Testing Theories Of The Choice To Vote Via Predictive Analytics: A Prospectus

Introduction

The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!', but 'That's funny …'

–Attributed to Isaac Asimov

In its latest incarnation, the general social survey includes over 950 variables, allowing researchers to test their hypotheses using the same set of data for decades. Private sources have even larger collections of political and social data, and larger datasets have become more common throughout political science. But while it is possible to develop hypotheses surrounding some of the data, testing the effects of a few variables on a dependent variable of interest, the majority of the data is never considered. We are, as Yale Librarian Rutherford Roger would put it "drowning in information, and starving for knowledge" (Hastie, Tibshirani, and Friedman 2009). In this proposed dissertation, I suggest a way to leave this state and build our understanding through, rather than in spite of, large data sets, increasing the knowledge that Roger longs for. My approach to adding knowledge through more efficient analysis of data involves applying an algorithm I created to the question of the choice to vote, using inductive pattern searching to turn the starvation Roger describes into a feast.

Within this dissertation, I will explore large data sets using an algorithm I created. This algorithm makes it possible to consider all available data to build models of the most useful predictors for a given dependent variable in a dataset. To deal with a large volume of available data, the algorithm gives each variable in the data set the chance to demonstrate its predictive ability. The algorithm parses the dataset in a manner that both enables the comparison of different theories' predictions and the

creation of a predictive model that will shed new insight on the choice to vote.

To retest theories surrounding the choice to vote, the dissertation will explain and demonstrate the algorithm's means of finding patterns, using inductive reasoning on large datasets and applying the reasoning to data concerning the decision to vote. The benefit of such an approach lies in the use of its predictive analytic technique, which can validate or dispute theorists' offerings independent of more typical approaches in statistical analysis. This technique enables testing theory as it pertains to the criterion of prediction, rather than gathering inferential statistics. Such an approach allows for the creation of models that open a new argument with theorists, due to their basis in a different conception of theoretical power. In the case of making predictions, statistical significance and correlation are irrelevant to theoretical power. Instead, insight is provided based on how well the theorized model predicts a set of data kept out of the sample for modeling. Such an approach builds more predictive models and a sense of which variables have relevance, allowing for a consideration of the choice to vote in a manner that improves on political science's typical approach to data analysis.

In quantitative political science, determining answers to questions such as whether someone decides to vote usually follows a simple format. A researcher hypothesizes a connection between two variables, adding some controls based on theoretical reasoning. The researcher then looks for or gathers data, often conducting a regression to test the hypotheses, and then comes to a conclusion about the theory's veracity based on the regression's results (Brians et al. 2011). Verifying or rejecting theory, however, usually depends on statistical significance, with a p-value below a certain threshold of α determining whether a connection is found (Verzani 2005). Due to the low threshold of α, and the potential for false positives and manipulation, political science's approach can lead to questionable conclusions.

Although our dependence on statistical significance requires a willingness to accept some uncertainty, it has some demonstrable problems in its practice, making conclusions drawn from it

suspect. The problem of relying primarily on significance testing to determine the veracity of a given theory partially lies in the use of the typical threshold for proof. Some level of uncertainty is accepted in all statistical tests, but the use of 95% as an accepted threshold of certainty ensures that one out of twenty tests will return a false result. Either the tests will return a false positive, suggesting that a result is wrongly significant when it shouldn't be, or a false negative, where a result that does predict the dependent variable is instead rejected. While some level of uncertainty should be accepted as part of performing analyses, the level of uncertainty typically used is problematic. Such uncertainty leaves the discipline vulnerable to researchers searching for large quantities of results and publishing on significant ones (Gelman and Lokkan 2013). While the discipline guards against this by demanding theoretical justifications, an alternative statistical approach, predictive modeling, can better prevent overreliance on significance tests.

Predictive Versus Exploratory Modeling

The use of predictive modeling can enable the development of models that do not rely on significance, and has been in some parts of the literature. Before discussing where such an approach has been used before, however, predictive modeling must be defined and contrasted with exploratory modeling, the typical approach in the social sciences. Exploratory modeling involves testing causal explanations using statistical models. A researcher conducting exploratory modeling uses hypotheses to determine what theoretical constructs should be transformed into measurements, which are then compared using statistical testing. Predictive modeling differs in its approach. Instead of developing causal explanations, predictive modeling attempts to apply statistical models for the purpose of predicting new or future observations (Shmuelli 2010). Although predictive modeling is less commonly used in the social sciences, this approach has found sanctuary in electoral politics, where predictive models have allowed for theoretical advancements.

