

# A simulation study of knowledge

Mark Fredrickson

John Ostrowski

December 16, 2010

## 1 Introduction

Comedian Jay Leno has a running feature on his late night television program entitled “Jaywalking.” In these segments, Leno and a camera crew go out into public places and interview passers-by, usually asking factual open-ended response questions, some of them political in nature. Audiences never see the people who give correct answers, but Leno’s producers never have a shortage of people answering the questions incorrectly. The conclusion the viewers draw, in between fits of laughter, is that the American public is poorly informed with regard to even the basic political institutions and most well-known political players.

Survey research corroborates this impression. Even when using instruments designed to elicit the most favorable responses, survey respondents display low levels of political knowledge (Mondak, 2001). Misinformation abounds as well. Often, the very individuals who are most certain of their beliefs are the most poorly informed (Kuklinksi et al., 2000).

But what if we could view a world in which knowledge was universally high? What would we observe? For starters, Leno would have fewer opportunities to score cheap laughs, but there may be deeper implications. Citizens’ political preferences might shift considerably. Citizens might support different policies than they do currently. If the differences between our world and the world of full knowledge are great, there might be a call for government interventions to raise collective levels of knowledge. If differences are small, we would have greater confidence in citizen competence, even with objectively low levels of knowledge.

Two previous studies have provided windows into the world of high knowledge. Using ANES data and a probit model, Bartels (1996) finds that raising knowledge levels would have a consistent

impact on electoral choices; on average there would be a two percent increase in Republican support in presidential elections. Althaus (1998) considers knowledge's role in policy preferences, again using ANES data with a logistic regression. He finds large changes in policy preference, on the order of changes of 10 percentage points or more in some sub-categories of respondents.

To believe these studies we must believe two things. First, that simulations of high knowledge produce results that are consistent with reality in our counterfactual world of a highly informed citizenry. Second, that their models do a good job predicting voter behavior. The first belief is inherently untestable, as all counterfactuals are. However, the second belief is not only testable but also necessary if one is to make the leap of faith that the first assumption requires. Both studies first build a model, and then exogenously set knowledge levels for all observations to their highest level and use the model to predict outcomes (electoral choice and policy positions, respectively). If the model is flawed, the predicted outcomes are flawed.

There are two ways in which we might be convinced of the models accuracy. First, strong theory might indicate a functional form as well as a series of variables to include in the model. Guided by theory, model construction is relatively simple. Even if the results do not predict outcomes in the sample data, theory dictates that the remaining noise is random error and only more precise measurement instruments could improve the accuracy of the model. Alternatively, in the absence of strong theory, a model that closely fits the data would be more convincing than a model that has a great deal of error in predicting observed outcomes in the training sample. This model would be believable not for any *a priori* reason, but because we are unable to find a more accurate model of the data.

In this paper, we present the second method with an aim toward achieving a better view of the world of full knowledge. We do not claim to have superior theory on the process of citizen preference formation. We do, however, claim to have a superior model building process; one that eschews *a priori* specification of functional form and draws on a wide range of variables, combining them in non-linear ways to reach the most accurate model possible. The accuracy of our different models do not reliably outperform the linear models relied on by past scholars, but the fact that each of our models makes different predictions about a simulated reality is an important finding. Looking specifically at

respondents' policy preference with regard to abortion rights – a question first tackled by Althaus (2003) – the increase in support varies across models, despite the fact that all models perform similarly on accuracy tests. This suggests that past findings may depend on the discipline's emphasis on linear regression. We also introduce in this paper a method for determining the importance of each variable in each model. The predictor variables were theoretically determined by the original researchers and since the emphasis of the research was on simulated results, the importance of each variable was ignored. Examining the importance of each variable across models, we can identify those attributes of respondents that factor into policy preference changes when information levels increase.

## 2 Data

Following Althaus (2003), we use the 1992 American National Election Study (ANES) as the primary data set for our study.<sup>1</sup> As a general picture of the data set, Figure 1 shows the distribution of the sample on gender, income<sup>2</sup>, age and race (restricted to Caucasian and African-American respondents as in later analysis). While the 1992 ANES was a nationally representative sample, our primary concern is making in-sample inferences, so it may be better to consider the respondents members of an experimental study pool rather than a typical survey. As such, we will not spend much time describing the demographics of the subjects.

### 2.1 Knowledge Scale

We also directly adopt the knowledge scale of Althaus (2003), an additive index of 18 political knowledge questions on the ANES survey combined with the interviewer's subjective assessment of the respondent's level of political knowledge. The variable ranges in value from 0 to 23. Figure 2 shows the distribution of the knowledge scale in the 1992 sample.

---

<sup>1</sup>In fact, we are indebted to Scott Althaus for his generously supplying us with the coding scheme used in his book.

<sup>2</sup>Income has been transformed from an ordinal measure into a quasi-continuous variable, with each respondent categorized by income percentile

