# Predictive, Agent-based, and Causal Machine Learning Models of U.S. Congressional Elections

By

Parker Quinn

Thesis Project

Submitted in partial fulfillment of the

Requirements for the degree of

MASTER OF SCIENCE IN PREDICTIVE ANALYTICS

June 2018

Dr. Joe Wilck, First Reader

Dr. Nick Mastronardi, Second Reader

## ABSTRACT

**Predictive, Agent-based, and Causal Machine Learning Models of U.S. Congressional Elections**

**Parker Quinn**

This paper uses district-level data to highlight similarities and differences of a variety of U.S. congressional election models ranging from predictive statistical learning models to causal structural models and new causal statistical models. First, several statistical learning models are produced with high out-of-sample accuracy. Second, the causal impacts of candidate spending, incumbency, and voter registration are estimated using a structural agent-based model as in Kretschman and Mastronardi (2010). Finally, a new environment for election modeling – machine learning for causal inference – is introduced. A demonstration is provided in the form of a "causal forest," which uses machine learning to produce accurate predictions and estimate heterogeneous treatment effects. The causal impact estimates from the agent-based and causal forest models are used to develop counterfactual arguments describing the efficiency of allocated candidate spending across districts. The results from this research are valuable to many stakeholders, including candidates, party committees, and voters.

*Dedicated to*

My parents – my biggest advocates, who taught me that anything can be done with good planning.

*"There is no reason why good cannot triumph as often as evil. The triumph of anything is a matter of organization." – Kurt Vonnegut*

Colette – my partner in all adventures, who has given me more than I could ever ask for.

*"That's what I consider true generosity: You give your all and yet you always feel as if it costs you nothing." –Simone de Beauvoir*

**Acknowledgments**

My readers were immensely helpful in this project's completion and in my professional development, taking time from their busy lives to provide valuable guidance and feedback. Special thanks to Nick Mastronardi, who has been a personal mentor of mine for years – I certainly would not be where I am today without his generosity. Thanks also to Kyle Kretschman, Julie Tibshirani, and Stefan Wager for providing additional insights on their research.

I would also like to thank my family and friends for their continued support, especially my siblings and their children. They have helped me through all of my endeavors and constantly remind me of the important things in life.

Finally, I owe a debt of gratitude to my phenomenal colleagues at AMC/A9, who encouraged me to pursue this opportunity and helped me manage my time between school and work. Investing in people's personal growth, as they do, is the mark of a first-rate organization.

## Table of Contents

## List of Tables and Figures

# Chapter 1 Introduction

The 2016 elections were the most expensive on record, with over $6 billion spent by candidates and committees (Center for Responsive Politics). Campaign expenditures are a common topic of discussion during election cycles, but surprisingly, candidates and party committees often seemingly misallocate funds. In Maryland's 8th congressional district, one that heavily favors Democrats, Democratic candidates spent a cumulative $22 million during the 2016 campaign. Conversely, Democrats narrowly lost the congressional election in California's 49th, a district that voted for Hillary Clinton, having spent relatively little during the campaign. Republicans missed similar opportunities while overspending in districts like Louisiana's 3rd, where they spent nearly $11 million and won easily, securing over 80% of votes. Candidates and parties could avoid these mistakes and draw attention to other important aspects of elections with innovative statistical models.

To make such normative statements confidently, analysts must employ statistical learning models with practical predictive accuracy, but also control for potentially confounding causal factor impacts using structural academic models. For example, before assessing the efficiency of partisan spending, careful causal modeling of the effect of expenditures controlling for district-specific environmental characteristics is necessary. This process is typically performed separately from any predictive modeling efforts. However, some recent causal machine learning models seem to deliver both the accuracy of predictive statistical learning models and the causal confidence of structural models, resolving this dichotomy.

The need for modeling methodology reconciliation is not unique to politics. Large e-commerce companies use predictive models to forecast total sales in warehouse re-supply models and causal structural models to assess the efficacy of marketing campaigns on sales and

profitability. Modern analysts should be comfortable with both approaches, but need clarification on which environments are most appropriate for each and a deeper understanding of their differences.

Politics is an ideal environment to investigate these approaches. Because ballots are private and only aggregate vote shares are observed, rigorous structural attribution modeling is essential for quality causal analyses. However, these models have failed to incorporate the predictive accuracy and up-to-date reporting that the national stage and intense 24-hour news cycle demand from the industry. Consequently, a separate class of predictive political models has emerged to satisfy non-rigorous observers focused only on outcomes, polarizing the field of political election modeling.

Lewis-Beck and Tien (2016) divide forecasters of the recent past into modelers, poll users, marketers, and experts. They say, "To oversimplify, the modelers traditionally use single-equation statistical models derived from election theories, the poll users follow single- or aggregated-survey estimates, the marketers study candidate trading prices in political stock markets, and the experts employ judgments from the campaign trail." These models effectively fall into two categories – those that stress predictive accuracy and those that explain election outcomes. Regardless of the model used, the political modeling community persists with the notion of an inevitable trade-off between accuracy and research value.

The models presented and estimated in this paper focus on elections of the U.S. House of Representatives. Political science researchers have long attempted to explain the determinants of electoral outcomes, particularly in national elections like those for the House, Senate, and Presidency. Pioneering work by Kramer (1971) and Tufte (1978) on economic voting led to a growth in understanding about the relationship between economic factors and election outcomes.

Others worked to uncover truths about the pattern of presidential party seat loss in midterm congressional elections, most notably Tufte (1975), Lewis-Beck and Rice (1984), and Campbell (1985). Levitt and Wolfram (1997) decomposed the sources of incumbency advantage in U.S. House elections. Perhaps the most contentious area of research has been in the realm of campaign spending in congressional elections, with contributions from Jacobson (1978, 1990), Green and Krasno (1988), Gerber (2004), and Levitt (1994). Although scholars disagree on the effects of these factors on election outcomes, the accumulation of their work has produced a body of theory from which election models can emerge.

Despite this thorough examination by scientists, relatively few political forecasts predict congressional elections, and three substantial gaps remain in the existing body of work. First, election models that use the vast amount of publicly available district-level data and recent advances in statistical learning techniques to predict outcomes in individual districts are scarce. Second, structural models of congressional elections that identify determinants of election outcomes are still necessary considering the lack of consensus in certain areas. Furthermore, most contemporary literature relies on reduced-form econometric models, rather than experimental or structural approaches. Third, it is not clear that an adequate middle ground exists – models that are highly accurate, based on sound theory, and allow for estimation of causal effects. Most modelers consider themselves as part of one group or the other, focusing on either predictive accuracy or electoral theory. Both approaches have merits, but no attempt has been made to bridge this gap despite the rapid development of novel statistical techniques.

### Section 1.1 Guiding Research Questions

The goals of this research align with those of empirical election modeling in general – to accurately predict elections and to understand the factors that impact the outcomes. This paper

contributes to each goal separately while addressing the questions: How do models from these two perspectives compare, and is there an approach that combines their respective advantages?

The following objectives provide a framework for addressing the goals and questions of this paper:

1. Using rich sources of district-level data available before Election Day, develop accurate machine learning models to predict individual congressional election outcomes.
2. Build on academic knowledge of election mechanisms by quantifying the causal impacts of campaign spending, incumbency, and voter registration using an agent-based discrete choice model. Use the causal impact estimates to analyze candidate spending in individual districts.
3. Combine the advantages of 1 and 2 with a new application for a statistical technique that estimates causal effects from a predictive machine learning algorithm.

**Section 1.2 Justification**

The predictive modeling portion of this research tests the efficacy of statistical learning methods in the arena of political forecasting, which has been underserved in this regard. Election models are useful to many stakeholders who have a vested interest in the outcomes. Producing accurate and parsimonious predictions can inform the decisions of candidates, media outlets, voters, donors, and other political establishments. They also satisfy our natural curiosity and desire to quantify uncertain events.

The structural modeling portion of this research aims to add to existing knowledge of election mechanisms, with emphasis on the impacts of candidate expenditures, incumbency, and the distribution of registered voters. Concern about the consolidation of political power is deeply ingrained in the American national identity, and the issue of campaign finance reform has received renewed attention. If money is influential in electoral outcomes, then citizens would rightly be worried about wealthy factions wielding extreme political power. Conversely, if spending is inconsequential to election outcomes, perhaps policy advocates should focus elsewhere, such as the incumbency advantage or districting process. These are vital aspects of a healthy republic, and policymakers and citizens have a stake in all of these factors. Furthermore, the estimates produced by causal models can be used to develop counterfactual arguments – changes to the environment that would have led to a different outcome – that can inform decision-makers such as candidates or party committees.

It is possible that many modelers are not aware of one or more of these methods (or weaknesses of their preferred method). Although each of these modeling efforts is useful on their own, the portion of this research that uses machine learning for causal inference combines the advantages of both. With further research in this area, it may be possible to satisfy the desire for highly accurate predictions in a way that also adds to the understanding of election mechanisms.

## Chapter 2 Review of Literature

Louis Bean achieved fame for predicting the Truman over Dewey upset in the 1948 presidential election in what is widely considered the first empirical political forecast. Since then, statistical modeling of elections has become commonplace, and many scientists have made significant contributions to the field. The work of Michael Lewis-Beck is highly regarded - Lewis-Beck (2005) and Lewis-Beck and Tien (2016) provide historical narratives, contrasting perspectives, modeling techniques, and tested guidelines for scholars interested in studying elections empirically.

In machine learning, models are designed to maximize predictive accuracy by finding hidden patterns in the data. However, it is also helpful to specify patterns known to be important, which can be informed by the current body of knowledge.

Two factors that are known to affect congressional elections are incumbency and midterm loss. Levitt and Wolfram (1997) show that incumbents typically enjoy reelection rates of over 90%, and ascribe this advantage mostly to their ability to deter quality challengers. Others have tried to measure and explain the consistent loss in congressional seats experienced by the sitting President's party during midterm elections. Tufte (1975), Lewis-Beck and Rice (1984), and Campbell (1985) offer several possible explanations. They describe a "regression to the mean" theory, in which the presidential party performs better in on-year House elections, and a "referendum" theory, which attributes midterm loss to presidential popularity and economic conditions. Erikson (1988) evaluates these theories and also contributes the idea of a "presidential penalty," arguing that the electorate tends to penalize the "in" party.

Individual congressional candidates, especially incumbents, may also be evaluated on economic conditions. Kramer's (1971) seminal article established a link between economic and electoral outcomes, and his findings were supported by Tufte's (1978) later work. Both theorized that voters viewed economic factors as a way to hold a referendum on the President and his party. Lewis-Beck and Stegmaier (2000) review research in this area and conclude that economic conditions often drive voters to reward or punish those currently in power.

Debate on the effects of campaign spending on election outcomes has been very contentious as scholars explore methods to address the issue of endogeneity, often finding contradictory or counterintuitive results. Campaign finance has been a focus of election research since the groundbreaking work of Jacobson (1978). Using OLS regression, he found that challenger spending has a much stronger impact than incumbent spending. A decade later, Green and Krasno (1988) question Jacobson's assumptions and refute his theory by using instrumental variables to show that the effects of incumbent and challenger spending are roughly equal. Jacobson (1990) responds with an explanation of how this criticism comes up short. Levitt (1994) proposes a new method for controlling for candidate quality, finding that campaign spending by any candidate has a minimal impact on election outcomes. More recently, Gerber (2004) uses field experiments to measure spending effects, ultimately agreeing with Jacobson's original argument. Scholars and citizens agree that this is a necessary topic of research, especially as the link between money and politics becomes more apparent, but have not reached a consensus.

The issue of ideological extremity among elected officials has become a topic of national discussion in recent years and is starting to gain the attention of researchers. Butler, Lee, and Moretti (2004) investigated whether voters affect ideological positions of candidates, or if voters

merely choose among candidates that independently take policy views. They find that voters elect policies and that an exogenous shift in electoral strength for a particular party does not affect candidate ideology. Pyeatt (2014) looks at the relationship among incumbent ideology, district ideology, and challenger quality in congressional elections. He finds that these two factors produce diverging incentives for incumbents hoping to discourage strong challengers from entering the election. Candidate ideology is still a relatively new area of study for election research, and many unknowns remain.

In addition to this research, current statistical predictions of elections can guide the decisions regarding the variables to include in the machine learning models. Klarner (2008) and Hummel and Rothschild (2014) develop statistical models of various elections and include predictive variables such as past election results, ideology, and economic variables. Some political experts also forecast congressional elections using their professional judgment. Lewis-Beck and Tien (2014) consider models that mix traditional statistical approaches and expert political knowledge, called "Structure-X" models, finding that combining methods can dramatically reduce prediction error.

When choosing the models themselves, many options are available for both classification and regression problems. James, Witten, Hastie, and Tibshirani (2015) provide an excellent guide for best practices when working with typical machine learning models such as regression, trees, and support vector machines. Additionally, Polley (2018) presents a method and programming package for combining several machine learning algorithms into one "Super Learner," which can increase predictive accuracy and inform modeling choices.

Most research into electoral mechanisms uses reduced-form statistical models. The work described above by Jacobson, Green and Krasno, and Levitt on the relationship between

campaign spending and election outcomes all use reduced-form models. Kretschman and Mastronardi (2010), on the other hand, use a structural model to estimate the impacts of congressional election spending with data from the 2002, 2004, 2006, and 2008 elections. Their model is based on a structural model of discrete demand by Berry, Levinsohn, and Pakes (1995), which has been used across many industries to measure the causal impact of product characteristics. Using incumbency and expenditures as product characteristics, and voter registration statistics to account for consumer heterogeneity, Kretschman and Mastronardi find small effects of spending on outcomes.

Models that allow for causal inference, such as the one by Kretschman and Mastronardi, are essential to research in many disciplines. Policymakers in medicine, business, and government use the results from these models to make informed decisions and explore counterfactuals. However, limitations to causal models often arise in analyses with many population subgroups and high dimensionality in the data. Furthermore, they often require the analyst to strictly define a structural equation by relying on in-depth knowledge of the system being modeled. Treatment effect estimates are often averages or only apply to specific groups of the population. Researchers in machine learning have created models that can make sense of high-dimension data and find hidden patterns in the data to make accurate predictions but cannot be used for causal inference. However, there has been a growing set of literature dedicated to machine learning methods for causal inference. These models do not impose requirements on the analyst, such as a structural model, but can still estimate causal impacts while maintaining high predictive accuracy. Athey and Wagner (2017) present a method for “causal forests” that use a conventional machine learning algorithm – the random forest – to estimate causal impacts. This

model is an example of a new way to approach election forecasting with causal machine learning.