Predictive modeling, while uncommon in political science, is advanced in at least one subdiscipline of political science: forecasting elections. As Mayer points out in his summation of electoral forecasting in American politics, predictive modeling has empowered further understanding of historical factors in deciding elections (Mayer 2014, Abramowitz 2008;2012). Electoral forecasting has also provided a sense of how information affects vote choice and the ability of polls to approximate outcomes (Gelman and King 1993). Debate continues on the role of campaigns in deciding elections (Belanger and Soroka 2012), as well as the best approaches for approximating electoral outcomes (Gelman and King 1993, Silver 2012). Despite these open questions, however, predictive modeling provided the underpinnings that allowed these discoveries to happen, making a contribution to the literature that can be expanded.

Despite predictive modeling's track record in contributing to our understanding of campaigns and elections, and its potential for contributing in other areas, its role in political science remains controversial. Robinson points out that the use of modeling to make prescriptions is dependent on the modelers' assumptions, biasing results in favor of preordained outcomes (1992). Silver dissents from this position, arguing that the modeler's perspective needs to be incorporated, and should be stated explicitly in his advancement of Bayesian statistical approaches to predicting elections (2012). While the debate between Bayesians and Frequentists remains unsettled (Bayarri and Berger 2004), arguing that prior biases affect predictions remains insufficient to discount predictive modeling's potential for contributing to political science.

In addition to the potential effect of prior biases, predictive modeling has been discounted in the literature due to questions of its applicability. For instance, Shapiro points out the difficulty of prediction across political science, as well as the fact that political theory is often broad enough to predict conflicting outcomes. Despite this theoretical capriciousness, Shapiro argues profoundly for the necessity of theory, as without it, empirics are blind (2002). Shapiro is correct, but prediction's

difficulty means it should not be discarded, but embraced, as it allows us to demand increased rigor from theory. When a theory's chosen indicators are capable of predicting out of sample data, that theory gains additional credibility as an explanation, and both theorists and methodologists benefit from such measurement (King 2014). Predictive modeling suggests a set of results that prevent the theoretical capriciousness that Shapiro decries. It does this by requiring a set of logically consistent outcomes from theory, as theory tested by predictive modeling cannot only suggest the importance of a particular predictor, especially when its prediction is compared to analytics' findings. Despite its difficulty, the rewards of predictive modeling outweigh the drawbacks, and the need for such an approach has never been greater.

In addition to the solutions to problems posed by the literature, there has been an increased call for applied social sciences coming from practitioners. The need for additional accuracy not only adds reinforcement to theoretical predictions, but it also enables political science to satisfy the demand for applied theory. Work on political predictions has caught the attention of campaigns (Issenberg 2012), and the increased demand for accurate predictions has driven prior research (Gerber, Green, and Larimer 2008). While predictions can be affected by the researchers' subjectivity and prior beliefs (Robinson 1992, Silver 2012), improving our approaches towards predictive analytics will enable an increase in testable predictions. Added methodological rigor is only one benefit, however, as predictive modeling enables an enhancement in our discipline's relevance, especially in areas where applied social science is required.

Despite the potential for prediction to contribute to questions of applied social science, and to enable further theoretical rigor, the problem of predictive modeling involves the difficulty in applying its techniques. While supervised learning techniques such as gradient boosted modeling might enable the creation of predictive models, these approaches require an inappropriate concentration in statistics for political scientists. To alleviate this problem, I have created a variation on a supervised learning

techniques, as discussed by Hastie, Friedman, and Tibshirani (2009), based on the cross-validation techniques described by Kuhn and Johnson (2013). This approach, known as the Best Subset in Validation Algorithm (BeSiVa, for short), should enable predictive modeling as a means of making predictions and testing theory, looking specifically at the choice to vote.