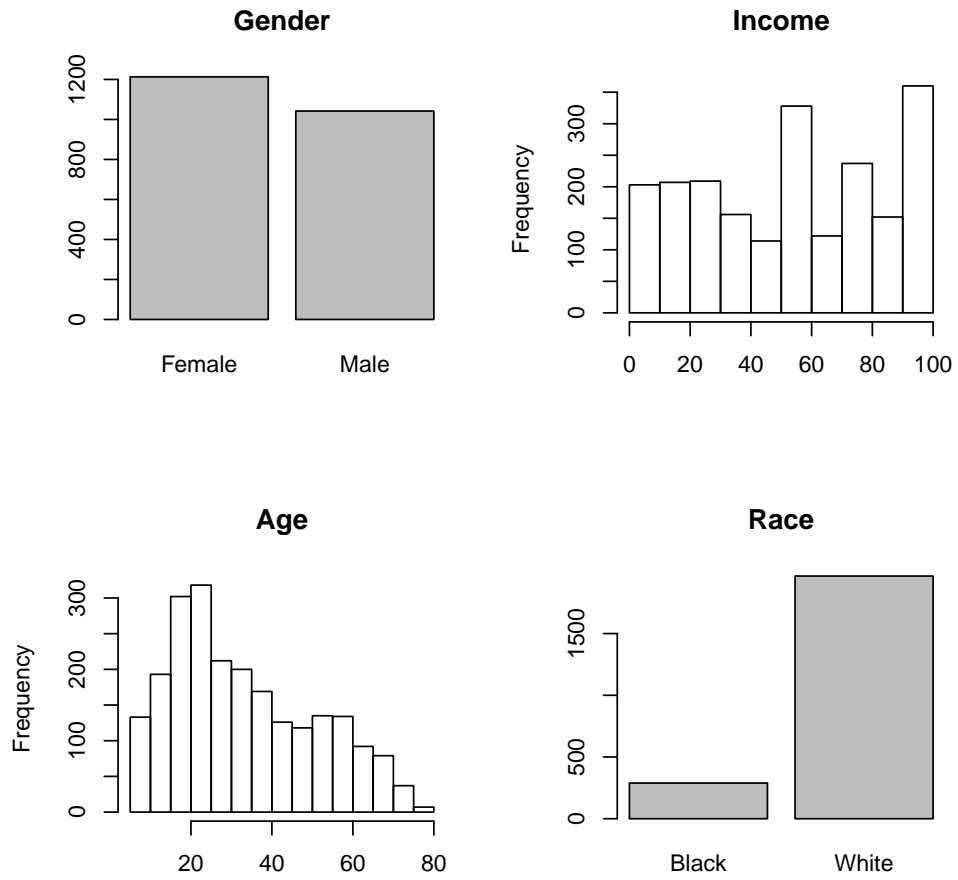


Figure 1: Basic demographic distributions from the 1992 NES

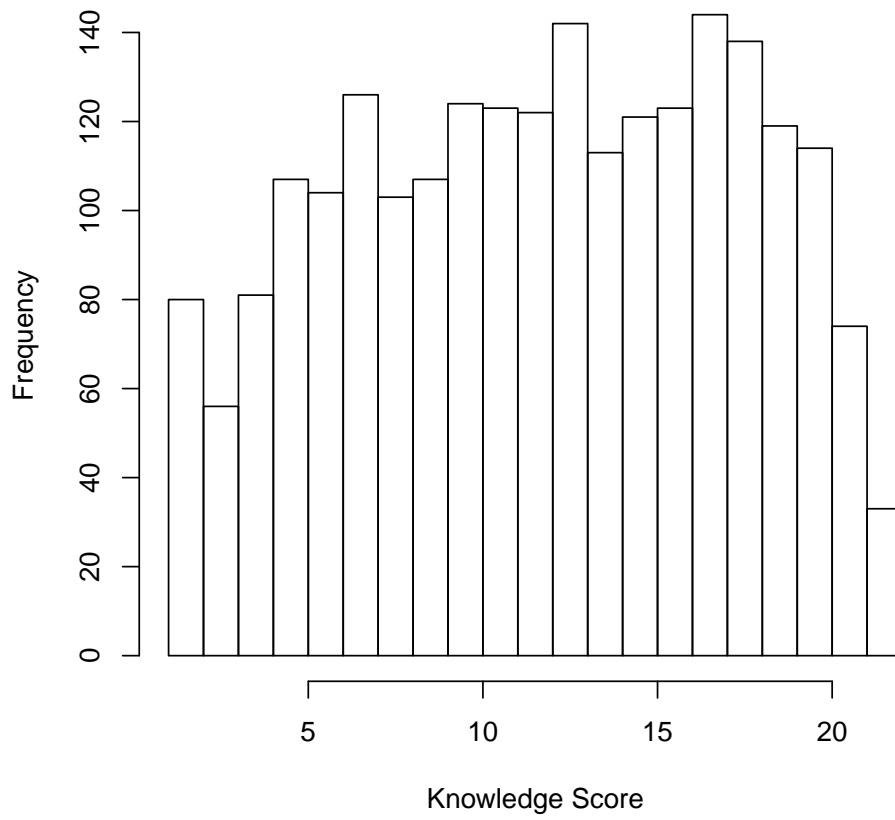


Figure 2: Distribution of knowledge scale on 1992 ANES

## 2.2 Outcome

In order to most directly address previous research, we consider the relationship of political knowledge with support for abortion rights (Althaus, 2003, p. 109 and p. 327). On the ANES survey respondents were given four possible responses (they could also choose to offer their own answer or decline to answer). The first three responses, which all include some sort of limitation of abortion rights (ranging from complete prohibition to an ambiguous limitation on unnecessary abortions). As in Althaus (2003), a binary variable is created by collapsing the first three answers. Thus, respondents expressing a preference for any type of limitation on abortion are coded 0 and those who express support for no limitations at all are coded 1. Table 1 shows the overall support for abortion rights in the entire 1992 ANES sample, as well as support within several sub-categories of the data. Figure 3 provides the full sample support as percentages.

	Total	Support	Oppose
Entire Sample	2228	1030.00	1198
Men	1042	479.00	554
Women	1213	551.00	644

Table 1: Support for abortion rights in 1992 ANES (there is some non-response, so rows do not sum).

## 2.3 Missing Data

Missing data in the sample are non-trivial. We consider the problem of missing outcomes (e.g. support for abortion) differently from missing predictors (e.g. the respondent refused to supply his or her income). In the case of missing outcomes, we exclude those observations from the model building process, but include them in the simulation process, following (Althaus, 2003).

To minimize our reliance on imputation, we use a simple imputation strategy using the mean for numeric variables and the mode for categorical variables. When interacted variables are used in models, we first compute interactions, marking any observations as missing if one or both of the interacted variables are missing (von Hippel, 2009). After computing the interactions, we fill in all missing variables (both original and interacted) with the mean or mode for the column, as

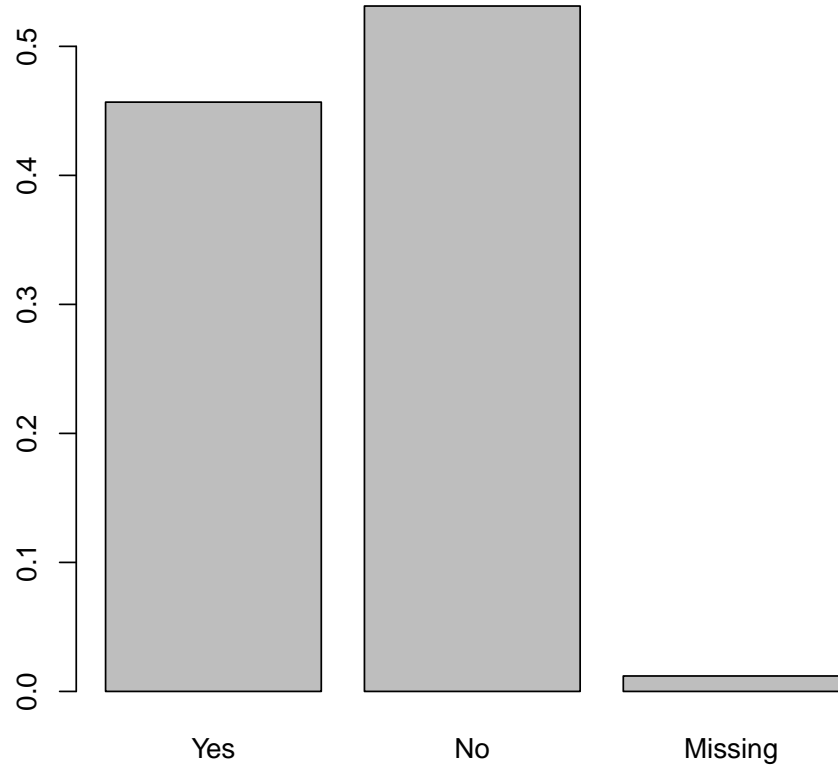


Figure 3: Observed support of abortion rights.

appropriate.