## Chapter 3 Data Collection and Preparation

The raw data for this research consist of one observation for each U.S. House of Representatives election from 2006 to 2016. The features include candidate information (incumbency status, years in office, party, ideology score, campaign expenditures), election information (party in control of the presidency, midterm status), and district information (partisan lean, demographics, economic indicators, previous election results). A detailed table of the data collected is shown in Appendix 1. Most of the data come from open sources or agencies such as the Federal Election Commission (FEC) and U.S. Census Bureau, which make the information publicly available on their websites. Acquiring voter registration data can be arduous and expensive, so registration data for only the 2016 elections were purchased from Catalist, a for-profit company that aggregates and sells registration data.

The raw dataset went through a series of cleaning steps to remove erroneous observations and repair inaccuracies. Data sources have unique ways of encoding their information, so inconsistencies required further research on individual districts, elections, or candidates. Additionally, some district demographics were not available during certain election years or were aggregated across several years. There were also complications when accounting for the redistricting that occurred in 2010. Where this occurs, the most recently available information was used, or the observations were excluded in modeling. Several additional features were created by combining the collected data into new variables for modeling (Appendix 2). These features are informed by theory and findings from previous research. Features are often re-scaled from existing variables so that they are better suited for modeling or converted to reflect how well the district and the incumbent candidate are matched. Note that although many additional features were created, not all of them end up being used (Table 1).

| *Dataset* | *Observations* | *Variables* | *Notes* |
|---|---|---|---|
| *Initial (raw)* | 2,610 | 86 | 435 districts x 6 election years = 2,610 elections |
| *Cleaned* | 2,271 | 86 | Elections without challengers excluded, anomalies removed |
| *Reduced for modeling* | 2,271 | 55 | Non-relevant variables excluded |

*Table 1 - Dataset cleaning and reduction*

The final dataset is suitable for modeling and can be manipulated depending on the types of models created. For example, it may be appropriate to exclude elections without a challenger or exclude voter registration data, which is only available for the 2016 elections. Finally, it is good practice to exclude extraneous variables to preserve parsimony. These decisions are made on a case-by-case basis for the models described in the sections below. For simplicity, special subsets of dependent variables were created for these models (Table 2), and exploratory graphs for a selected subset of continuous variables are in Appendix 3.

| *Subset A* | *Subset B* |
|---|---|
| ▪ Previous election incumbent vote share<br>▪ Third party indicator<br>▪ Percent change in median household income<br>▪ Change in unemployment rate<br>▪ Midterm effect<br>▪ Incumbent expenditures (log transformation)<br>▪ Partisan Voter Index (PVI) effect<br>▪ Expenditure difference<br>▪ Sabato score<br>▪ Open election indicator | ▪ All variables from Subset A<br>▪ Incumbent's party<br>▪ Election year |

*Table 2 - Dependent variable groupings*

Four variables are of particular interest in this dataset: the expert forecast produced by Larry Sabato, and the three features that were created to capture the midterm effect, partisan voter index (PVI) effect, and the incumbent ideology effect. The impact of midterm elections on

the incumbent's vote share appears to be substantial (Figure 1). Incumbents in the same party as the sitting President during a midterm election typically see much lower vote shares than those in the opposite party, or when it is an on-cycle election.

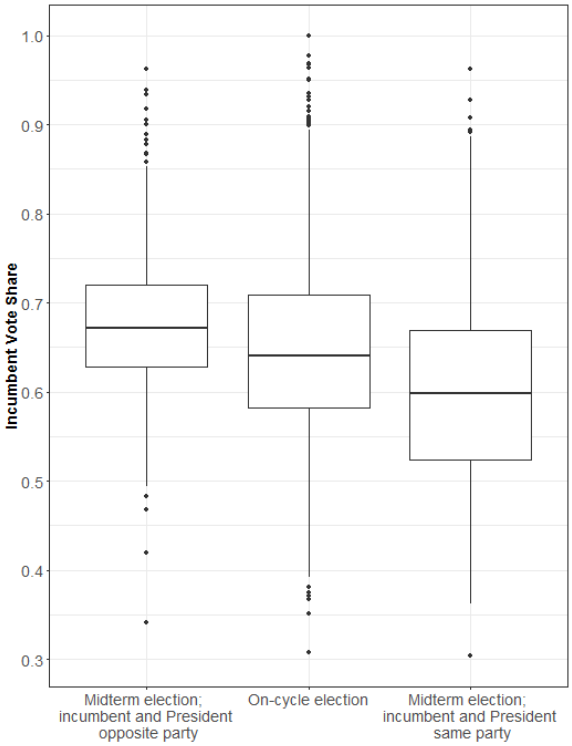



*Figure 1 - Midterm effect versus incumbent vote share*

The inclusion of district partisan lean and incumbent ideology are also critical to model accuracy – they have strong relationships with actual outcomes. In general, as PVI increases, the Democratic candidate's vote share decreases (Figure 2). Incumbent ideology scores can be normalized by district partisan lean to create the ideology effect, which has a positive relationship between incumbent ideological extremity and vote share (Figure 3). As expected, incumbents with ideological scores that go against their district's ideology (i.e., the ideology effect variable is negative) lose more frequently.

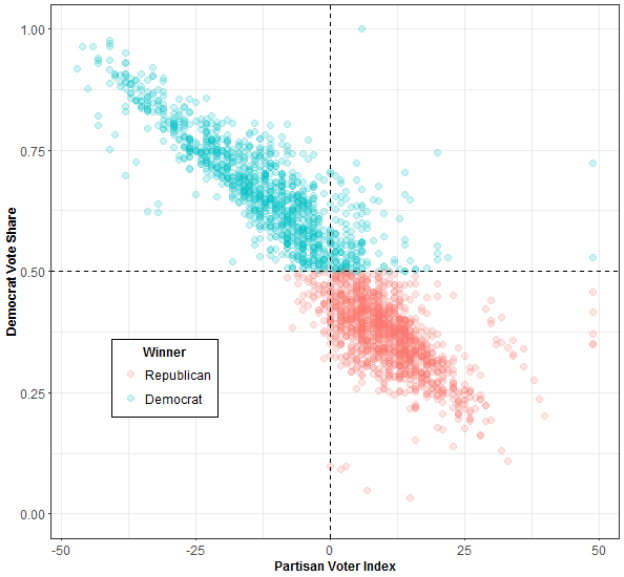



*Figure 2 - District PVI versus vote share*

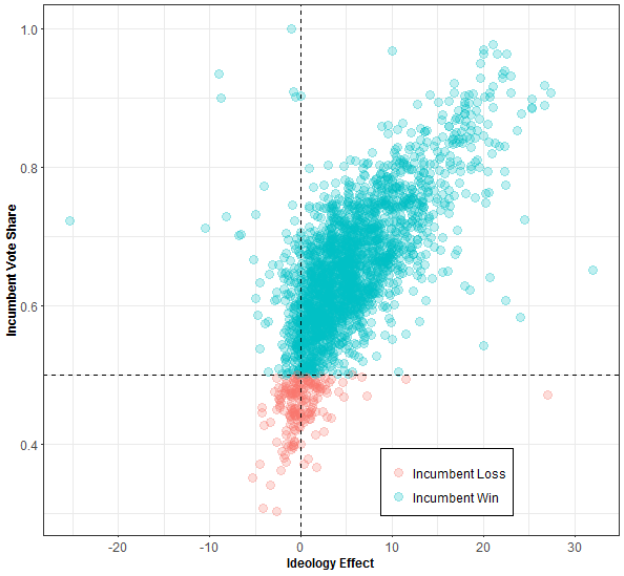



*Figure 3 - Ideology effect versus incumbent vote share*

Finally, the expert forecasts from Larry Sabato appear to have an imperfect but relatively strong relationship to actual outcomes. Sabato's predictions lose accuracy as they get closer to a "toss-up," indicating that his methodology is very consistent (Figure 4).

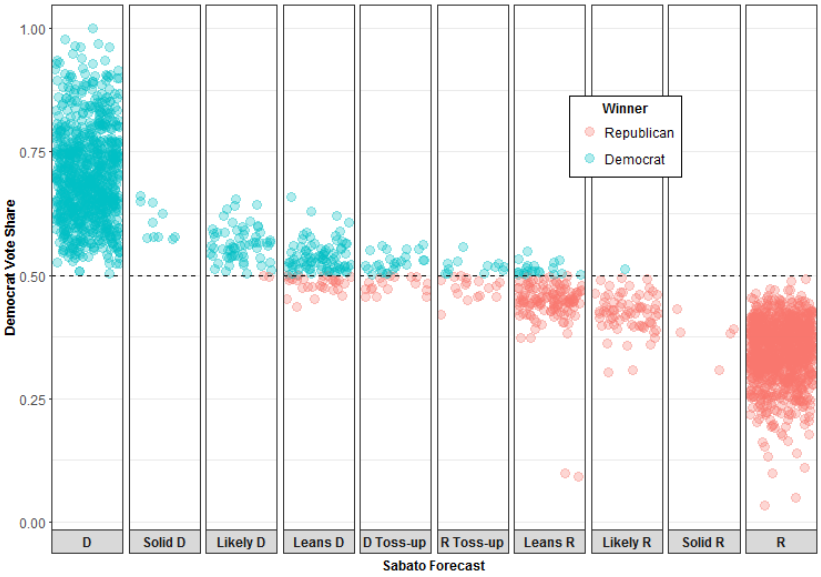



*Figure 4 - Sabato forecast accuracy*

# Chapter 4 Predictive Models

In this chapter, several statistical learning models are built with a focus on predictive accuracy of U.S. House elections.

## Section 4.1 Methods

Recalling the "No Free Lunch" theorem (Wolpert 1996), no single machine learning algorithm outperforms all others for any given problem. Hence, it is important to try several models (and variations of these models) to arrive at acceptable levels of interpretability and accuracy. With hundreds of models to choose from, this research relies heavily on the models described in "An Introduction to Statistical Learning" (James, Witten, Hastie, and Tibshirani 2015), an influential book in this field.

The models in this section predict outcomes for the incumbent party's candidate (henceforth referred to as the "incumbent"). Across the entire raw dataset, incumbents enjoyed a significant edge, winning 2,418 of 2,610 of elections (Table 3).

| | *Win* | *Loss* |
|---|---|---|
| *Incumbent outcome* | 2,418 | 192 |

*Table 3 - Incumbent outcomes*

As a result, a naïve model (one that predicts an incumbent victory in every district) election has a baseline accuracy of about 93%. Any model that is less accurate than this is worse than guessing, so this paper focuses on incumbent candidate outcomes to maximize predictive accuracy. Two types of models are produced – classification models, which predict whether the incumbent is expected to win or lose, and regression models, which predict the incumbent's two-party vote share.

Elections without a challenger from the opposite major party (usually indicated by an incumbent that receives 100% of the two-party vote share) and that occurred in districts with significant redistricting were excluded. These models are trained on data from 2006 to 2012 and evaluated in the 2014 and 2016 elections, allowing the models to learn on two-thirds of the data and test their accuracy on the remaining third to avoid overfitting. It would be within common practice to use only 15-20% of the data for the hold-out dataset, but here data from one midterm election (2014) and one on-cycle election (2016) are used for evaluation.

Each of these models uniquely estimates the response and makes different assumptions about the explanatory variables and error term. Furthermore, the models have their own strengths and weaknesses with regard to accuracy and interpretation of results. As a result, some models include fewer variables than others and are valuable in different ways (Table 4). For example, in logistic or linear regressions, it is better to include fewer variables (all else equal), but this is less of a concern in a random forest. Although regression algorithms are more restrictive in size, it is generally easier to interpret their results.

| *Model* | *Dependent Variables* | *Methods and Tuning Parameters* |
|---|---|---|
| *Linear Regression (VS)* | Subset A | N/A |
| *Logistic Regression (WL)* | Subset A | N/A |
| *k-Nearest Neighbors (WL)* | Subset A | Scaled variables<br>k = 7 |
| *Random Forest (VS)* | All variables | Variables sampled per node = 15<br>Number of trees = 1000 |
| *Random Forest (WL)* | All variables | Variables sampled per node = 15<br>Number of trees = 1000 |
| *Tree (VS)* | Subset B | N/A |
| *Tree (VS)* | Subset B | Cross-validation and pruning<br>Nodes = 6 |
| *Tree (WL)* | Subset B | N/A |
| *Gradient Boosting (VS)* | Subset B | Number of trees = 5000<br>Interaction depth = 3<br>Cross-validation folds = 5 |
| *Gradient Boosting (WL)* | Subset B | Number of trees = 3000<br>Interaction depth = 3<br>Cross-validation folds = 5 |
| *Adaptive Boosting (WL)* | Subset B | Number of iterations = 3000 |
| *Naïve Bayes (WL)* | Subset B | N/A |
| *Principal Components Regression (VS)* | All variables | Scaled variables and cross-validation<br>Validation type = "MSEP"<br>Components = 11 |
| *Lasso (VS)* | All variables | Cross-validation<br>Alpha = 1<br>Lambda = 0.0005 |
| *Support Vector Machine (WL)* | Subset A | Kernel = linear<br>Tuned, cost = 0.01 |
| *Super Learner (VS)* | Subset A | Models = Neural Net, GLM, Random Forest, LM, SVM, XGBoost |
| *Super Learner (WL)* | Subset A | Models = Neural Net, GLM, Random Forest, LM, SVM, XGBoost |

*Table 4 - Predictive models. See previous section and appendices for a description of dependent variables. WL = Incumbent win/loss classification model, VS = Incumbent 2-party vote share model.*

This research includes models used for classification (i.e., logistic regression, k-nearest-neighbors), regression (linear regression, support vector machine), and both (random forest, gradient boosting). All of these models are selected from "An Introduction to Statistical Learning," or are otherwise well-documented and widely used in the predictive modeling community. Finally, a "Super Learner" algorithm is employed, which assesses the performance

of several models, then creates a weighting system to produce a final prediction based on a combination of those models. Like the random forest, this is an "ensemble" learning method, which uses multiple learning models determine the final predictions. The super learner has two main strengths: 1) it increases the accuracy of the predictions, and 2) its weighting system shows which models are preferred, and to what extent they contribute to the final predictions.

