of committee contributions correlated strongly with winning elections, party contributions had little to no effect on the election outcome. In an attempt to shed light on a different perspective, Figures 3.5 and 3.6 were generated to visualize the relationship between operating expenditures and net contributions. Figure 3.5 is color-coded according to seat status and Figure 3.6 is color-coded according to election outcome. A noticeable observation is that the clusters look very similar in these graphs. In particular, the elections in which the candidate was running for an open seat display a very similar trend to the

elections in which the candidate lost. Likewise, the elections in which the candidate was an incumbent display a very similar trend to the elections in which the candidate won, with this cluster trending towards higher numbers of total contributions. Thus, there seems to be a relationship between the fact that a candidate is an incumbent, a candidate's ability to raise larger amounts of money (perhaps due to name recognition and existing reputation), and a candidate's ability to win an election.

The third set of descriptive elements were box plots, which serve as a very useful visual of summary statistics. Figure 3.7 shows the differences in total contributions between candidates of different seat statuses, and clearly concludes that incumbent candidates on average gained larger amounts of contributions than challengers and open seats. Additionally, Figure 3.8 plots total contributions in the context of election outcome and major political party (Republicans and Democrats). While there seems to be little difference in contributions between the major parties, it is evident that candidates that won their elections, on average, gained much larger amounts of contributions, as the first quartiles of the winners in both parties are higher than the means of the losers in both parties.

Unfortunately, the Principal Component Analysis that was performed did not produce desirable results. One of the weaknesses of PCA is that it relies on linear relationships to explain observations about the data set. Our data set did not have many linear relationships when we ran the analysis and we were not able to then cluster the data via a clustering algorithm. The biplot showed some correlation between data but did not show any distinct groupings of data as would be desired.

Finally, we performed predicting modeling on the dataset. We first fit simple logistic regression models on one and two variables, where we computed 85% accuracy on a simple testing set. Interpreting the coefficients of these models, we again saw the issue of

autocorrelation in our dataset, as some variables which have relatively strong positive relationship with election wins have negative signed coefficients. While our model interpretations would thus be limited moving forward, we decided to move forward with all fields (as opposed to doing variable selection) and used more advanced methods which were already somewhat black box in nature, thus making our sole goal prediction accuracy.

We next used a 10-fold cross validation approach to select a model from the following methods: Logistic, LDA, QDA, K-Nearest-Neighbors (with both 3 and 5 neighbors), and both Linear and RBF SVM. Some models, such as KNN, were limited in their use of our categorical variables, but we included the variable Seat_Status as a natural ordering does occur: 0 for challenger, 1 for open seat, and 2 for incumbent. The average error rates across the folds are displayed in Figures 3.12 and 3.14.

We then moved forward with the RBF SVM method as it had the lowest estimated mis-classification rate of 3.57%. We used the whole dataset to train the final model, as is recommend by Hsu, Chang, and Lin in “A Practical Guide to Support Vector Classification.” We tuned the models hyperparameters over a finer range, which we selected upon observing the error over the different cost and gamma combinations tested in the cross validation portion of our analysis. Our final model boasts a 98.5% prediction accuracy rate, misclassifying just 22 of the 1,427 observations in our dataset. The model only holds 127 support vectors, a significant reduction from the original observations. Finally, through some graphical analysis, we found that Democrats appear to face a lower election probability than their Republican and third party counterparts at similar net contribution levels. Election challengers face a similar scenario against incumbent candidates. Overall, House elections were very predictable based upon campaign finance data for the 2016 election cycle.

5.0 References

"Campaign Finance versus Election Results." *Kaggle*, Kaggle, 7 Dec. 2016, www.kaggle.com/danerbland/electionfinance

"Metadata Description." *Metadata for Candidate Summary*, Federal Election Commission, classic.fec.gov/finance/disclosure/metadata/metadataforcandidatesummary.shtml.

Hsu, Chih-Wei, et al. A Practical Guide to Support Vector Classification. www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.