### 3 Models

Many political science studies are well-suited for linear models. Linear models are well-understood, easily interpreted, and are most valuable in testing theories about the data. As a trade off, linear models are less flexible in detecting non-linear relationships, are limited in the number of variables that can be included, and may not produce the most accurate predictions of novel data. For most work, political scientists would be happy to make these trade offs, but for studies for which theory testing is not the primary purpose, as is the case with investigating counterfactual worlds, we might seek a modeling solution that emphasizes prediction accuracy at the cost of interpretation. To put it another way, we would be happy with a black box, provided that black box does a good job of predicting outcomes.

We now consider several possible models, including linear models. For each model, we discuss how we fit the model to the data described in the previous section. We also run a cross-validation process to assess how accurate each model is at predicting novel outcomes.

#### 3.1 Linear Models

##### 3.1.1 Logistic Model

We begin with a replication of the model of abortion rights presented in Althaus (2003): a logistic model with all predictor variables interacted with the knowledge scale. While we do not put much importance on the estimated coefficients or associated standard errors, Table 2 presents the standard regression table that corresponds to Table B.1 in Althaus (2003).

While the standard regression model provides a picture of the relative importance of the different variables, to evaluate many different models using a common frame work we consider what happens if we permute the data while performing cross-validation. From the universe of possible models built from 95% of the data and tested on 5% of the data, we sample 20 units. We repeat this process for all the linear methods and summarize them in Table ??.

	Main Effects	Info Interaction
(Intercept)	-1.597 (0.807)*	
info	0.21 (0.065)**	
educyrs	-0.005 (0.009)	0 (0.001)
zincome	0.015 (0.006)**	0 (0)
age80yrs	-0.005 (0.012)	-0.001 (0.001)
rep	0.021 (0.368)	-0.062 (0.031)*
dem	-0.351 (0.338)	0.024 (0.029)
black	-0.228 (0.403)	0.041 (0.034)
female	-0.66 (0.294)*	0.071 (0.022)**
married	0.112 (0.292)	-0.034 (0.022)
union	-0.154 (0.375)	0.001 (0.028)
homeown	-0.011 (0.305)	-0.009 (0.023)
parent	0.187 (0.38)	-0.007 (0.028)
worseoff	0.44 (0.272)	-0.027 (0.021)
prot	0.033 (0.399)	-0.066 (0.03)*
cath	0.175 (0.483)	-0.036 (0.037)
othrel	-0.395 (0.513)	-0.031 (0.04)
east	0.879 (0.449)	-0.061 (0.032)
midwest	0.289 (0.405)	-0.044 (0.029)
south	-0.033 (0.396)	-0.034 (0.029)
urban	-0.095 (0.308)	-0.002 (0.022)
rural	-0.263 (0.331)	0.012 (0.026)
retired	1.001 (0.522)	-0.042 (0.04)
homemakr	0.774 (0.441)	-0.066 (0.037)
execprof	0.549 (0.451)	-0.026 (0.03)
clerical	1.134 (0.461)*	-0.065 (0.037)
techsale	0.051 (0.532)	0.006 (0.038)

Table 2: Replication of Table B.1 from Althaus (2003)

### 3.1.2 Ridge Regression

We might also wish to shrink the analysis to prevent the model from over fitting the data. On the entire data set, the full logistic regression will always minimize the error, but for testing models built on subsets (i.e., cross-validation), shrinkage techniques might prevent over-fitting and actually provide a greater amount of accuracy in our tests. The first shrinkage method we assess is “ridge regression.” Ridge regression falls into the class of penalized linear regression methods, which combine a least squares term with a penalization term based on model size. Ridge employs a  $L_1$  penalization term. In practice, ridge regression shrinks the magnitude of coefficients, rather than

completely dropping them, as some of the later methods will employ.

### 3.2 Variable Selection via AIC

Before assessing the importance of variables via permutation and cross-validation, it is instructive to look at a simpler method – variable selection using AIC. Of course, we do not expect that using AIC to guide us in selecting our variables will improve the accuracy. Rather, it is simply a way to adhere to Occam’s Razor. It will also be instructive in that it will give us something with which to compare the permutation results. We would expect that the Generalized Importance Scores (GIS) for the full logistic model and variable selection via stepwise AIC would lead to similar conclusions about variable importance. Table 3 shows the results of the selection process.

	x
(Intercept)	-0.922 (0.358)**
info	0.155 (0.025)***
zincome	0.011 (0.002)***
age80yrs	-0.014 (0.004)***
rep	0.228 (0.272)
female	-0.6 (0.285)*
married	0.134 (0.259)
prot	0.115 (0.3)
cath	0.549 (0.38)
othrel	-0.641 (0.193)***
east	0.551 (0.356)
midwest	-0.304 (0.144)*
south	-0.494 (0.141)***
retired	1.04 (0.409)*
homemakr	0.649 (0.423)
clerical	1.05 (0.44)*
info:rep	-0.078 (0.02)***
info:female	0.067 (0.021)**
info:married	-0.042 (0.019)*
info:prot	-0.064 (0.023)**
info:cath	-0.07 (0.027)**
info:east	-0.036 (0.025)
info:retired	-0.048 (0.029)
info:homemakr	-0.062 (0.035)
info:clerical	-0.063 (0.035)

Table 3: Variable Selection with AIC

We can also use the results of the AIC selection as a model unto itself. To score the accuracy of the model, we use the variables selected in the global selection process, but fit a model at each iteration of the accuracy scoring procedure to the training data (see Section 3.7 for more details on the scoring procedure). While the each subset of the dataset might suggest keeping different variables, we use the global results to make the model scoring consistent over iterations.

### 3.2.1 LASSO

The primary tuning parameter of the LASSO regression is the penalization parameter  $\lambda$ . To find an optimal value for this parameter on this dataset, we perform a 20-fold cross validation. We find that setting  $\lambda = 0.00223$  results in the best performance. We use this value in fitting our LASSO regression model used in assessing accuracy, variable importance, and simulation results.