A Brief Introduction to the BeSiVa Algorithm

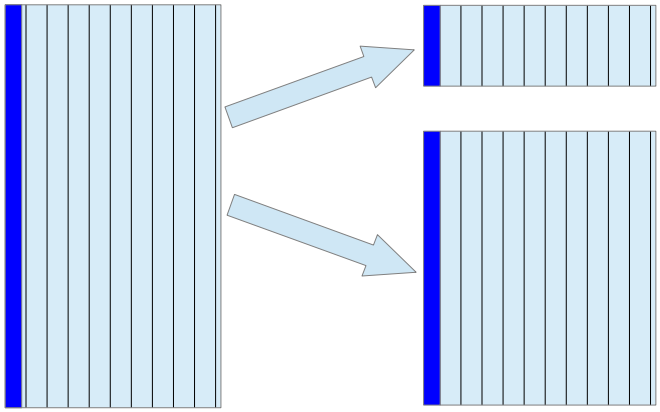
The BeSiVa algorithm was created to solve a specific problem I experienced while attempting to model the choice to support a particular candidate for a campaign. I was in possession of a large data set, without a strong idea of how the majority of the data should be used. There were typical variables, like age, race and income, but also some unusual and esoteric predictors, such as whether someone was likely to smoke heavily or subscribe to hunting magazines. The goal was to find a way to use this information to create a better prediction of someone's support for the candidate. The following sections concern the problem of plenty of data, and little insight, describing the algorithm, and explaining how the algorithm will be incorporated into my dissertation.

How the Algorithm Works

The BeSiVa algorithm originated from the request to test every created model's predictions on a subset of the data, known as a test or validation set (Kuhn and Johnson 2013). As seen in figure 1, the algorithm begins by taking the data and subsetting it into two sets of observations, a test set, and a training set, which would be used in separate parts of the algorithm. The algorithm uses the training data to estimate models, and the test data is then used to see how well those models perform in a later step. This split is necessary, as a failure to divide the data and use the whole set for estimation and

testing may lead to overfitting. When data are overfit, idiosyncrasies in the whole of the data are predicted better than general trends, and making a prediction on the test set prevents this problem (Clark 2004). The observations were split at random between training and test sets at a ratio of 90-10, as per the request of the campaign. Having randomly divided the data into these two subsets, the algorithm can proceed.

Figure 1: A diagram showing the first step of the BeSiVa algorithm. The data are divided into these two subgroups at random, in order to make predictions, and to make sure that the predictions are not based on overfitting the data.



After the data has been split into test and training sets, the algorithm begins estimating models based on the variables in the data. Each model in the campaign data was estimated using logistic regression on the training data, making it easy to determine how well the models made predictions on the test set. In the first step, the dependent variable is regressed against every other variable using the training data, making a set of models. These models are then used to make predictions with the test set of data. Since the regressions employed are logistic, determining model performance consists of taking predictions made on the test set and comparing them to the test set's recorded values for the dependent variable. This measure of model performance is called the percent correctly predicted, or PCP. The model with the highest PCP was noted, and its independent variable was permanently added to future models.

Once the algorithm finds the the best predictor, or the independent variable whose model yielded the highest PCP, it keeps that variable in all models it makes afterwards. The best predictor is removed from the list of variables that can be added to models, and the algorithm repeats the same process, adding all other variables. The algorithm is now looking for the second best predictor, the independent variable whose addition to the model from the last step yields the new highest PCP. Once the two variable model with the highest PCP was found, those two variables were stored. This process could be repeated as many times as desired, adding the number of variables the user prefers, or stopping once the PCP fails to increase above a specified threshold.

To reiterate:

1. Divide the data at random into two subsets: training (to estimate models) and test (to determine how well the models work) sets
2. Estimate one variable models using the training data for each variable in the data set
3. Test each model by predicting the dependent variable outcomes in the test data
    1. Get the percent of dependent variables correctly predicted for each model
4. Record whichever predictor has the highest PCP (the “best” predictor)
    1. remove this predictor from the list of predictors to iterate over
5. Return to step two, but keep the best predictor in all estimated models.
6. Repeat steps 2-5 for as many variables as desired, adding the best predictors to the model
    1. (Optional: State a minimum increase in PCP. If no variable increases the PCP by the minimum, stop the algorithm)

To demonstrate the functionality of the BeSiVa algorithm, I used it on a question concerning the choice to vote in the 2012 presidential election in the general social survey. Based on the algorithm's determination, the most capable predictor of whether someone chose to vote in 2012 was whether they chose to vote in 2008 (vote08), which led to a correct prediction of the choice to vote in

86% of the cases. In its second iteration, the algorithm chose education as a predictor (educ), which brought up the percent of observations correctly predicted to 88%. At this point, multiple variables started predicting the same fraction of observations correctly, suggesting the need for additional data.