The variables included were chosen based on the capabilities, strengths, and weaknesses of each model. Subsets A and B, which include only the variables that were deemed relevant based on the existing literature, are small portions of the overall data set. These subsets are preferred in models that are sensitive to multicollinearity (such as linear regression), overfitting (naïve Bayes, trees), or lead to excessive computation times in high dimensions (support vector machine). On the other hand, the principal components regression and lasso are designed for dimensionality-reduction, and the random forest performs very well with high-dimensionality. These models are trained on the full data set.

Many of these models have tuning parameters or require model specifications, and these choices were made with the following goals in mind (ordered by priority): 1) maximize test-set accuracy, 2) minimize model complexity, and 3) minimize model run time. Additionally, cross-validation was used to inform the choices for tuning parameters whenever possible. These methods are mostly consistent with standard practices in the analytic community.

**Section 4.2 Results**

The predictive models varied in both classification accuracy and mean absolute error (MAE) against the testing dataset (Table 5).

| *Model* | *Classification accuracy* | *Mean Absolute Error (MAE)* |
|---|---|---|
| *k-Nearest Neighbors (WL)* | .981 | - |
| *Gradient Boosting (WL)* | .977 | - |
| *Support Vector Machine (WL)* | .977 | - |
| *Random Forest (WL)* | .975 | - |
| *Tree (WL)* | .975 | - |
| *Super Learner (WL)* | .975 | - |
| *Adaptive Boosting (WL)* | .975 | - |
| *Logistic Regression (WL)* | .974 | - |
| *Naïve Bayes (WL)* | .952 | - |
| *Super Learner (VS)* | .978 | .0305 |
| *Random Forest (VS)* | .976 | .0305 |
| *Gradient Boosting (VS)* | .977 | .0314 |
| *Linear Regression (VS)* | .977 | .0347 |
| *Lasso (VS)* | .974 | .0348 |
| *Tree (VS)* | .976 | .0371 |
| *Tree (VS), cross-validated* | .976 | .0407 |
| *Principal Components Regression (VS)* | .960 | .0417 |

*Table 5 - Test set classification accuracy and mean absolute error of predictive models; classification accuracy is the sum of true positives and true negatives divided by the number of test set observations*

Most of these models have similar overall performance on the test set. In this section, the results from three models above are described in detail – one classification model, one regression model, and one super learner.

### Section 4.2.1 k-Nearest-Neighbors

Among classifiers, for which classification accuracy is the preferred performance metric, the k-nearest-neighbor (kNN) algorithm performs best, correctly classifying 98.1% of outcomes with an equal split among false positives and false negatives (Table 6).

| | *Actual Loss* | *Actual Win* |
|---|---|---|
| *Predicted Loss* | 24 | 7 |
| *Predicted Win* | 7 | 691 |

*Table 6 - kNN prediction versus outcome table*

Other than its high accuracy (best among all models tested), an additional benefit of the kNN is its simplicity – test set observations are classified based on the outcomes of the seven most similar training set observations. However, the disadvantage of the kNN model is its lack of meaningful insight into variable importance.

### Section 4.2.2 Random Forest

Among regression models, for which both classification accuracy and MAE are important, the random forest performs adequately in classification accuracy and has the lowest MAE. The random forest had a classification accuracy of 97.6%, a good split between false positives and false negatives (Table 7), and tightly grouped, normally distributed errors (Figures Figure 5 and Figure 6).

| | *Actual Loss* | *Actual Win* |
|---|---|---|
| *Predicted Loss* | 20 | 9 |
| *Predicted Win* | 8 | 683 |

*Table 7 - Random forest prediction versus actual table*

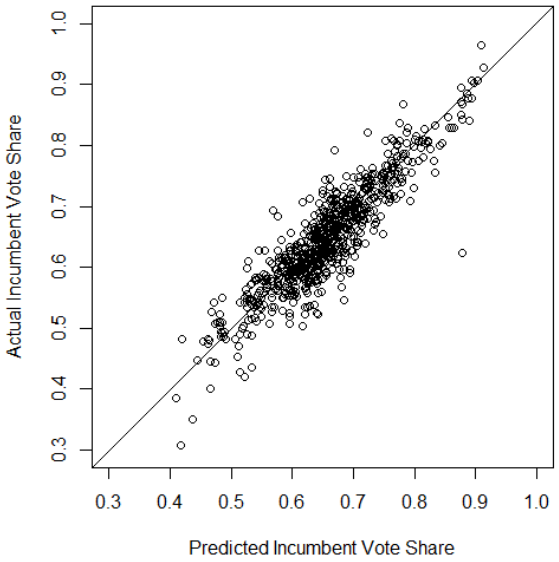



*Figure 5 - Random forest prediction versus actual plot*

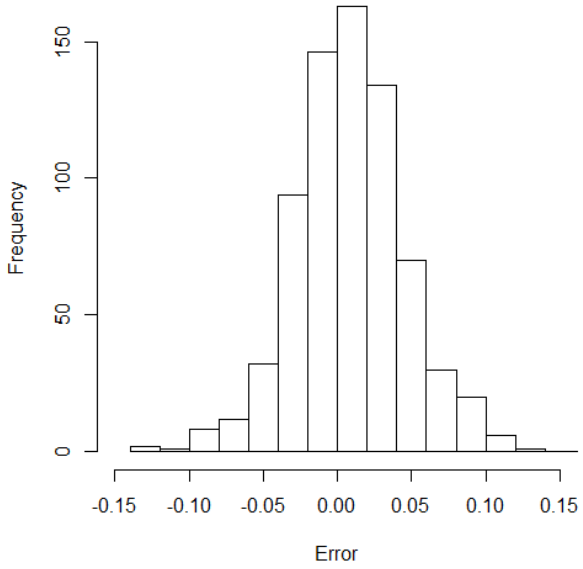



*Figure 6 - Random forest histogram of errors (predicted minus actual)*

Although the random forest model is not as intuitive as the kNN, it does have some distinct advantages. First, it predicts vote share, which allows for a sense of "closeness" in any particular election. Second, it measures the importance of particular variables in the model, including the impact on mean-square-error (MSE) and node purity (Figure 7).

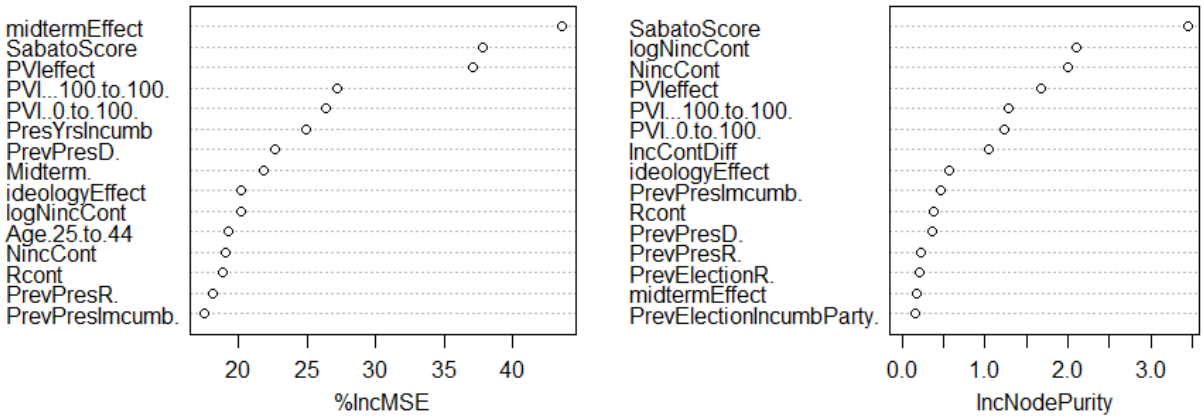



*Figure 7 - Random forest variable importance plot*

Sabato Score, based on the election rating produced by the political expert Larry Sabato, is very influential in the predictions made by the random forest model. An advantage of this is that expert judgment can account for factors not included in the data, such as redistricting.

Furthermore, it appears that the feature creation process of this research was fruitful, as the midterm effect and PVI effect variables are also highly influential to the predictions.

The random forest model is highly accurate in both predicted vote share and election outcome. The predicted vote share and variable importance produced by this model provide useful information about the uncertainty of outcomes and the driving factors behind the predictions.

### Section 4.2.3 Super Learner

Among super learner models, the vote-share model is preferred because outperforms the win-loss super learner in classification accuracy (Table 8) and it includes predictions for vote share with low error (Figures Figure 8 and Figure 9).

| | *Actual Loss* | *Actual Win* |
|---|---|---|
| *Predicted Loss* | 23 | 8 |
| *Predicted Win* | 8 | 690 |

*Table 8 - Super Learner prediction versus actual table*

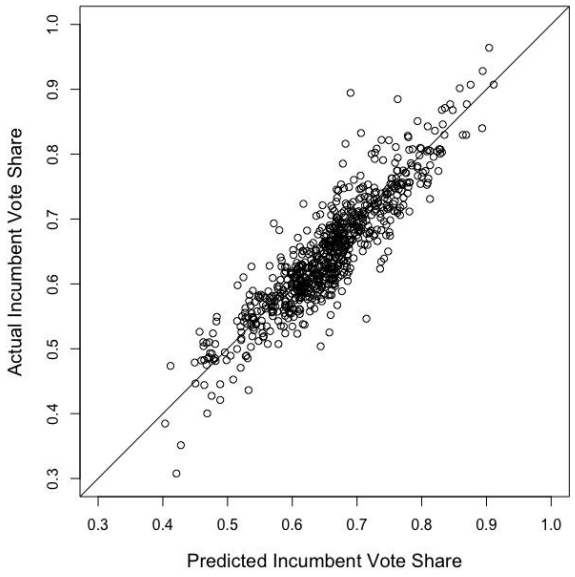



*Figure 8 - Super Learner predicted versus actual plot*

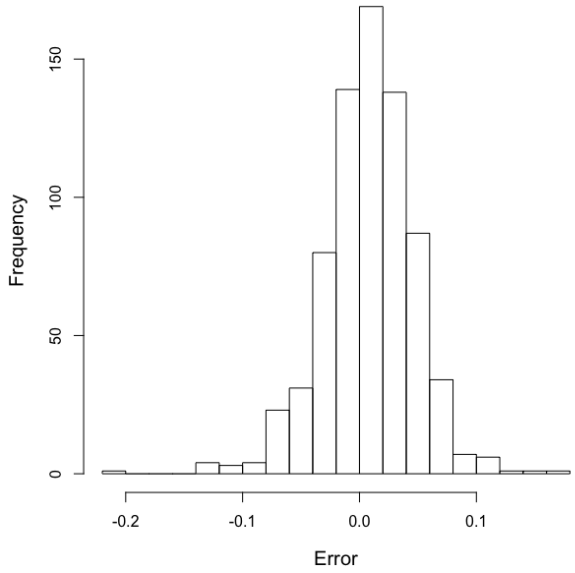



*Figure 9 - Super Learner histogram of errors (predicted minus actual)*

Because multiple models are estimated, and their results are combined to create final predictions, the super learner model does not provide insight into variable importance. However, the weights given to each model show the impact of each model on the final predictions, which can be informative when choosing among several models (Table 9).

| *Super Learner model* | *Weight* |
|---|---|
| *Neural net* | 0.006 |
| *Generalized linear model* | 0.000 |
| *Random forest* | 0.411 |
| *Linear model* | 0.000 |
| *SVM* | 0.311 |
| *XGBoost* | 0.272 |

*Table 9 - Super Learner model weights*

Both linear models received weights of zero in this Super Learner, indicating that they do not contribute to the final predictions. The random forest, support vector machine, and boosting model account for most of the contribution to the final predictions. The Super Learner's outperforms the best individual models in both classification accuracy and absolute error. Furthermore, the model weights are useful information when evaluating multiple models, and

including additional models in the Super Learner cannot reduce its accuracy. This demonstrates why a Super Learner can be a convenient tool when facing a problem with many potential models.

# Chapter 5 Structural Model

This chapter describes a structural econometric model that was built with a focus on quantifying U.S. House election mechanisms and producing relevant insights for stakeholders.

## Section 5.1 Methods

This research moves beyond reduced-form or "black box" models to estimate the causal impacts of candidate campaign expenditures, incumbency, and voter registration on final vote share. Instead, structural estimation depicts the system that produces votes for a particular candidate. Modeling from the perspective of the decision-maker – here, an individual voter – and optimizing their utility function, causal effects can be estimated. Furthermore, causal modeling allows for the generation of counterfactual arguments. For example, it is possible to estimate the level of campaign expenditures by a particular candidate that would have led to a victory instead of the observed defeat.

With this structural approach, an agent-based discrete demand model that depicts an individual utility-maximizing voter facing a discrete choice among candidates is developed. This model is an application of the discrete product-demand estimation in Berry-Levinsohn-Pakes (1995) in which: the consumers are voters, the market is a congressional district, the products are the candidates, the product characteristics are incumbency status and expenditures, the market share is the final vote share, and consumer heterogeneity is captured by the distribution of registered voters within the district. With this method, the causal impacts of campaign expenditures and incumbency on final vote share are estimated.

The derivation of this model is found in Kretschman and Mastronardi (2010). The structural modeling process begins by assuming a utility-maximizing voter choosing among candidates with observable characteristics ($\vec{x}_j$) – incumbency status and campaign expenditures.

$$max_{x_j \in \{x_0, x_1, x_n\}} U_i(x_j) = \vec{x}_j \cdot \vec{\beta}' + \delta_j + \varepsilon_{ij}$$

Where $\vec{\beta}'$ is the vector of coefficients explaining the utility impact of candidate characteristics, $\delta_j$ is the mean utility from unobserved candidate characteristics, and $\varepsilon_{ij}$ is the utility error component to voter $i$ from voting for candidate $j$.

Kretschman and Mastronardi present the conditions that would lead to a voter choosing a particular candidate (or abstaining), and go through a series of steps that lead to the general parameter estimation equation for two candidates:

$$min_{\Theta=(\vec{\beta},\vec{\delta})} \sum_{i=1}^{N} [b_0^{data} - \frac{\exp(\vec{x}_0 \cdot \vec{\beta}' + \delta_0)}{1 + \exp(\vec{x}_0 \cdot \vec{\beta}' + \delta_0) + \exp(\vec{x}_1 \cdot \vec{\beta}' + \delta_1)}]^2$$

Where $b_0^{data}$ is the observed vote share for the Democratic candidate, $\vec{x}_0$ and $\vec{x}_1$ are the vectors of observable characteristics of the Democratic and Republican candidates (respectively), and $\delta_0$ and $\delta_1$ are the respective unobserved candidate characteristics.