### 3.2.2 Discriminant Analysis

Thus far, we have considered applying regression techniques to classification. Discriminant analysis provides a linear classification scheme in which the classification boundary is selected using a linear method. In certain cases, discriminant analysis may be necessary (i.e., the outcome is not binary). However, in our case, the outcome is indeed binary, and so the choice between discriminant analysis and logistic regression requires a comparison of advantages and disadvantages. Logistic regression is much more easily interpretable and does not rely on the same strict normality assumptions that discriminant analysis relies on. Thus, logistic regression has been selected as our primary classification scheme. However, discriminant analysis is useful (and may provide better estimates) than logistic regression when the ratio of cases to predictor variables is low. Including interaction terms, this ratio is close to 50. We choose to begin with regularized discriminant analysis as it attempts to draw on the strengths of both quadratic (flexibility) and linear (stability) discriminant analysis.

In order to maintain consistency throughout the GIS process, the regularization parameters are selected prior via 20-fold cross validation on the entire dataset. The optimal parameters are  $\gamma = 0.00000272$  and  $\lambda = 0.999$ . Given these regularization parameters, we are essentially using

LDA.

### 3.3 Naïve Bayes

The Naïve Bayes algorithm is well known as a simple, yet effective, algorithm. The end goal of the algorithm is to fit the joint distribution of outcome and predictors using the empirical distribution in the training data. If we believe there is some level of dependence between the variables (and this is easily validated by checking the observed correlations), fitting the joint distribution can be a difficult task (thus techniques that smooth and simplify, such as linear regression). As the name suggests, Naïve Bayes makes a very large assumption: all of the predictors are mutually independent. While this assumption is almost certainly false outside of carefully controlled experimental settings, in practice it performs quite well on real world data.

In our implementation, we consider the assumption of independence to imply that the interaction terms used in the linear models should be excluded. Therefore, we fit the model exclusively on the main effects.<sup>3</sup> The algorithm has no tuning parameters, so there is no need to use cross validation on this model.

### 3.4 CART

“Classification and regression trees”...

One interesting property of this data is a sensitivity to the minimum deviance allowed in the splitting criterion. Figure 4 shows our tree model on the left (minimum deviance set to 0.002) and on the right a version fit with minimum deviance increased to 0.004. In the left hand tree information plays a relatively small role. In fact, information only matters to college educated, non-republicans, which in our dataset comprise 507 cases.

### 3.5 Random Forest

Random forests expanded upon the idea of a decision tree and have seen success in settings where accurate classification of novel data is paramount (Hastie et al., 2008; Siroky, 2009). The goal of a

---

<sup>3</sup>Might be interesting to try with explicit interactions included.



1. The user specifies a total number of trees to build ( $N$ ).
2. For each tree, a bootstrapped sample is drawn from the observations, leaving some “out-of-bag” observations behind.
3. The bootstrapped sample is used to build a single decision tree. At each node in the decision tree, a random sample of variables is drawn. The algorithm uses these variables to maximally separate the observations on the outcome (e.g. finding the best split on the given variables between Democrats and Republicans).
4. The “out-of-bag” observations are classified by the decision tree, and the results are recorded. These records form a running tally of forest accuracy at classifying novel data.

To classify data, observations are fed into the forest one at time. Each tree votes on the observation, recommending a class. The majority winner is returned as the class for that observation.

The random forest has several properties that make it advantageous for this study. First, a measure of accuracy is built directly into the model building process. Our primary goal is developing a model that is more accurate than preceding linear models. Moreover, this measure is based on novel data. Trees are tested on data that was not involved in their construction. Second, random forests can handle many variables and combine them in non-linear ways. Each tree is drawing a sample from the larger population of potential variables. In expectation, for a large enough forest, every variable will be tested against every other variable multiple times and at different levels in the tree. This combination and nesting of variables creates natural non-linear coverage of high-dimensional data.

We use a random forest implementation for the R statistical environment (Liaw and Wiener, 2002).

In building a random forest, the primary parameters are the number of trees in the forest, the number of variables used at each tree node, and the sampling plan for choosing observations to include in building the tree. As a general rule for a categorical outcome such as voter turnout or party choice, the square root of the number of possible variables is used, which in this case translates to 5 variables used per tree node (sampled from the total possible options). For the final

parameter, the number of trees, a few candidate values were employed during the early analysis. There is a decreasing return on additional trees after roughly 250 for this dataset, so a value of  $250 \times 2 = 500$  was selected to ensure sufficient convergence in the forest.

### 3.6 K-Nearest Neighbor

The K-Nearest Neighbor (KNN) algorithm classifies novel data by averaging over the nearest observations from the training data. As the names suggests, this algorithm is tuned setting the size of the neighborhood  $k$  searched. To fit our KNN model, we use 10-fold cross-validation over the NES data to find the  $k$  parameter with the highest mean accuracy. We find that a value of 2 has the highest mean accuracy in a 10-fold cross-validation. This parameter is then used to evaluate the model accuracy and variable importance using the *Generalized Importance Score* algorithm previously described.

### 3.7 Model Summary

We now consider the performance of each model and the relative contribution of each of the included variables to the model’s ability to correctly classify data. While overall accuracy is simple to estimate using cross-validation, it is more difficult to assess variable importance over a wide variety of models. Most readers will be familiar with regression tables showing estimated coefficients, standard errors, and  $p$ -values. While this is a useful method for assessing linear regressions, not every model has an easy to read equivalent (which makes linear regressions’ dominance as an explanatory tool understandable). Comparing across models, moreover, requires a common method of assessing a variable’s importance. To this end, we propose extending a method of variance importance scoring already used in random forest models (Kursa and Rudnicki, 2010; Breiman, 2002).

The intuition behind the technique is simple: if a variable is important to a model’s ability to predict an outcome, if “noise” is inserted into the classifier instead of true values, the classifier should perform worse than when it sees true data. Larger decreases in accuracy when seeing “fuzzed” data should correspond to more importance. We formalize this intuition by relying on sampling theory.

Consider a finite data set  $D$  consisting of an outcome  $Y$  and a set of predictors  $X$ . Fix  $K$  as value much less than the number of observations in  $D$ , say for example 10%. We can now think of a universe of vectors formed by taking all the sized  $K$  subsets of  $D$ ,  $T$ , along with the remaining data  $D - T$ , and classifiers fit on  $D - T$ . If we sample from this space, we can construct a sample of models from all possible models formed in this way.

We then use this sample to infer to the greater population. For each classifier in the sample we can insert noise for each variable and record its performance on the testing data  $T$ . In practice, we use a permutation of the observed values as our source of noise. From the observed decrease in accuracy of the classifier on the fuzzed data we can compute a confidence interval for the true value in the universe of all models fit on subsets of  $D$ . Since the variables we use as inputs to these models are the same, this technique allows us to compare the importance of variables across models. We call this method *Generalized Importance Scores*. While we do not claim to have invented this technique, we are unaware of a formal name.