For a more thorough demonstration of the algorithm with the same data, I removed whether someone voted in 2008 as a possible predictor. Once that variable was removed from consideration, the strength of an individual's identification with the political parties (called partyid in the GSS data) became the first variable to show relevance, leading to a correct prediction of 76.8% of observations in the holdout set. The next most relevant predictor was another measure of education (degree), leading to 79.7% correctly predicted among the holdout set (the original measure of education was included, but had slightly fewer correct predictions, leading to the prominence of the categorical predictor degree). The next two variables to maximize predictive accuracy were sex and race, which were tied for 80.2% of observations correctly predicted. Not sure what to do if a tie occurred, I added both variables to the next regression. The resulting variable that maximizes predictions (rplace) involves the respondent's relationship with or position as the head of the house. Only one additional correct observation is added by this variable, however, again suggesting the need for further research.

A Plan for the Dissertation

Chapter 1

My plan for this dissertation revolves around two distinct objectives: demonstrating the utility of BeSiVa to make predictions on collections of data, and comparing the algorithm against alternative approaches. These alternatives include theoretically driven models, as well as variations on the algorithm. In chapter one, I will lay out a more advanced description of the algorithm described above,

describing how it can be used to make predictions, and then making them on the GSS in a more systemic fashion. I will apply BeSiVa to the GSS and make comparisons with the alternative approaches using its predictions and inferential statistics. I will compare the algorithm's outputs to other approaches in the literature such as Campbell et al. (1960), Key (1966), and Rosenstone and Hansen (1993). In doing so, I will argue that by the criterion of prediction, some theories in the literature are more capable of predicting the choice to vote than others, demonstrating how predictive modeling and the BeSiVa algorithm add a new means of evaluating theory.

Chapter 2

In its consideration of the data, BeSiVa's proposed models currently tend to select fewer variables than the models proposed by theory. Despite this, a set of questions concerning the algorithm's use of the data remain, and will be explored in the second chapter. Chapter two will feature a discussion of BeSiVa as it compares to variations on the algorithm's design. This chapter will feature the predictions of the algorithm using different targets, datasets used for prediction. These variations will focus on what data that is used, especially looking at what happens when the entire data set is predicted, with no data held out separately for testing. Another variation features multiple holdout sets, dealing with the possibility of an unrepresentative set of held out data by adding a step to BeSiVa, in which several subsets of the data are tested, and the PCPs are averaged. All variations on the algorithm will continue to use the GSS data, making comparisons to the models proposed by the original algorithm and the theoretical approaches in chapter one. Through this comparison of algorithmic variations, chapter two will elaborate on the use of predictive approaches to exploring data, further testing the algorithm and the theoretical explanations of the choice to vote.

Chapter 3

Having explored the algorithm, as well as variations thereof, Chapter three will focus on using the algorithm to compare predictors of the choice to vote across different electoral contexts. Using the GSS's data, the choice to vote in different elections will be tested using the algorithm, noting whether the relevant predictors remain constant or vary. In addition to looking at the algorithm's predictions over time, the chapter will discuss how the algorithm's models compare to theory in terms of the accuracy of the predictions and the independent variables each recommends. This should allow for a comparison of the predictive models to the theoretical approach and offer a suggestion of the theoretical models' veracity, with concurrence indicating further verification of theory. This comparison represents a demonstration of predictive modeling's capability to explain the choice to vote in terms of what variables lead to the best predictions on the data. In doing so, BeSiVa, in its current or modified form (based on the findings of Chapter two), will challenge or verify theory in order to create a new understanding of predictive modeling and the choice to vote.

The BeSiVa algorithm represented an attempt to solve a specific problem for a campaign, involving a question of many variables and little direction. Using this algorithm, my solution, on publicly available data for the first time, I uncovered a set of findings that suggest both verification of and challenges to existing theory. For this reason, I will examine GSS data with my algorithm, using the results as a means of reconsidering theory. I also plan to evaluate the algorithm in terms of differing conceptions of predictive modeling, looking at alternative data to make predictions. In doing so, I intend to challenge theory, using my approach as a new means of evaluating theorists' arguments. This approach challenges theory on a new dimension, using predictive modeling to force explanations to be held up not only to inferential statistics, but to a new criterion: prediction. Such an approach demands additional rigor, requiring explanations that hold up to further scrutiny. For my dissertation I plan to

turn a means of applying this criterion to explore relevant theories and use the models the approach suggests to verify or challenge theories that relate to the choice to vote. This algorithm I created, designed to determine a set of useful predictors from a large collection of possibilities, will allow me to use large data sets as a means of reconsidering the question of the choice to vote.