It is possible that each of the candidate characteristics may influence heterogeneous voters differently. In Berry-Levinsohn-Pakes (1995), a normally distributed random coefficient was introduced to the regression equation to account for a continuum of heterogeneous consumers. In this analysis, the random coefficients are simpler because only three types of consumers exist – those registered as Democrats, Republicans, or Other – and the distribution of these consumers in each district is known. Using the district voting registration statistics to

account for voter heterogeneity and inserting candidate characteristics into this equation, the utility function for a voter from voting for the Democratic candidate ($x_0$) is:

$$U(x_0) = x_{r,0} \cdot \left(x_{I_0} \cdot \beta_{I_{0,0}} + x_{e_0} \cdot \beta_{e_{0,0}} + \delta_{0,0}\right) + x_{r,n} \cdot \left(x_{I_0} \cdot \beta_{I_{0,n}} + x_{e_0} \cdot \beta_{e_{0,n}} + \delta_{0,n}\right) + x_{r,1} \cdot \left(x_{I_0} \cdot \beta_{I_{0,1}} + x_{e_0} \cdot \beta_{e_{0,1}} + \delta_{0,1}\right) + \varepsilon_0$$

In the first set of terms, $x_{r,0}$ represents the percentage of registered voters in the Democratic party, $x_{I_0}$ and $x_{e_0}$ indicate the incumbency status and expenditure level of the Democratic candidate (respectively), $\beta_{I_{0,0}}$ and $\beta_{e_{0,0}}$ are the utility impacts on Democratic registered voters from the Democratic candidate's incumbency status and expenditures (respectively), and $\delta_{0,0}$ is the mean utility to voters registered Democrat from the unobserved characteristics of the Democratic candidate. The remaining terms can be interpreted in the same manner based on their notation, where $n$ represents voters registered "Other."

Inserting the random coefficient equations into the general estimation equation and assuming that the error term is distributed extreme-value, the final closed-form equation for the predicted final vote share of the Democratic candidate is:

$$b_0(\vec{\beta}, \vec{\delta}) = \frac{\exp\left(x_{r,0} \cdot \left(x_{I_0} \cdot \beta_{I_{0,0}} + x_{e_0} \cdot \beta_{e_{0,0}} + \delta_{0,0}\right) + x_{r,n} \cdot \left(x_{I_0} \cdot \beta_{I_{0,n}} + x_{e_0} \cdot \beta_{e_{0,n}} + \delta_{0,n}\right) + x_{r,1} \cdot \left(x_{I_0} \cdot \beta_{I_{0,1}} + x_{e_0} \cdot \beta_{e_{0,1}} + \delta_{0,1}\right)\right)}{1 + \sum_j \exp\left(\vec{x}_j \cdot \vec{\beta}_j{}' + \vec{\delta}_j\right)}$$

Though this is the same model specification as Kretschman and Mastronardi, this paper uses data from the 2016 election cycle. Many districts do not ask for party affiliation on their voter registration forms, so not all 435 congressional districts are included in the analysis. Non-competitive elections (those without a major party challenger) were also excluded. The resulting dataset contains 232 observations. Using the final predicted vote share equation and the available data, a non-linear regression is specified, and a grid search is performed to find the coefficients

that minimize the sum of squared errors. These coefficients provide the estimation for causal impacts of incumbency and expenditures.

**Section 5.2 Results**

A benefit of the discrete-demand model is that it produces highly specific and interpretable parameters – each parameter estimates the impact of a factor on a specific group of voters. However, this means that a large number of parameters are produced, which can be a problem with so few observations. Restrictions can be imposed on the parameters to mitigate this issue and improve standard errors, and it is possible to relax these restrictions when more data becomes available. Based on the findings of Kretschman and Mastronardi (2010), the following restrictions are necessary and within reason:

$$\delta_{0,0} = \delta_{1,1}, \delta_{0,1} = \delta_{1,0}$$

$$\beta_{e_{0,0}} = \beta_{e_{1,1}} = \beta_{e_{0,1}} = \beta_{e_{1,0}} = \beta_{e_{0,n}} = \beta_{e_{1,n}}$$

$$\beta_{I_{0,0}} = \beta_{I_{1,1}} = \beta_{I_{0,1}} = \beta_{I_{1,0}} = \beta_{I_{0,n}} = \beta_{I_{1,n}}$$

After modeling the non-linear regression with the restricted parameters (for the final vote share of the Democratic candidate), the following results are obtained:

| *Parameter* | *Estimate (Std. error)* |
|---|---|
| $\beta_{I_0}$ *Democrat Incumbent* | 1.630*** (0.136) |
| $\beta_{ln(e_0)}$ *Democrat Expenditures (log)* | 0.046*** (0.007) |
| $\beta_{I_1}$ *Republican Incumbent* | -6.000 (15.705) |
| $\beta_{ln(e_0)}$ *Republican Expenditures (log)* | 0.356*** (0.050) |
| $\delta_0$ *Democrat Unobserved* | -1.003*** (0.107) |
| $\delta_1$ *Republican Unobserved* | -4.230*** (0.748) |

*Table 10 - Summary of results from discrete-demand model*

Positive coefficients for Democratic incumbency and campaign expenditures and the negative coefficient for Republican incumbency are consistent with the conventional wisdom of electoral systems. The positive coefficient for Republican expenditures is somewhat counter-intuitive. Note that a log transformation of candidate expenditures is used under the assumption that the impact on vote share is non-linear with diminishing marginal returns. These results are largely consistent with those found by Kretschman and Mastronardi. The plot of residuals is presented in Appendix 4 and show that the regression has reasonable goodness of fit and does not violate model assumptions.

The interpretation of these results is somewhat ambiguous, but at minimum illustrates the difficulty of congressional campaign spending research. First, the coefficients for candidate expenditures indicate a very small marginal impact, especially in costlier elections. For example, at the mean level of expenditures by the Democrat ($1.14M), an additional $1.95M leads to only a 4-point Democratic swing. Second, spending by the Republican candidate appears to have a stronger impact on the Democrat's final vote share, but the effect is positive. This leads a non-intuitive interpretation of candidate spending impact: it is possible that it does more to increase

voter turnout than influence voter preference. Although these results agree with the findings of Kretschman and Mastronardi, voter turnout is a controversial topic, and the policy implications are not clear. A final interpretation of the discrete-demand model is that even though the impacts of candidate spending are important, they are insignificant compared to the district's partisan demographics. Lawmakers are often concerned about campaign finance laws, but these results suggest that their efforts would be better spent on the districting process. Those that determine district boundaries appear to hold significant power, which can be abused for political gain.

One of the benefits of this causal model is that it can be used to produce counterfactuals – scenarios that describe how the observed results are expected to change based on changes in the factors estimated in the model. Counterfactuals can inform policy or campaign decisions beyond mere prediction of the results. Here, changes in final vote share based on varying levels of candidate expenditures are estimated. This information is not only useful to candidates of individual district elections, but also to decision-makers for political party committees, who are managing their resources across all districts.

A party committee may imagine the set of congressional districts as investments in a portfolio and are interested in how well they maximized the return on their investment. That is, how efficiently campaign funds were used to maximize the number of seats won. Although both parties spent similar amounts overall, inefficiencies in their allocation of funds are likely – tight races with low expenditures and blowouts with high expenditures (Figure 10 - 2016 congressional elections spending and outcomes).

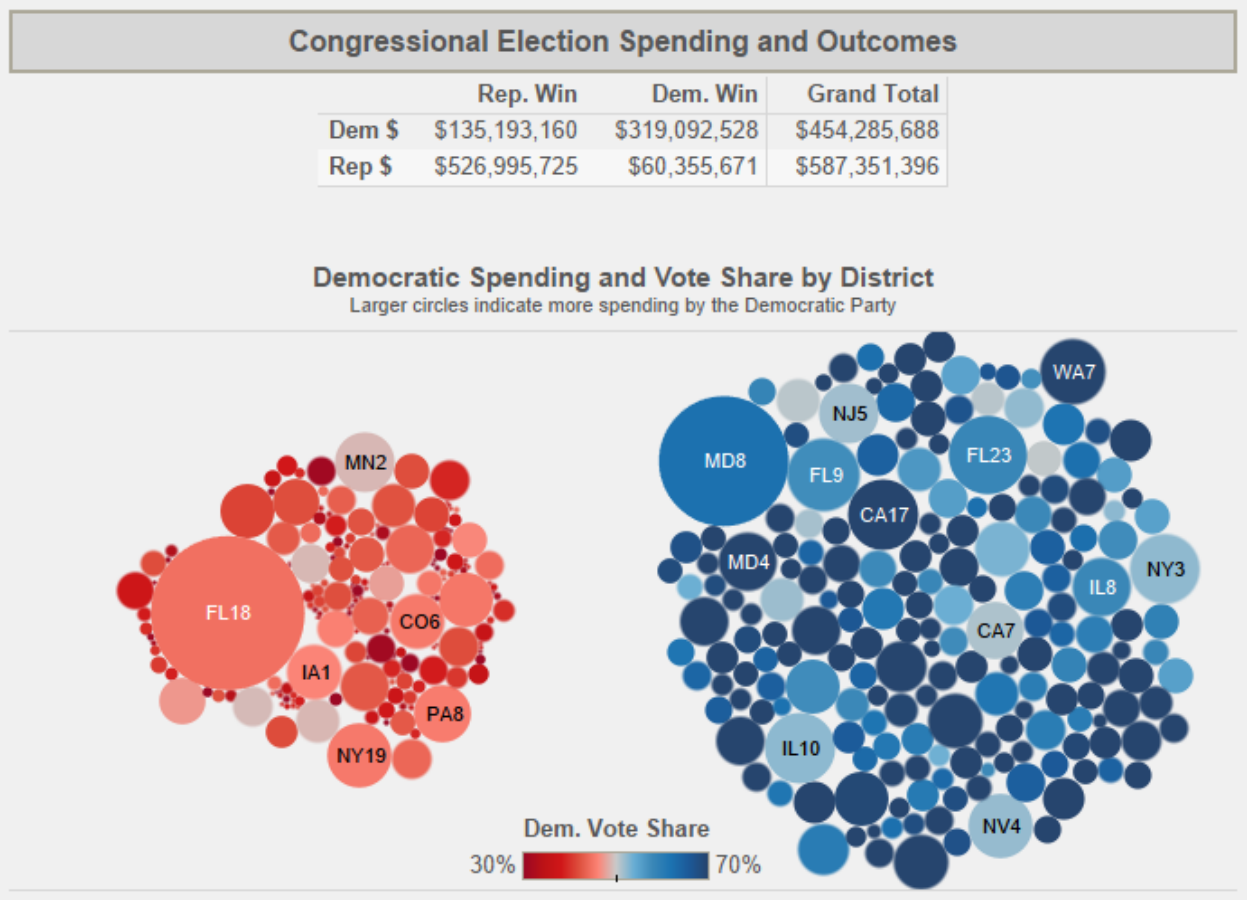



*Figure 10 - 2016 congressional elections spending and outcomes*

On their road to securing 241 seats in the House, Republicans spent $587 million. However, a good portion of this money was wasted on eventual Democratic victories. In fact, to take control of the House, Republicans only needed to win the 218 cheapest elections, totaling about $360 million (Figure 11).