For each model, we set  $K$ , the amount of data reserved for testing, at 25% and take 20 samples. The amount of data withheld is an arbitrary choice, but variations on the exact percentage have little impact on the results. Twenty samples are chosen in correspondence with the general rule of thumb for invoking the Central Limit Theorem. For each model, Table 4 and Table 5 show 95% confidence intervals for the model accuracy as well as a 95% confidence interval for the mean decrease in accuracy for each variable for linear and non-linear models, respectively. Both confidence intervals represent two-sided alternative hypotheses.

## 4 Simulation

To simulate behavior under full knowledge, we set every observation's information level to the maximum value of the scale: 23. Then we predict outcomes for all observations (even those with missing outcomes in the original sample) using each model. As noted in Althaus (2003), the correct simulation method for aggregate preferences is to use the average probability of support for abortion rights, rather than the average of the individual expected outcomes (i.e. considering individuals with an estimated probability of less than 0.5 to oppose full abortion rights and individuals with

	Logistic	Ridge	LASSO	RDA	AIC
accuracy	(0.637, 0.653)***	(0.463, 0.48)	(0.648, 0.662)***	(0.643, 0.661)***	(0.654, 0.669)***
info	(0.036, 0.048)***	0	(0.033, 0.05)***	(0.038, 0.053)***	(0.044, 0.063)***
educyrs	(-0.004, 0.001)	0	(-0.004, 0.001)	(-0.005, 0.001)	0
zincome	(0.017, 0.029)***	(0, 0)	(0.019, 0.034)***	(0.019, 0.037)***	(0.019, 0.031)***
age80yrs	(0.001, 0.011)*	0	(0.004, 0.015)**	(0.005, 0.017)**	(0.006, 0.02)**
rep	(0.033, 0.045)***	(0, 0)	(0.031, 0.051)***	(0.041, 0.059)***	(0.045, 0.058)***
dem	(-0.001, 0.006)	(0, 0)	(-0.001, 0.004)	(-0.002, 0.006)	0
black	(0, 0.007)	(0, 0)	(0, 0.005)	(0.002, 0.007)**	0
female	(0.002, 0.015)**	0	(-0.003, 0.005)	(0.001, 0.01)*	(0.003, 0.011)**
married	(0.001, 0.01)*	0	(0.002, 0.013)*	(-0.001, 0.008)	(0.006, 0.017)***
union	(-0.006, 0.001)	0	(-0.003, 0)	(-0.002, 0.003)	0
homeown	(-0.004, 0.004)	0	(-0.003, 0.003)	(-0.004, 0.002)	0
parent	(-0.002, 0.003)	0	(0, 0.005)	(-0.002, 0.004)	0
worseoff	(-0.005, 0.002)	0	(-0.004, 0.001)	(-0.002, 0.003)	0
prot	(0.025, 0.041)***	(-0.002, 0)*	(0.02, 0.037)***	(0.016, 0.031)***	(0.021, 0.037)***
cath	(-0.003, 0.005)	0	(-0.001, 0.007)	(0, 0.008)	(-0.003, 0.004)
othrel	(0.013, 0.025)***	(-0.002, 0)*	(0.012, 0.025)***	(0.013, 0.024)***	(0.012, 0.023)***
east	(-0.001, 0.008)	0	(-0.002, 0.002)	(-0.004, 0)	(-0.004, 0.002)
midwest	(-0.002, 0.004)	(0, 0)	(-0.002, 0.006)	(0, 0.009)	(-0.002, 0.006)
south	(0.001, 0.01)*	(-0.001, 0)	(0.003, 0.016)**	(0.005, 0.016)***	(0.007, 0.019)***
urban	(-0.005, 0.001)	0	(-0.003, 0.002)	(-0.005, 0)*	0
rural	(-0.003, 0.004)	0	(-0.001, 0.003)	(0, 0.005)*	0
retired	(-0.001, 0.008)	0	(-0.004, 0.004)	(-0.003, 0.004)	(-0.003, 0.006)
homemakr	(-0.002, 0.004)	0	(-0.002, 0.002)	(-0.003, 0.001)	(-0.001, 0.005)
execprof	(0, 0.006)	0	(0, 0.006)	(0, 0.003)	0
clerical	(-0.001, 0.006)	0	(-0.001, 0.005)	(0.001, 0.006)**	(0.001, 0.006)*
techsale	(-0.002, 0.003)	0	(-0.002, 0.003)	(-0.002, 0.003)	0

Table 4: *Linear Models*: 95% confidence intervals for model accuracy and mean decrease in accuracy for each variable. Significance stars are included for the 0.05, 0.01, and 0.001 levels. The null hypothesis for accuracy is the same as guessing the mean number of abortion supporters (.464). The null hypothesis for the variables is that they have zero decrease in accuracy.

	CART	Random Forest	K-Nearest Avg	Naive Bayes
accuracy	(0.589, 0.61)***	(0.627, 0.649)***	(0.489, 0.506)***	(0.604, 0.625)***
info	(0.016, 0.036)***	(0.021, 0.038)***	(-0.012, 0.012)	(0, 0.014)
educyrs	(0.01, 0.03)***	(0.014, 0.028)***	(0, 0.018)	(-0.002, 0.005)
zincome	(0.009, 0.027)***	(0.01, 0.027)***	(-0.016, 0.011)	(-0.001, 0.01)
age80yrs	(-0.008, 0.011)	(-0.005, 0.008)	(-0.001, 0.028)	(-0.005, 0.007)
rep	(0.023, 0.044)***	(0.016, 0.028)***	(-0.001, 0.006)	(0.002, 0.011)**
dem	(-0.004, 0.002)	(0.001, 0.012)*	(-0.003, 0.003)	(-0.004, 0.003)
black	(-0.001, 0.001)	(0, 0.004)*	(-0.003, 0)	(-0.004, -0.001)**
female	(-0.003, 0.004)	(-0.006, 0.002)	(0.001, 0.009)*	(-0.002, 0.001)
married	(0.001, 0.008)*	(-0.003, 0.007)	(-0.003, 0.001)	(-0.004, 0.002)
union	(-0.002, 0.001)	(-0.004, 0.001)	(-0.003, 0.001)	(-0.005, 0)
homeown	(-0.003, 0)	(-0.002, 0.006)	(-0.004, 0.002)	(-0.006, 0.001)
parent	(-0.002, 0.001)	(-0.006, 0)	(-0.002, 0.002)	(-0.006, 0.004)
worseoff	(-0.005, 0)	(-0.002, 0.005)	(-0.002, 0.003)	(-0.002, 0.002)
prot	(0.001, 0.012)*	(-0.002, 0.009)	(-0.002, 0.003)	(-0.005, 0.004)
cath	(-0.003, 0.003)	(-0.002, 0.005)	(0, 0.005)	(-0.003, 0.004)
othrel	(-0.001, 0.006)	(-0.004, 0.005)	(-0.002, 0.003)	(-0.004, 0.003)
east	(-0.002, 0.003)	(-0.004, 0.004)	(0, 0.006)	(0.003, 0.012)**
midwest	(-0.005, 0.001)	(-0.004, 0.003)	(-0.001, 0.007)	(0, 0.004)
south	(-0.002, 0.009)	(-0.003, 0.008)	(-0.002, 0.005)	(-0.002, 0.01)
urban	(-0.002, 0.002)	(-0.002, 0.004)	(-0.003, 0.004)	(-0.005, 0.001)
rural	(-0.002, 0.002)	(-0.005, 0.002)	(-0.003, 0.003)	(-0.005, 0.008)
retired	(-0.002, 0.002)	(-0.003, 0.002)	(-0.001, 0.002)	(-0.003, 0.002)
homenakr	(-0.003, 0)	(-0.004, 0.002)	(-0.003, 0)*	(-0.006, 0.004)
execprof	(-0.001, 0.001)	(-0.003, 0.004)	(-0.001, 0.005)	(-0.004, 0.007)
clerical	(-0.002, 0.001)	(-0.003, 0.002)	(-0.001, 0.002)	(0.002, 0.011)*
techsale	(-0.001, 0.001)	(-0.003, 0.003)	(-0.002, 0.001)	(-0.005, 0.005)