Appendix: Tables

Table 1: preliminary logistic regressions from the BeSiVa algorithm. The results of whether someone turned out to vote in 2012 are displayed below (including PCP), and the algorithm finds the primary driver of vote choice to be whether someone voted in 2008, with education in years as the second most predictive independent variable.

	Iteration 1 Estimate (S.E.)	Iteration 2 Estimate (S.E.)
(Intercept)	0.908***	0.712***
	(0.008)	(0.036)
Did not vote in 2008	-0.771***	-0.740***
	(0.016)	(0.017)
Ineligible in 2008	-0.473***	-0.455***
	(0.032)	(0.032)
Education (years)	.	0.014***
		(0.002)
N	2125	2125
Deviance	213.100	210.069
PCP	86.5%	88.6%
$-2LLR(Model\chi^2)$ **	236.124**	239.154*

$*p \leq 0.05$** $p \leq 0.01$***$p \leq 0.001$

Table 2: logistic regressions from BeSiVa when vote choice is not an option. The strength and direction of partisanship is the best predictor, with other relevant predictors including educational status race, and relationship to the head of the household.

	Iteration 1 Estimate (S.E.)	Iteration 2 Estimate (S.E.)	Iteration 3 Estimate (S.E.)	Iteration 4 Estimate (S.E.)
(Intercept)	0.897***	0.643***	0.617***	0.632***
	(0.022)	(0.033)	(0.037)	(0.037)
Not Strong Democrat	-0.231***	-0.218***	-0.204***	-0.199***
	(0.033)	(0.031)	(0.032)	(0.032)
Lean Democrat	-0.236***	-0.232***	-0.212***	-0.210***
	(0.034)	(0.033)	(0.034)	(0.034)
Independent	-0.469***	-0.415***	-0.395***	-0.389***
	(0.031)	(0.030)	(0.031)	(0.031)
Lean Republican	-0.204***	-0.198***	-0.176***	-0.173***
	(0.037)	(0.036)	(0.037)	(0.037)
Not Strong Republican	-0.151***	-0.163***	-0.146***	-0.145***
	(0.035)	(0.034)	(0.035)	(0.035)
Strong Republican	0.011	0.003	0.019	0.017
	(0.037)	(0.036)	(0.037)	(0.037)
Other Party	-0.243***	-0.258***	-0.235***	-0.234***
	(0.064)	(0.061)	(0.062)	(0.062)
High School	.	0.201***	0.196***	0.201***
		(0.030)	(0.030)	(0.030)
Junior College	.	0.272***	0.261***	0.259***
		(0.042)	(0.042)	(0.042)
Bachelor's Degree	.	0.384***	0.381***	0.376***
		(0.034)	(0.034)	(0.034)
Graduate Degree	.	0.433***	0.432***	0.426***
		(0.038)	(0.038)	(0.038)
Female	.	.	0.028	0.032
			(0.018)	(0.019)
Black	.	.	0.042	0.041
			(0.027)	(0.027)
Other Race	.	.	-0.090**	-0.082*
			(0.033)	(0.033)
Spouse	.	.	.	-0.033
				(0.024)
Child	.	.	.	-0.127**
				(0.039)
Son or Daughter-in-law	.	.	.	-0.091
				(0.413)
Grandchild or Great-grandchild	.	.	.	0.129
				(0.238)
Parent or In-law	.	.	.	-0.190
				(0.240)
Other Relative	.	.	.	-0.090
				(0.065)
Non-relative	.	.	.	-0.160***
				(0.048)
N	2112	2112	2112	2112
Deviance	392.048	360.174	357.796	353.919
PCP	76.8%	79.7%	80.2%	80.6%
$-2LLR(Model\chi^2)$ ****	55.190****	87.064***	89.442**	93.319*

$*p \leq 0.05$** $p \leq 0.01$***$p \leq 0.001$