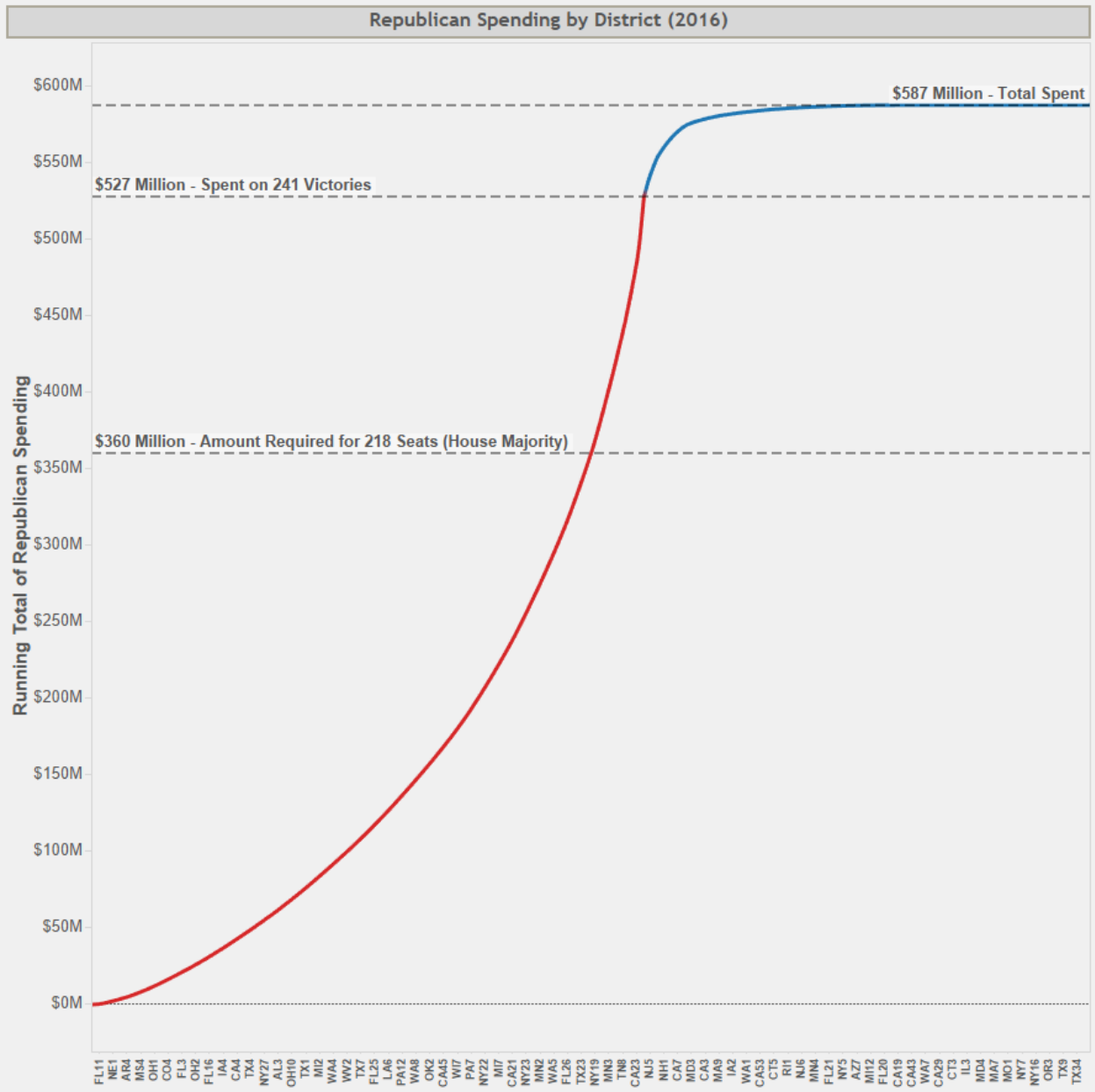
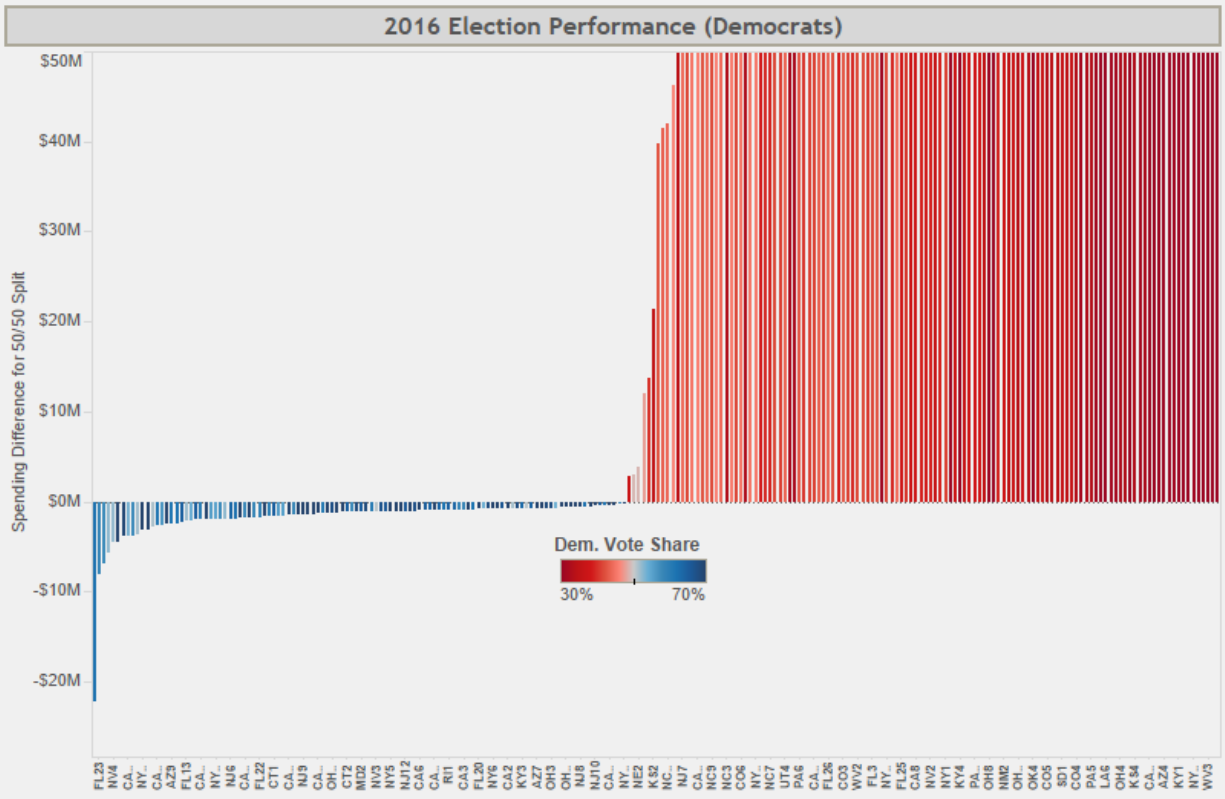



*Figure 11 - Republican spending by district (2016)*

By finding the marginal change in the predicted final vote share based on a unit increase in candidate expenditures, and extrapolating the rate of change, it is possible to estimate the levels of spending that would lead to a 50/50 split in final vote share. This allows for a more detailed view of the elections in which campaign funds were misallocated – the cheapest additional seats and costliest victories (Figures Figure 12 and Figure 13).



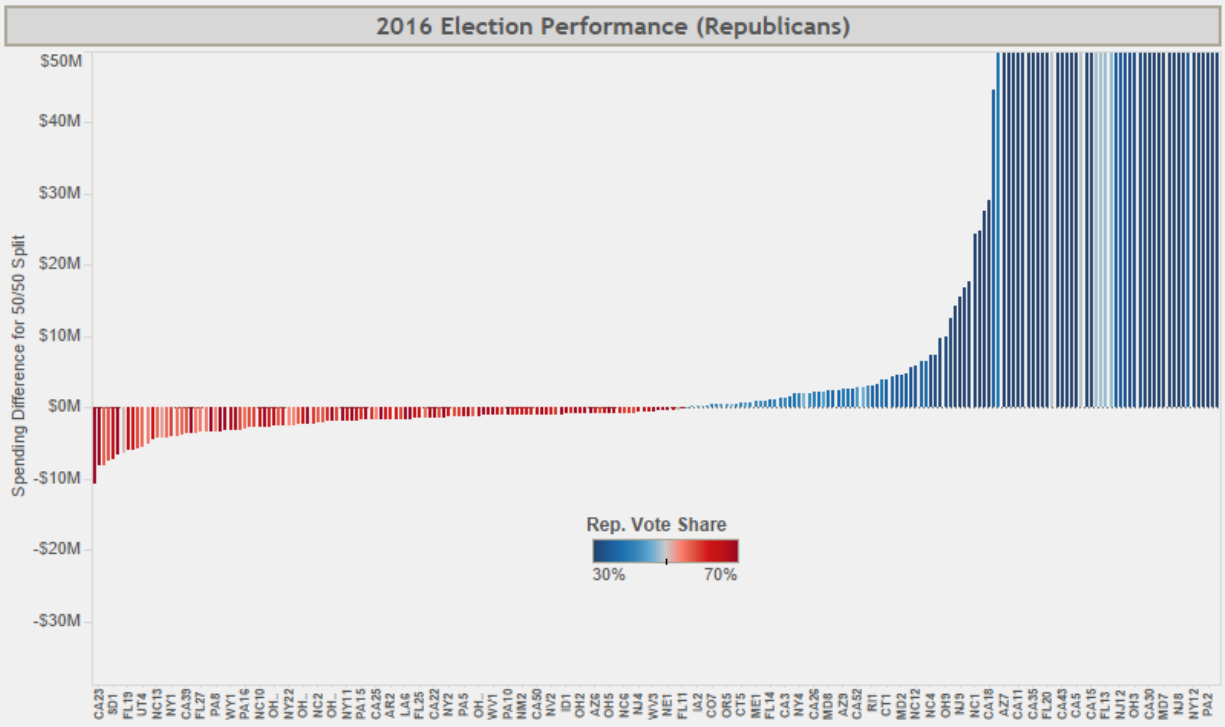

**Costliest Victories**

| | Dem. % | Dem. $ | Delta $ |
|---|---|---|---|
| MD8 | 62.5% | $22.2M | -$22.2M |
| FL23 | 58.3% | $8.1M | -$8.1M |
| FL9 | 57.5% | $7.0M | -$7.0M |
| NY3 | 52.4% | $6.5M | -$5.7M |
| NV4 | 52.1% | $5.4M | -$4.6M |
| MD4 | 77.4% | $4.5M | -$4.5M |
| CA24 | 53.2% | $4.1M | -$3.8M |
| DE1 | 57.5% | $3.8M | -$3.8M |
| MD5 | 69.4% | $3.8M | -$3.8M |
| NJ5 | 51.7% | $4.7M | -$3.7M |
| NY13 | 92.8% | $3.1M | -$3.1M |
| NY14 | 82.6% | $3.1M | -$3.1M |
| CA7 | 51.2% | $4.2M | -$2.7M |
| CA36 | 61.4% | $2.6M | -$2.6M |
| CA52 | 56.4% | $2.6M | -$2.6M |

**Cheapest Additional Seats**

| | Dem. % | Dem. $ | Delta $ |
|---|---|---|---|
| PA9 | 36.6% | $0.0M | $2.8M |
| CA49 | 49.0% | $2.1M | $3.0M |
| NE2 | 48.9% | $2.5M | $3.8M |
| CA10 | 47.6% | $1.6M | $12.1M |
| NJ11 | 40.0% | $0.0M | $13.7M |
| KS2 | 34.7% | $0.0M | $21.4M |
| FL15 | 42.5% | $0.0M | $39.8M |
| NC2 | 43.3% | $0.1M | $41.4M |
| NC13 | 43.9% | $0.2M | $42.0M |
| NY22 | 46.3% | $1.7M | $46.3M |
| PA4 | 33.9% | $0.0M | $51.9M |
| NJ7 | 44.3% | $0.3M | $56.6M |
| NC5 | 41.5% | $0.0M | $78.1M |
| FL27 | 45.1% | $1.0M | $79.0M |
| CA25 | 45.8% | $1.9M | $79.5M |
| CA39 | 42.3% | $0.1M | $97.6M |

*Figure 12 - Democratic congressional election performance (2016), spending differences based on discrete-demand model from Section 5.1*




**Costliest Victories**

| | Rep. % | Rep. $ | Diff $ |
|---|---|---|---|
| LA3 | 82.1% | $10.7M | -$10.7M |
| CA23 | 70.0% | $8.2M | -$8.2M |
| AZ2 | 56.7% | $8.0M | -$8.0M |
| AK1 | 58.1% | $7.4M | -$7.4M |
| SD1 | 64.1% | $7.3M | -$7.3M |
| LA4 | 71.3% | $6.5M | -$6.5M |
| CA49 | 51.0% | $6.3M | -$6.3M |
| FL19 | 65.9% | $6.0M | -$6.0M |
| AZ5 | 63.1% | $5.9M | -$5.9M |
| FL6 | 58.6% | $5.7M | -$5.7M |
| UT4 | 56.2% | $5.6M | -$5.6M |
| NY19 | 54.7% | $5.2M | -$5.2M |
| PA9 | 63.4% | $4.5M | -$4.5M |
| NC13 | 56.1% | $4.3M | -$4.3M |
| CA10 | 52.4% | $4.2M | -$4.2M |
| FL18 | 55.5% | $7.6M | -$4.1M |
| KS3 | 55.8% | $4.1M | -$4.1M |
| NY1 | 59.0% | $4.1M | -$4.1M |

**Cheapest Additional Seats**

| | Rep. % | Rep. $ | Diff $ |
|---|---|---|---|
| NH2 | 47.7% | $0.2M | $0.1M |
| IA2 | 46.3% | $0.2M | $0.2M |
| PA17 | 46.2% | $0.2M | $0.2M |
| CA7 | 48.8% | $1.4M | $0.3M |
| CO7 | 42.0% | $0.1M | $0.4M |
| FL9 | 42.5% | $0.0M | $0.4M |
| NY25 | 44.2% | $0.2M | $0.5M |
| OR4 | 41.8% | $0.1M | $0.5M |
| NY3 | 47.6% | $1.3M | $0.6M |
| OR5 | 44.6% | $0.3M | $0.6M |
| CA9 | 42.4% | $0.1M | $0.6M |
| CT5 | 42.1% | $0.1M | $0.6M |
| FL23 | 41.7% | $0.2M | $0.8M |
| NY18 | 44.8% | $0.5M | $0.8M |
| DE1 | 42.5% | $0.2M | $0.9M |
| FL22 | 41.1% | $0.2M | $0.9M |
| ME1 | 42.1% | $0.2M | $0.9M |
| CT4 | 39.9% | $0.2M | $1.2M |
| FL14 | 38.2% | $0.0M | $1.2M |
| CO2 | 39.7% | $0.2M | $1.3M |
| CA3 | 41.8% | $0.4M | $1.4M |
| CA47 | 37.5% | $0.1M | $1.6M |

*Figure 13 - Republican congressional election performance (2016), spending differences based on the discrete-demand model from Section 5.1*

For the Democrats, there were only a few additional seats that could have been gained through reasonable levels of additional spending. Conversely, Republicans had many opportunities in which a relatively small increase in spending would have led to an expected victory in the district. This result may be an artifact from an election year that favored Republican candidates, but may also indicate an issue with the districting process. Nonetheless, both parties (and individual candidates) had inefficiencies and missed opportunities in their allocation of campaign funds.

These counterfactuals are an example of the type of insights that are only possible through causal modeling. Although predictive models sometimes produce similar coefficients and often have high predictive accuracy, structural models from the perspective of the decision-maker are often more insightful because they estimate causal impacts.

# Chapter 6 Causal Machine Learning Models

This chapter provides an example of the application of machine learning for causal inference to congressional elections.

## Section 6.1 Methods

The model used in this section is based on the work of Athey and Wager (2017), who develop a random forest model for causal inference – a "causal forest." The causal forest is a non-parametric method for estimating heterogeneous treatment effects and is superior to nearest-neighbor methods in both bias and variance. An advantage of the causal forest is that it performs well with a high number of covariates and produces estimates for treatment effects across a continuum. Thus, the causal forest can potentially combine the strengths of the models in chapters 4 and 5.

Athey and Wager develop their method based on the idea of "causal trees." Whereas standard decision trees classify points based on observations in the same leaf, causal trees treat points in the same leaf as random experiments to estimate the treatment effect. From there, many causal trees with random subsamples of training observations can be combined into a causal forest to reduce variance. The primary assumption imposed on the individual trees is that for each training observation, the response may only be used to estimate the treatment effect or to determine where to split the tree, but not both.

Athey and Wager use these findings to create a statistical package for building causal trees and analyzing the resulting treatment effects. To produce a causal forest, the user must define the outcome variable, the treatment variable, covariates, the number of variables to use when assembling each tree, and the number of trees to create. These are similar arguments to

those of a random forest. In this research, the available dataset is smaller than preferred for a causal forest but is nonetheless suitable for this statistical package. Furthermore, the flexibility in the treatment variable argument allows for exploration of the causal effects of many variables (Table 11).

| *Model* | *Covariates* | *Treatment Variable* | *Methods and Tuning Parameters* |
|---|---|---|---|
| *Causal Forest* | All numeric variables | Incumbent expenditures (log)<br>Challenger expenditures (log)<br>Incumbent ideology<br>Change in unemployment rate<br>District PVI effect | Variables per node = 15<br>Number of trees = 1000 |

*Table 11 - Causal forest model*

The resulting models are not only highly accurate (like the random forest model in Section 4.2.2) and provide causal estimates (like the structural model in Section 5.1), but also estimate heterogeneous treatment effects.