Table 5: *Non-linear Models*: 95% confidence intervals for model accuracy and mean decrease in accuracy for each variable. Significance stars are included for the 0.05, 0.01, and 0.001 levels. The null hypothesis for accuracy is the same as guessing the mean number of abortion supporters (.464). The null hypothesis for the variables is that they have zero decrease in accuracy.

probabilities greater than 0.5 to all be supporters). Consider a group of respondents, all with an estimated 0.6 probability of supporting abortion rights in the simulation. If we sum the expectation of each individual, we would conclude that 100% of the group supports abortion rights, which would be overstating the level of support. If responses are truly random with a  $p = 0.6$ , we would actually expect 60% of the sample to support full abortion rights, not 100%.

In this simulation, we consider only models for which the confidence interval included at least 60% accuracy. This is an arbitrary cut-off, but it does allow us to focus on fewer models, each of which is better than guessing that the respondent does not support abortion rights (which would net an accuracy rate of 53%). This cut off excludes ridge regression and k-nearest neighbor, but includes logistic regression, LASSO, Naive Bayes, CART, and Random Forest.

Figure 5 shows the results of the simulation for the models listed in Section ?? for the full sample. Also shown is the observed support for full abortion rights. We repeat this procedure breaking apart the sample into men and women. Figure 6 presents these results.

All of the models show increased support for abortion rights by both men and women. Are these increases “significant” in a statistical sense? To answer this question we treat our sample as an experimental pool that is undergoing a manipulation to increase political knowledge. In this framework, the predicted support of abortion for each observation using the model can be considered a pre-test measuring the subject’s support of abortion on a  $[0, 1]$  scale prior to the manipulation. After the manipulation (setting every observation’s information score to 23), we use the model to measure the probability of supporting abortion rights. All of the methods show significant increases in probability, but the substantive changes vary greatly.

The different methods disagree with respect to substantively large changes in the predicted probability of support, but do they paint the same picture with respect to the initial level of information? That is, who benefits the most from an increase in information: those at the bottom or middle of the scale (as naturally high information individuals are unlikely to change in this “experiment”). Figure 7 and Figure 8 show the mean change in predicted support for across the observed political knowledge scale for linear and non-linear methods, respectively. While all the methods show larger gains for those at the low end of the observed political knowledge scale, linear

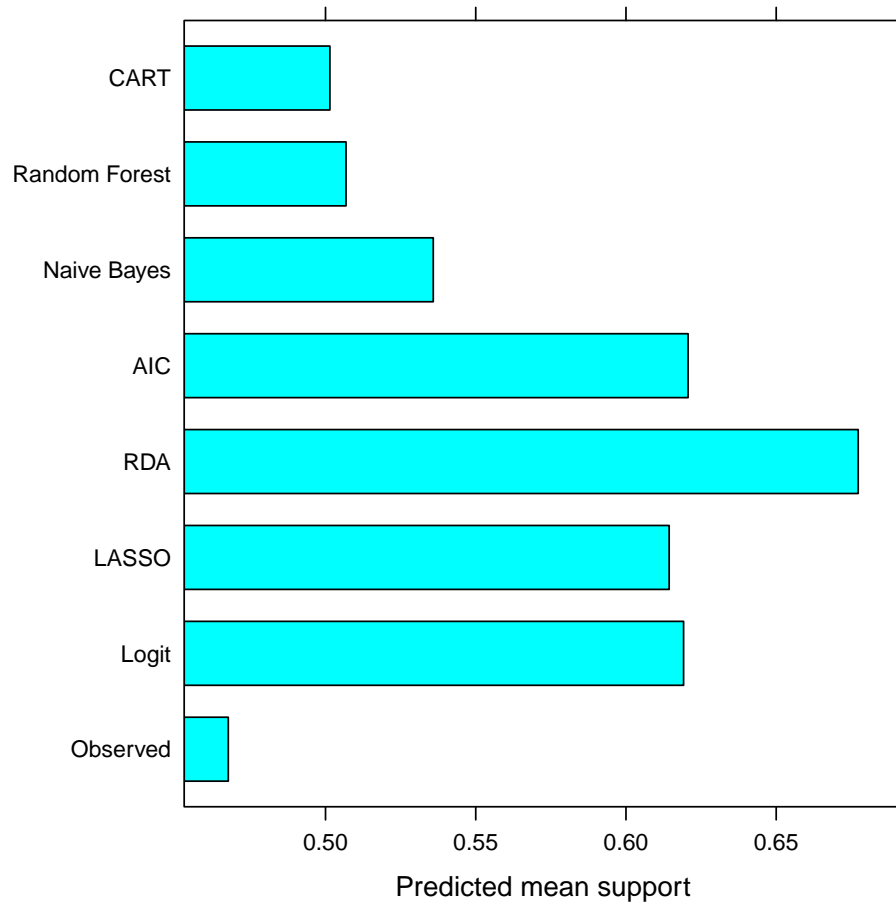


Figure 5: Observed and simulated support of abortion rights.