**Section 6.2 Results**

Using incumbent expenditures as the treatment variable, the causal forest produces an accurate model of the data with roughly 95% training accuracy for classification (Table 12). A goodness of fit assessment for the treatment effects and model predictions can be found in Appendix 5.

| | *Actual Loss* | *Actual Win* |
|---|---|---|
| *Predicted Loss* | 107 | 54 |
| *Predicted Win* | 24 | 1343 |

*Table 12 - Causal forest prediction versus actual table*

Additionally, the causal forest model estimates heterogeneous treatment effects of incumbent spending (Figure 14). The resulting treatment effect estimates also show that additional candidate spending often has a negligible impact on final vote share, and may even

curtail vote share. However, it is possible that an increase in candidate spending can be the difference between victory and defeat in very close elections, especially when incumbent expenditures were low.

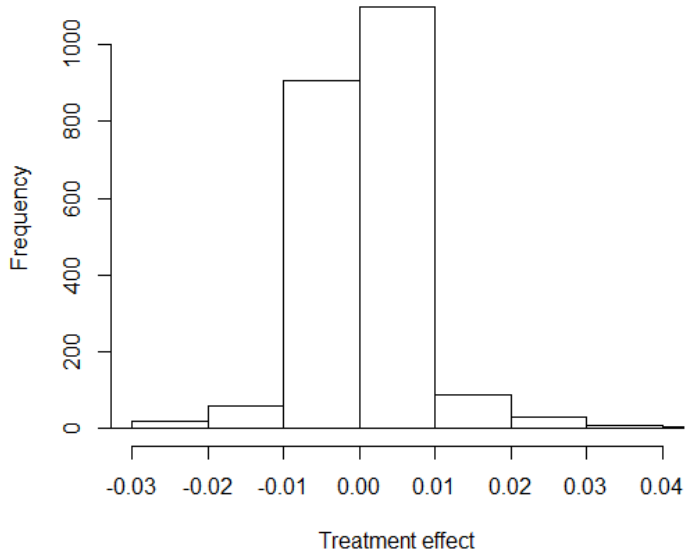
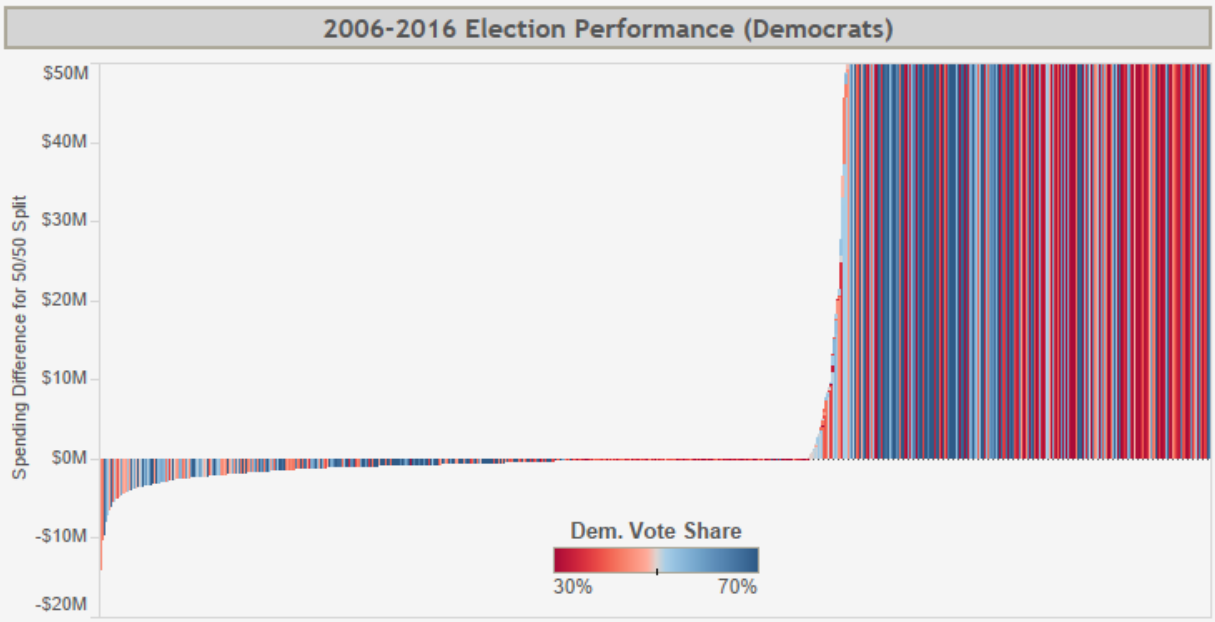



*Figure 14 - Histogram of causal forest treatment effect estimates (incumbent spending)*

The heterogeneous treatment effects can be used to produce counterfactuals, similar to those in Section 5.2. To compare the causal estimates of the causal forest to those from the discrete-demand model, the treatment effects of the challenger's spending must also be estimated. The treatment effects for these two candidates (incumbent and challenger) can then be converted into estimates for Republican and Democratic spending in each election from 2006 to 2016, similar to Section 5.2 (Figures Figure 15 and Figure 16).



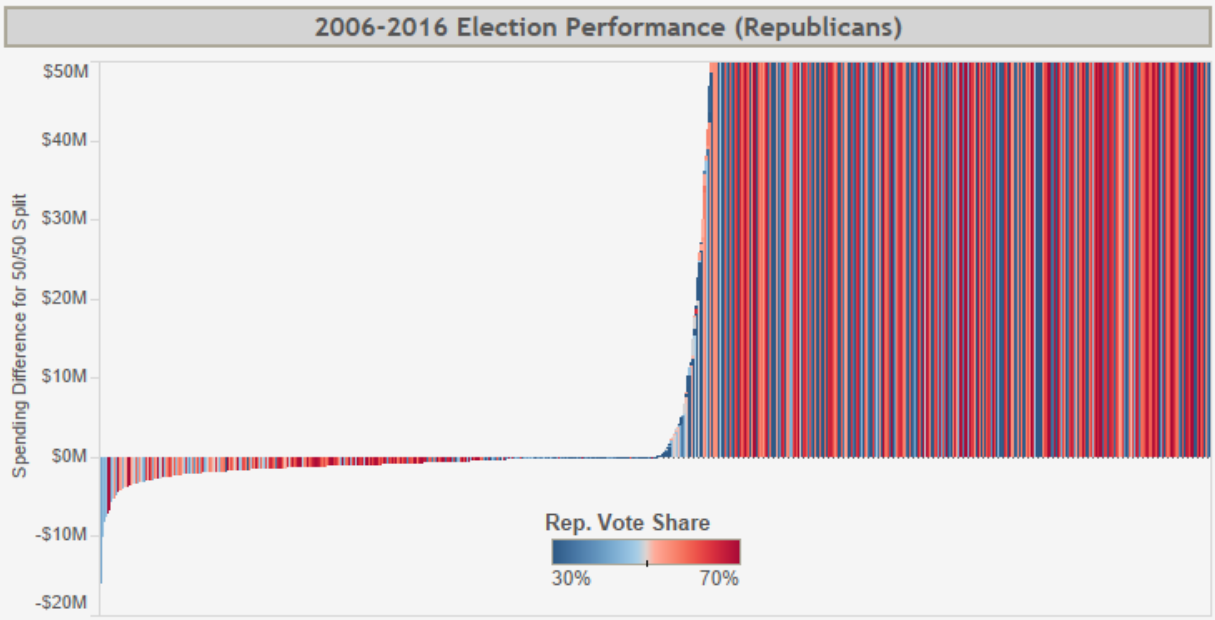

**Costliest Victories**

| | | Dem. % | Dem. $ | Delta $ |
|---|---|---|---|---|
| 2008 | SD1 | 67.6% | $6.1M | -$6.1M |
| | IN7 | 65.1% | $5.0M | -$5.0M |
| 2010 | IL5 | 73.6% | $4.9M | -$4.9M |
| | MD5 | 65.0% | $4.7M | -$4.7M |
| 2012 | NV1 | 66.8% | $12.8M | -$12.8M |
| | NM1 | 59.2% | $10.3M | -$10.3M |
| | HI2 | 80.5% | $9.2M | -$9.2M |
| | NJ9 | 74.7% | $6.6M | -$6.6M |
| | MD6 | 60.8% | $5.8M | -$5.8M |
| | FL9 | 62.5% | $5.4M | -$5.4M |
| | MA4 | 63.0% | $4.9M | -$4.9M |
| | DE1 | 65.8% | $4.5M | -$4.5M |
| 2014 | MI14 | 79.8% | $12.1M | -$12.1M |
| | PA13 | 67.1% | $9.6M | -$9.6M |
| | HI1 | 51.9% | $7.3M | -$7.3M |
| | NJ6 | 60.6% | $5.4M | -$5.4M |
| | MA6 | 57.2% | $5.2M | -$5.2M |
| | DE1 | 61.7% | $6.0M | -$6.0M |
| | NY4 | 57.0% | $4.9M | -$4.9M |
| 2016 | FL23 | 58.3% | $8.1M | -$8.1M |
| | FL9 | 57.5% | $7.0M | -$7.0M |
| | NV4 | 52.1% | $5.4M | -$5.4M |
| | IL8 | 58.1% | $4.6M | -$4.6M |

**Cheapest Additional Seats**

| | | Dem. % | Dem. $ | Delta $ |
|---|---|---|---|---|
| 2008 | NC10 | 42.4% | $0.7M | $2.6M |
| | CA2 | 42.1% | $0.1M | $4.7M |
| | SC2 | 46.2% | $0.7M | $5.9M |
| | CA25 | 42.3% | $0.0M | $7.2M |
| | SC4 | 38.1% | $0.1M | $9.3M |
| 2010 | MN8 | 49.2% | $0.0M | $0.0M |
| | IL8 | 49.9% | $2.5M | $0.5M |
| | NY24 | 50.0% | $1.9M | $0.8M |
| | GA8 | 47.3% | $1.8M | $1.7M |
| | TN1 | 17.5% | $0.0M | $7.9M |
| | NV3 | 49.7% | $2.7M | $8.7M |
| 2012 | MN6 | 49.4% | $0.0M | $0.0M |
| | IL13 | 49.8% | $1.8M | $0.3M |
| | SC3 | 33.3% | $0.0M | $0.8M |
| | IN2 | 49.3% | $6.9M | $3.6M |
| | OH2 | 41.4% | $0.1M | $4.0M |
| | TN2 | 21.7% | $0.0M | $4.2M |
| | GA1 | 37.0% | $0.1M | $8.6M |
| | MO7 | 32.6% | $0.0M | $11.7M |
| 2014 | SC3 | 28.8% | $0.0M | $9.5M |
| 2016 | TX24 | 41.1% | $0.0M | $4.0M |
| | SC1 | 38.6% | $0.0M | $5.4M |
| | NC5 | 41.5% | $0.0M | $6.3M |
| | SC2 | 37.3% | $0.1M | $13.2M |
| | NC6 | 40.7% | $0.1M | $15.2M |

*Figure 15 - Democratic congressional election performance (2006-2016), spending differences based on causal forest model from Section 6.1*




**Costliest Victories**

| | | Rep. % | Rep. $ | Delta $ |
|---|---|---|---|---|
| 2006 | CA50 | 55.0% | $11.6M | -$11.6M |
| | OK5 | 61.8% | $5.1M | -$5.1M |
| | NM1 | 50.2% | $4.9M | -$4.9M |
| | KY4 | 54.4% | $4.3M | -$4.3M |
| | OH4 | 60.0% | $4.2M | -$4.2M |
| 2008 | AK1 | 52.7% | $10.3M | -$10.3M |
| 2010 | IL10 | 51.1% | $18.4M | -$18.4M |
| | AK1 | 69.3% | $6.7M | -$6.7M |
| | KS4 | 61.7% | $4.8M | -$4.8M |
| | KS1 | 76.4% | $4.5M | -$4.5M |
| 2012 | ND1 | 56.8% | $9.6M | -$9.6M |
| | LA3 | 78.2% | $7.1M | -$7.1M |
| 2014 | OK5 | 62.3% | $7.4M | -$7.4M |
| | GA1 | 60.9% | $6.0M | -$6.0M |
| | CO6 | 54.7% | $4.7M | -$4.7M |
| 2016 | NY19 | 54.7% | $5.2M | -$5.2M |
| | CA10 | 52.4% | $4.2M | -$4.2M |
| | TX23 | 50.9% | $4.2M | -$4.2M |

**Cheapest Additional Seats**

| | | Rep. % | Rep. $ | Delta $ |
|---|---|---|---|---|
| 2006 | CT1 | 25.5% | $0.0M | $0.0M |
| | CT3 | 22.8% | $0.0M | $0.0M |
| | IL9 | 25.4% | $0.0M | $0.0M |
| | MS4 | 20.2% | $0.0M | $0.0M |
| | OH9 | 26.4% | $0.0M | $0.0M |
| | CA17 | 22.9% | $0.0M | $0.2M |
| | NY16 | 3.7% | $0.0M | $0.2M |
| | GA12 | 49.7% | $2.2M | $0.4M |
| | IL3 | 22.9% | $0.0M | $0.5M |
| 2008 | CA51 | 25.0% | $0.0M | $0.3M |
| | VA5 | 49.9% | $1.9M | $0.5M |
| 2010 | NY16 | 3.7% | $0.0M | $0.0M |
| | NY7 | 17.2% | $0.0M | $0.2M |
| 2012 | HI2 | 19.5% | $0.0M | $0.0M |
| | NJ10 | 10.8% | $0.0M | $0.7M |
| | NY14 | 14.2% | $0.0M | $0.1M |
| 2014 | CA29 | 25.4% | $0.0M | $0.1M |
| | SC6 | 26.1% | $0.0M | $0.1M |
| | TX16 | 30.2% | $0.0M | $0.1M |
| | MD7 | 27.9% | $0.0M | $0.6M |
| 2016 | OH3 | 31.7% | $0.0M | $0.0M |
| | TX28 | 32.1% | $0.0M | $0.0M |
| | CA30 | 27.6% | $0.0M | $0.2M |
| | TX34 | 37.3% | $0.0M | $0.3M |
| | SC6 | 28.3% | $0.0M | $0.5M |

*Figure 16 - Republican congressional election performance (2006-2016), spending differences based on causal forest model from Section 6.1*

Several interesting findings emerge from these counterfactuals. First, they generally agree with those from the discrete-demand model. Both parties appear to allocate close the optimal amount of campaign funds in many of the elections (especially those that are not expected to be close) but often overspend in elections that are not close or underspend in elections that are within reach. Additionally, many of the same districts are identified as costly victories or cheap additional seats by both the causal forest and the discrete-demand model. For example, both models estimate Democrats overspent in Florida's 23rd district in 2016 by about $8 million.

Second, these counterfactuals differ from those in Section 5.2 in one significant way – the causal impact estimates are sometimes negative, which changes the outcomes in the plots above. For example, Democratic candidates frequently overspent campaign funds in elections that were won by the Republican candidate (Figure 15). In these instances, the causal forest model estimates that campaign spending by the Democrat reduced their final vote share. For example, the discrete-demand model identifies the 2016 election in California's 10th district as a cheap additional seat for Democrats, calling for an additional $12.1 million to secure a victory. However, the causal forest instead identifies this election as a costly victory by the Republican, estimating that a lower spending level would have sufficed. In short, spending under the discrete-demand model always increases vote share, but this is not always the case in the causal forest estimates.

Third, there appear to be party differences in the counterfactuals. Democrats overspend more frequently (though perhaps less severely), and Republicans overall have more inexpensive opportunities to pick up additional House seats. Partisan districting may be a driving factor in these effects, allowing one party to spend campaign funds more efficiently in a particular district and allocate additional resources to more competitive ones.

Overall, the results and counterfactuals derived from the causal forest model showcase the benefits of machine learning for causal inference. The findings in this section are consistent with those from the discrete-demand model, yet do not require a structural model defined by the analyst. Furthermore, the causal forest model produces heterogeneous treatment effects and allows for causal inference of any variable of interest. Treatment effects of four additional variables are shown in Appendix 6.

Simultaneously measuring causal impact estimates of many variables in a system like political elections allows for richer analysis of individual factors and their interactions. Doing so in an environment that is also highly predictive of the outcomes is especially useful to stakeholders, such as candidates and party committees, and demands minimal additional effort from the analyst. These are the defining strengths of causal machine learning, a relatively new field of study. The causal forest described in this section, though not ideally suited for studying political elections, demonstrates these advantages.

# Chapter 7 Conclusions

This paper develops several predictive models of congressional election outcomes, a causal model of candidate expenditures, and reveals new territory for political election models. Scientists have long studied the mechanisms that drive voting behavior and used the findings to construct models that predict or explain election outcomes. The new media landscape provides a platform for easy public access to these models, creating an ideal environment for studying differences in modeling approaches.

Predictive models, such as poll-aggregation forecasts, aim for accuracy, often without regard for explanatory power. Without a reliable source of district-level vote intention polls, many congressional election models predict seat change using national polls or rely on expert knowledge to identify battleground districts. This paper uses machine learning techniques to quantitatively predict congressional election outcomes at the district level. Accurate predictive models often provide crucial information to decision makers by contextualizing current expected outcomes, and should be used to inform strategy and further action. Although they are a fast and inexpensive way to produce accurate predictions, these models do not measure specific factor impacts.

Econometric models are commonly used in literature to study the impacts of particular factors and sometimes develop forecasts ahead of an election. These models have contributed significantly to the collective knowledge of election mechanisms without focusing on maximizing predictive accuracy. However, few of these use structural methods to explain the causal impacts of specific factors, such as candidate spending. This paper develops a structural model of voter choice to explain the causal impacts of expenditures, incumbency, and voter registration statistics. Causal modeling of this type is more prescriptive in nature, and can

quantify expected outcomes of policy decisions or produce counterfactual arguments. However, these models are not always highly predictive and require specialized knowledge to develop.

Most election modelers either prefer one of the above families of election models or lack awareness of other available methods. This paper proposes a new space for election models to explore – machine learning for causal inference. These models use new statistical learning techniques to achieve high accuracy and produce causal impact estimates. An example of one such model, a causal forest, is also provided. The causal forest produces accurate predictions and factor impact estimates that concur with the structural model of voter choice, and can be easily expanded to study other variables. It is the first model of its kind, but many causal machine learning models are suitable for studying elections, and more research is needed in this area. More generally, causal machine learning appears to reconcile the dichotomy between predictive and structural modeling.

Currently, many industries appear to emphasize predictive modeling efforts, largely due to rapid increases in available data. However, as competitive advantages from accurate predictions diminish, causal modeling will likely become more prominent. These approaches will benefit from continued innovation in areas like causal machine learning, and can positively disrupt environments with many stakeholders, such as political elections.

## References

Athey, Susan, and Stefan Wager. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association*, 2017.

Bean, Louis. *How to Predict Elections.* New York: Knopf, 1948.

Berry, Steven, James Levinsohn, and Ariel Pakes. "Automobile price in market equilibrium." *Econometrica*, 1995: 841-890.

Butler, Matthew, David Lee, and Enrico Moretti. "Do Voters Affect or Elect Policies? Evidence From the U.S. House." *The Quarterly Journal of Economics*, 2004.

Campbell, James. "Explaining Presidential Losses in Midterm Congressional Elections." *The Journal of Politics*, 1985: 1140-1157.

Erikson, Robert. "The Puzzle of Midterm Loss." *Journal of Politics*, 1988: 1011-1029.

Gerber, Alan. "Does Campaign Spending Work?" *American Behavioral Scientist*, 2004: 541-574.

Green, Donald, and Jonathan Krasno. "Salvation for the Spendthrift Incumbent: Reestimating the Efects of Campaign Spending in House Elections." *American Journal of Political Science*, 1988: 884-907.

Hastie, Trevor, Gareth James, Robert Tibshirani, and Daniela Witten. *An Introduction to Statistical Learning.* New York: Springer, 2015.

Hummel, Patrick, and David Rothschild. "Fundamental models for forecasting elections at the state level." *Electoral Studies*, 2014: 123-139.

Jacobson, Gary. "The Effects of Campaign Spending in Congressional Elections." *The American Political Science Review*, 1978: 469-491.

Jacobson, Gary. "The Effects of Campaign Spending in House Elections: New Evidence for Old Arguments." *American Journal of Political Science*, 1990: 334-362.

Klarner, Carl. "Forecasting the 2008 U.S. House, Senate and Presidential Elections at the District and State Level." *PS: Political Science and Politics*, 2008: 723-728.

Kramer, Gerald. "Short-Term Fluctuations in U.S. Voting Behavior, 1896-1964." *The American Political Science Review*, 1971: 131-43.

Kretschman, Kyle, and Nick Mastronardi. "U.S. Congressional Vote Empirics: A Discrete Choice Model of Voting." 2010.

Levitt, Steven. "Using Repeat Challengers to Estimate the Effect of Campaign Spending on Election Outcomes in the U.S. House." *The Journal of Political Economy*, 1994: 777-798.

Levitt, Steven, and Catherine Wolfram. "Decomposing the Sources of Incumbency Advantage in the U.S. House." *Legislative Studies Quarterly*, 1997: 45-60.

Lewis-Beck, Michael. "Election Forecasting: Principles and Practice." *The British Journal of Politics and International Relations*, 2005: 145-164.

Lewis-Beck, Michael, and Charles Tien. "Congressional Election Forecasting: Struture-X Models for 2014." *PS: Political Science and Politics*, 2014: 782-785.

Lewis-Beck, Michael, and Charles Tien. "Election Forecasting: The Long View." *Oxford Handbooks Online*, 2016.

Lewis-Beck, Michael, and Mary Stegmaier. "Economic Determinants of Electoral Outcomes." *Annual Review of Political Science*, 2000: 183-219.

Lewis-Beck, Michael, and Tom Rice. "Forecasting U.S. House Elections." *Legislative Studies Quarterly*, 1984: 475-486.

Polley, Eric. "Super Learner Prediction." CRAN, March 12, 2018.

Pyeatt, Nicholas. "Incumbent ideology, district ideology, and candidate entry in U.S. congressional elections, 1954-2008." *The Social Science Journal*, 2014: 181-190.

The Center for Responsive Politics. *Open Secrets*. www.opensecrets.org (accessed May 9, 2018).

Tufte, Edward. "Determinants of the Outcomes of Midterm Congressional Elections." *The American Political Science Review*, 1975: 812-826.

—. *Political Control of the Economy*. Princeton: Princeton University Press, 1978.

Wolpert, David. "The Lack of A Priori Distinctions Between Learning Algorithms." *Neural Computation*, 1996: 1341-1390.

# Appendices

## Appendix 1 Primary data

| *Data Source* | *Years* | *Variables* |
|---|---|---|
| *Constituency-Level Elections Archive (CLEA), compiled by Data4Democracy* | 2006-2014 | Year, State, District, Total votes, Democrat votes, Republican votes, Other votes, Democrat vote share, Republican vote share, Other vote share |
| *Cook Political Report* | 2016 | Year, State, District, Total votes, Democrat votes, Republican votes, Other votes, Democrat vote share, Republican vote share, Other vote share |
| *FEC* | 2006-2014 | Incumbent candidate party |
| *Politico* | 2016 | Incumbent candidate party |
| *Catalist* | 2016 | Democrat registered voters, Republican registered voters, Other registered voters |
| *FEC* | 2006-2016 | Democrat campaign expenditures, Republican campaign expenditures, Other campaign expenditures |
| *Cook Political Report* | 2006-2016 | District Partisan Voter Index (PVI) – ranges from 0 to 100 for a partisan lean (e.g. D+8 or R+13) by comparing a district's results in the previous two Presidential elections to the national average |
| *U.S. Census Bureau* | 2006-2016 | Unemployment rate, Median household income, and Percent of population below poverty line, Aged 25 to 44, Aged 65 and over, White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or Pacific Islander, Some other race, Two or more races, Hispanic |
| *Voteview* | 2006-2016 | Incumbent DW-NOMINATE ($2^{nd}$ dimension) ideology score – a scaling procedure developed by Keith Poole and Howard Rosenthal to analyze voting behavior and assign an ideology score from -1 (liberal) to 1 (conservative) for the current term |
| *Center for Politics – Sabato's Crystal Ball* | 2006-2016 | Larry Sabato's expert prediction, can be one of: D, R, Solid D, Solid R, Likely D, Likely R, Leans D, Leans R, D Toss-up, or R Toss-up |

*Table 13 - Primary data sources*

**Appendix 2 Features**

| *Feature* | *Calculation* | *Values* |
|---|---|---|
| *Incumbent Two-party Vote Share* | Two-party vote share for the candidate whose party matches the "Incumbent" variable above | Numeric, 0 to 1 |
| *Incumbent Win* | Whether the incumbent won | Integer, 0 to 1 |
| *Incumbent Registered Voters* | Share of voters registered to the incumbent's party | Numeric, 0 to 1 |
| *Non-incumbent Registered Voters* | Share of voters not registered to the party | Numeric, 0 to 1 |
| *Incumbent Registration Difference* | Difference between the share of voters registered to the incumbent's party and other parties | Numeric, -1 to 1 |
| *Third Party* | Whether there is a third-party candidate | Integer, 0 to 1 |
| *Ideology Effect* | "PVI (-100 to 100)" multiplied by "Incumbent DW_NOMINATE Ideology" | Numeric, -100 to 100 |
| *Sabato Number* | Re-scaling of the Sabato rating; -5 is most likely Democrat and 5 is most likely Republican | Numeric, -5 to 5 |
| *Sabato Score* | Re-scaled Sabato Number to indicate likelihood that the incumbent candidate will win; -5 is least likely and 5 is most likely | Numeric, -5 to 5 |
| *Open* | Whether it is an open-seat election | Integer, 0 to 1 |
| *Challenged* | Whether the incumbent faces a challenger | Integer, 0 to 1 |
| *Midterm* | Whether it is a midterm election | Integer, 0 to 1 |
| *Midterm Effect* | "Midterm" multiplied by "President/Incumbent Same Party" | Integer, -1 to 1 |
| *PVI (0 to 100)* | Re-scaled PVI; 0 is most Democratic-leaning, 100 is most Republican-leaning, and 50 is neutral | Numeric, 0 to 100 |
| *PVI (-100 to 100)* | Re-scaled PVI; -100 is most Democratic-leaning, 100 is most Republican-leaning, and 0 is neutral | Numeric, -100 to 100 |
| *PVI Effect* | "PVI (-100 to 100)" re-scaled to indicate district's lean toward (positive) or against (negative) incumbent candidate's party | Numeric, -100 to 100 |
| *Incumbent Campaign Expenditures* | Incumbent's campaign expenditures | Numeric, zero or positive |
| *Non-incumbent Campaign Expenditures* | Non-incumbent's campaign expenditures | Numeric, zero or positive |
| *Incumbent Campaign Expenditures Difference* | Difference between incumbent and non-incumbent's campaign expenditures | Numeric |
| *Log Incumbent Campaign Expenditures* | Log transformation of incumbent's campaign expenditures | Numeric, zero or positive |
| *Log non-incumbent* | Log transformation of non-incumbent's | Numeric, zero or |

| | | |
|---|---|---|
| *Campaign Expenditures* | campaign expenditures | positive |
| *President Years Incumbency* | Years that the sitting President has held office | Integer, 0 to 8 |
| *President's Party* | Party of the sitting President | Character, "D" or "R" |
| *President/Incumbent Same Party* | Whether the President and district incumbent are from the same party | Numeric, -1 or 1 |
| *Previous Election Republican Vote Share* | Vote share for the Republican candidate in the previous district election | Numeric, 0 to 1 |
| *Previous Election Democrat Vote Share* | Vote share for the Democratic candidate in the previous district election | Numeric, 0 to 1 |
| *Previous Election Incumbent Vote Share* | Vote share for the incumbent's party in the previous district election | Numeric, 0 to 1 |
| *2nd Previous Election Republican Vote Share* | Vote share for the Republican candidate in the second-most recent district election | Numeric, 0 to 1 |
| *2nd Previous Election Democrat Vote Share* | Vote share for the Democratic candidate in the second-most recent district election | Numeric, 0 to 1 |
| *2nd Previous Election Incumbent Vote Share* | Vote share for the incumbent's party in the second-most recent district election | Numeric, 0 to 1 |
| *Previous Presidential Election Republican Vote Share* | District vote share for the Republican candidate in the previous Presidential election | Numeric, 0 to 1 |
| *Previous Presidential Election Democrat Vote Share* | District vote share for the Democratic candidate in the previous Presidential election | Numeric, 0 to 1 |
| *Previous Presidential Election Incumbent Vote Share* | District vote share for the incumbent's party's candidate in the previous Presidential election | Numeric, 0 to 1 |
| *Previous President Party* | Party of the previous President | Character, "D" or "R" |
| *Previous President/ Incumbent Same Party* | Whether the incumbent and previous President are from the same party | Integer, 0 to 1 |
| *Years of Incumbent Party's District Control* | Years that the incumbent's party has controlled the district's House seat | Numeric, zero or positive |
| *Terms of Democratic Party's District Control* | Years that the Democratic party has controlled the district's House seat | Numeric, zero or positive |
| *Terms of Republican Party's District Control* | Years that the Republican party has controlled the district's House seat | Numeric, zero or positive |
| *Delta Unemployment Rate* | Change in the district's unemployment rate from the previous census | Numeric |
| *Percent Delta Median Household Income* | Percent change in the district's median household income | Numeric |