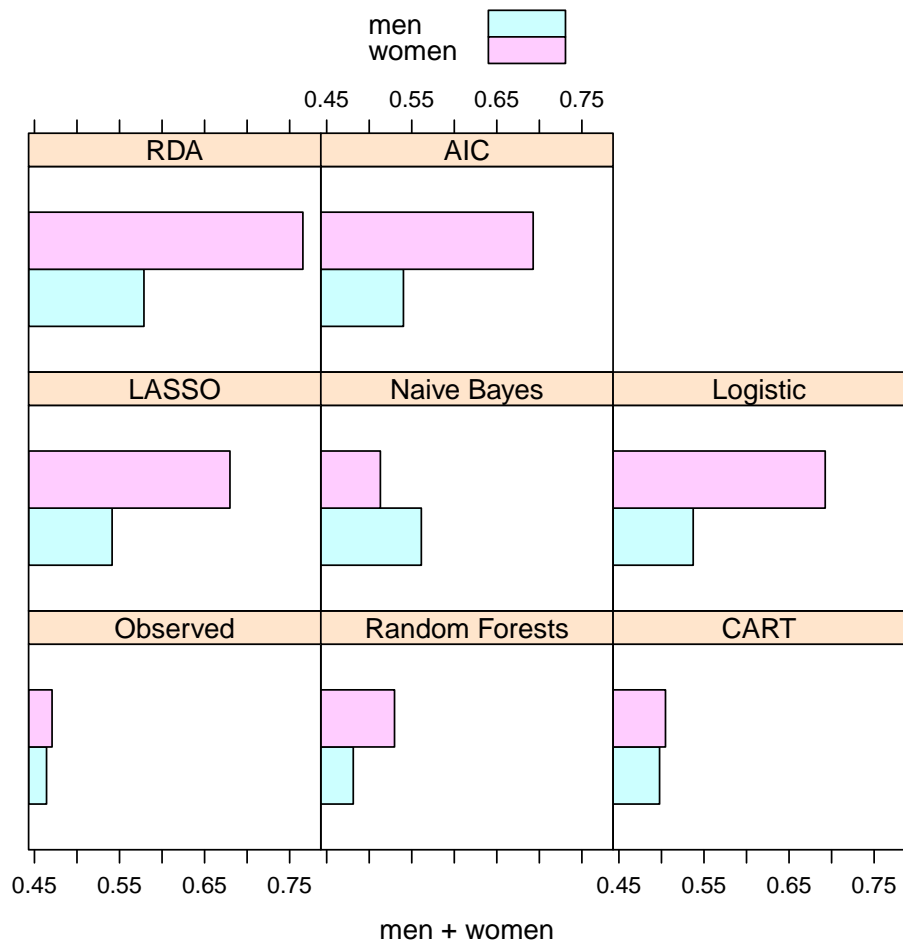


Figure 6: Observed and simulated support of abortion rights broken down by men and women.

	Men	Women
Logit	(0.063, 0.083)***	(0.209, 0.235)***
Random Forest	(0.008, 0.029)***	(0.054, 0.076)***
CART	(0.019, 0.034)***	(0.032, 0.049)***
LASSO	(0.072, 0.087)***	(0.197, 0.217)***
Naive Bayes	(0.05, 0.058)***	(0.066, 0.075)***
RDA	(0.188, 0.242)***	(0.336, 0.395)***
AIC	(0.067, 0.085)***	(0.21, 0.234)***

Table 6: 95% confidence intervals of change in predicted probabilities. Stars for the 0.05, 0.01, and 0.001 levels. Paired t-test.

methods predict large gains for low information respondents while non-linear methods indicate smaller changes when information is set to its highest value. Perhaps more interesting is that the two tree methods, CART and Random Forest, predict *decreases* in average support probabilities for those in the slightly above average to high information levels. Another outlier is the RDA classification method, which predicts extremely large, positive changes in information effects for low information subjects. Given that almost all other methods show much more modes effects, if any, we must question the results from RDA.

## 5 Conclusion

Unlike previous simulation studies, which have relied exclusively on widely known and well-understood linear regression techniques, we have used a variety of machine learning techniques to do two things:

- Identify which, if any, methods outperform logistic regression in accurately classifying novel data. Prediction accuracy has generally been given little attention in previous studies, but accurately predicting reality is a prerequisite for accurately predicting a counterfactual reality.
- Identify the variables that play a large role in accurately predicting support for abortion. This gives us insight into the attributes of individuals that are likely to help predict a counterfactual reality. Previous literature in this area has focused on information effects and so it would be worth exploring in future research whether or not those variables that are deemed important by our GIS measure play an interactive role with knowledge.

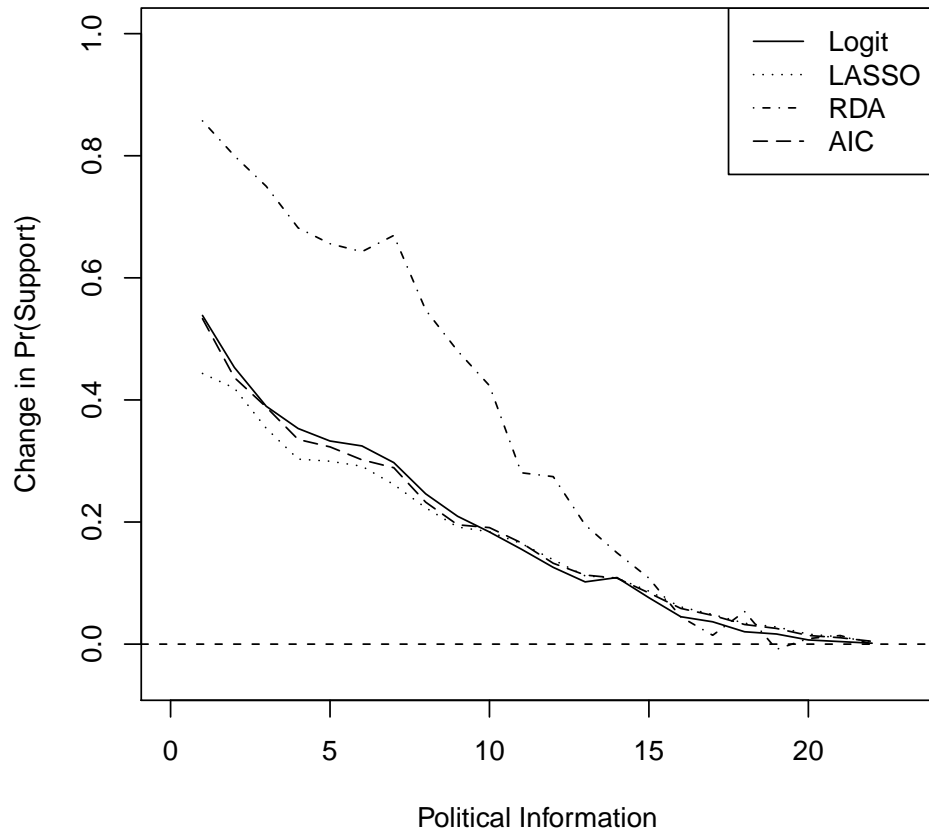


Figure 7: Average change in support across the observed knowledge scale by model.