*Table 14 - Features created*

## Appendix 3 Exploration of select continuous variables

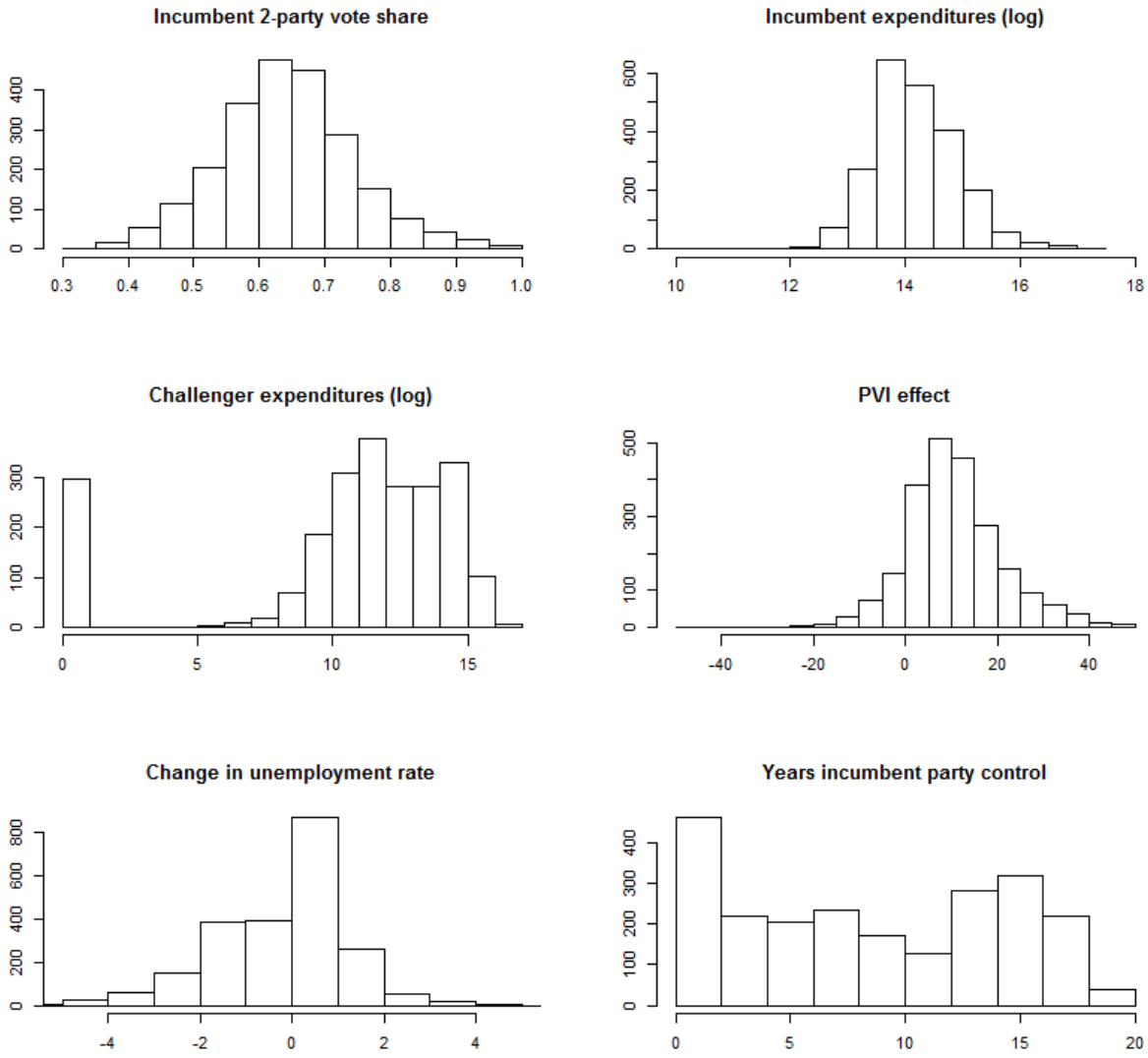



*Figure 17 - Histograms of select continuous variables*

## Appendix 4 Structural model residual plot

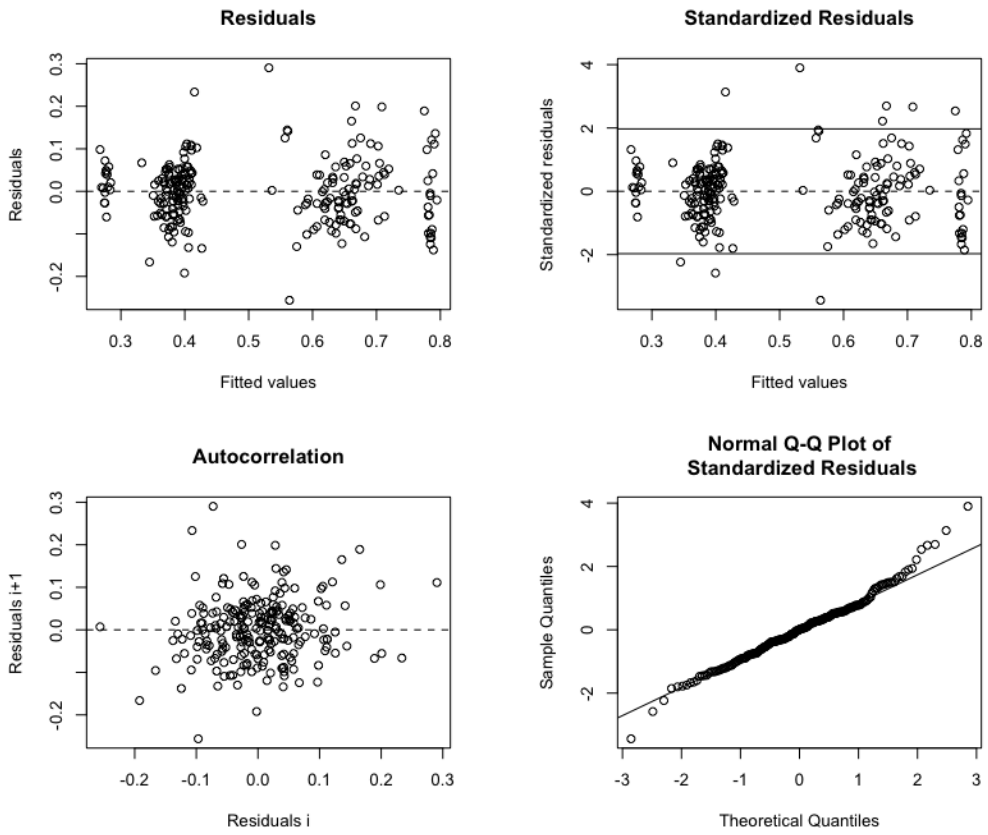



*Figure 18 - Discrete-demand model residuals. There does not appear to be any violation of the heteroscedasticity, autocorrelation, or normality assumptions.*

## Appendix 5 Causal forest goodness of fit

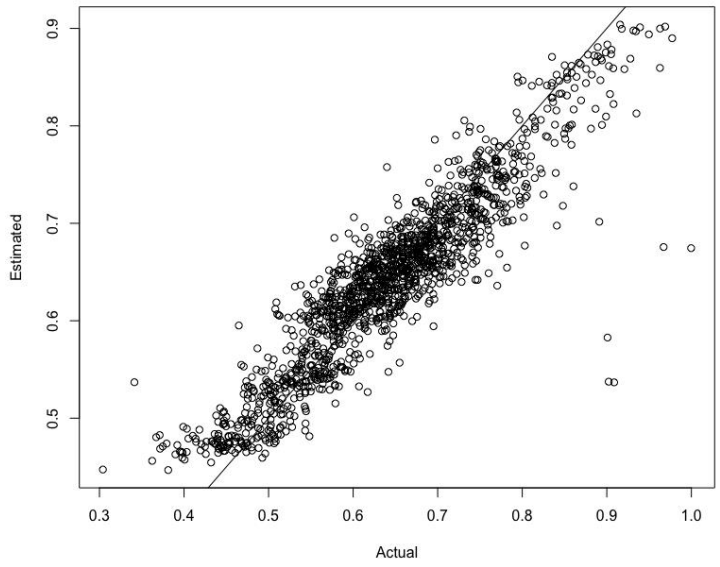



*Figure 19 - Causal forest estimated versus actual treatment (incumbent spending)*

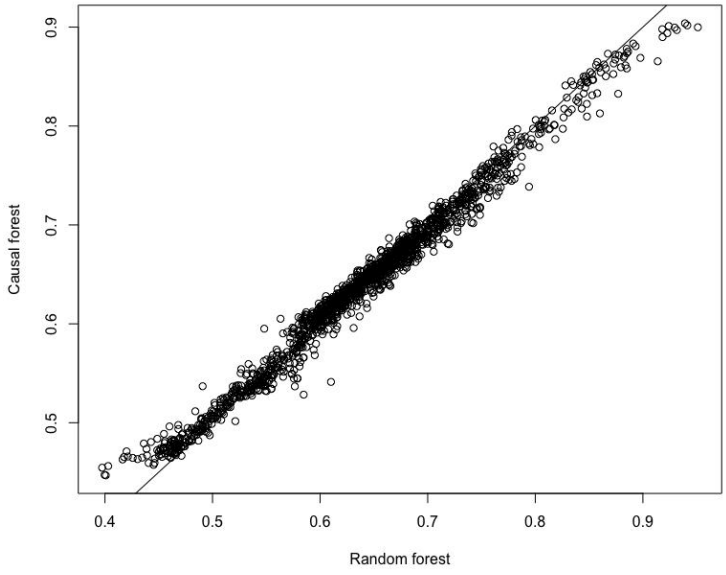



*Figure 20 - Causal forest versus random forest vote share predictions*

## Appendix 6 Additional causal forest treatment effects

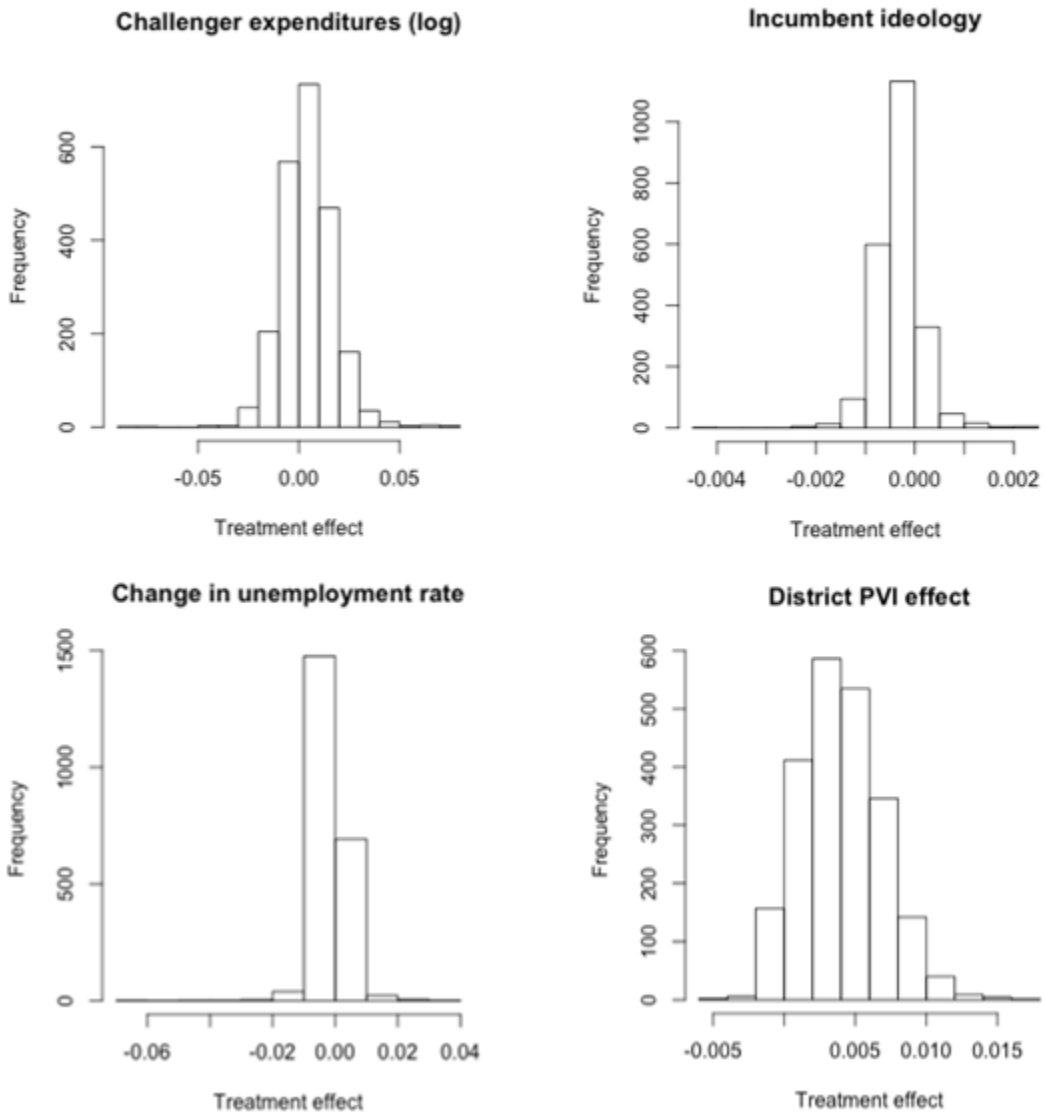



*Figure 21 - Causal forest treatment effect estimates for challenger expenditures, incumbent ideology, change in unemployment rate, and district PVI effect*