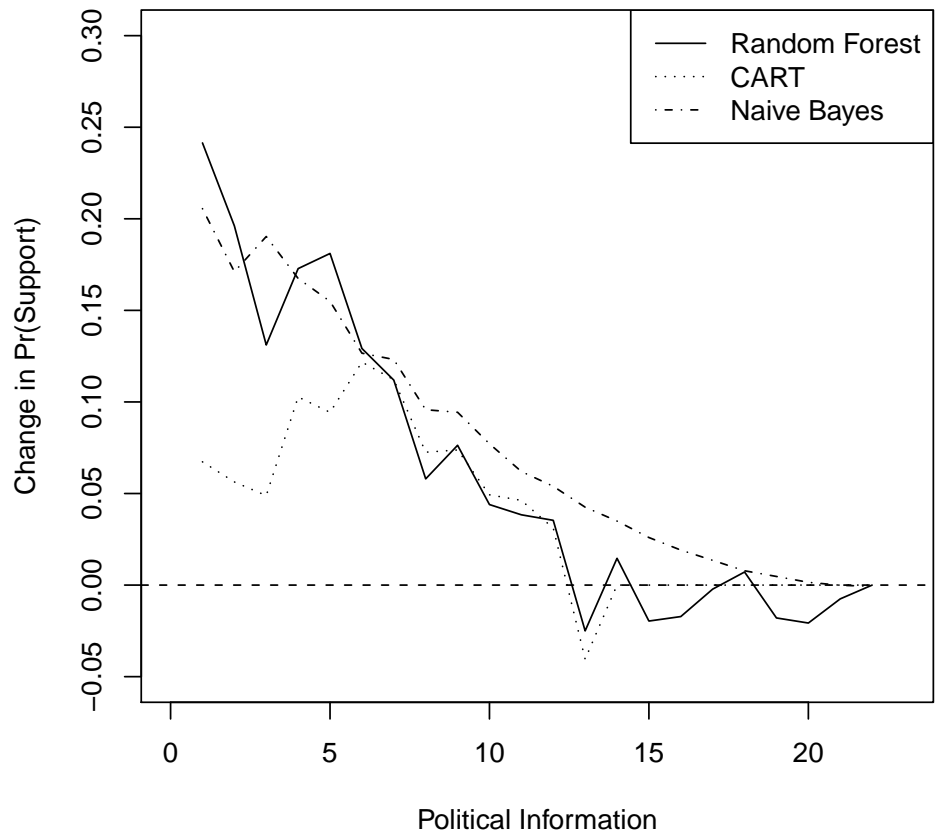


Figure 8: Average change in support across the observed knowledge scale by model.

In fact, the machine learning techniques employed here did not clearly outperform logistic regression. In general, all methods accurately predicted approximately 2/3 of the cases. This may be a function of the specific dataset we used or it may reflect the unpredictable nature of human behavior. Despite an inability to outperform logistic regression in terms of accuracy, the various machine learning techniques offered slightly different estimates of simulated support. Thus, it may be the case that certain conclusions about simulated preferences are conditional on the use of logistic regression. Given the emphasis on theory in the political science discipline, is our use of machine learning techniques justifiable? Can we claim that different simulated predictions call into question the strength of past findings? Variables in past studies were chosen for specific reasons. They are all demographic variables and thus are unlikely to correlate highly with every outcome variable studied (though we are investigating only one outcome variable), with the possible exception of party preference. Nonetheless, our techniques do not introduce new variables but simply penalize (in a variety of ways) those variables that do not help us to predict support for abortion. Theoretically, then, we have not "cheated" in any sense, as we have not introduced new variables that are likely to help our prediction accuracy. Thus, it seems reasonable for us to conclude that if various techniques (all with their own unique assumptions, none of which preclude their use here) predict different simulated results, then the strength of past findings may be called into question.

With regard to our second goal, the various techniques all identify several variables that are important across most models. Information level, income, age, party preference, and religion seem to have a large bearing on accuracy. Simply observing significance levels in logistic regression or the remaining variables in a stepwise AIC search is not enough to make conclusions about variable importance as these both rely on the same underlying linear framework. Rather, seeing the same variables bearing heavily on accuracy across different methods strengthens our belief that they are important. What remains to be seen is whether or not this is unique to this specific policy preference or if these variables translate across all policy preferences.

In light of these findings, future research could proceed in several directions. First, the machine learning methods employed in this paper can be applied to other variables in the ANES dataset (electoral choice, voter turnout, other policy preferences, etc.) as well as to other datasets. This

would allow us, as in this paper, to attempt to improve accuracy as well as compare simulated results. Second, variable importance scores, especially if they are consistent across questions, may lead researchers to focus on a smaller subset of predictor variables. Though including a large number of demographic variables no doubt helps to improve accuracy, a smaller subset of important variables may achieve the same level of accuracy. It would also allow researchers to focus in on those attributes of respondents that seem to be related to information levels and policy preferences.

Regardless, the use of machine learning techniques has long been ignored in political science. This makes sense when it comes to questions that are strongly motivated by theory. Here, though, theory plays a limited role and the overall goal is accuracy. In similar cases, researchers should look to machine learning techniques to boost accuracy (and therefore confidence in their findings) or to put past findings (which have relied on straightforward linear regression) into perspective.

## References

- Althaus, S. L. (1998). Information effects in collective preferences. *The American Political Science Review*, 92(3):545–558.
- Althaus, S. L. (2003). *Collective preferences in democratic politics: opinion surveys and will of the people*. Cambridge University Press, Cambridge.
- Bartels, L. (1996). Uninformed votes: Information effects in presidential elections. *American Journal of Political Science*, 40(1):194 – 230.
- Breiman, L. (2002). Looking inside the black box. Wald Lecture, Banff, Alberta, CA.
- Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The elements of statistical learning*. Springer-Verlag, New York, second edition.
- Kuklinksi, J. H., Quirk, P. J., Jerit, J., Schwieder, D., and Rich, R. F. (2000). Misinformation and the currency of democratic citizenship. *Journal of Politics*, 62(3):790 – 816.
- Kursa, M. B. and Rudnicki, W. R. (2010). Feature selection with the boruta package. *Journal of Statistical Software*, 36(11):1–13.

- Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.
- Mondak, J. J. (2001). Developing valid knowledge scales. *American Journal of Political Science*, 45(1):224–238.
- Siroky, D. S. (2009). Navigating Random Forests and related advances in algorithmic modeling. *Statistics Surveys*, 3:147 – 163.
- von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, 39(1):265 – 291.