How to Cautiously Uncover the “Black Box” of Machine Learning Models for Legislative Scholars

Soren Jordan, Hannah L. Paul, and Andrew Q. Philips

Forthcoming

Abstract

Machine learning models, especially ensemble and tree-based approaches, offer great promise to legislative scholars. However, they are heavily underutilized outside of narrow applications to text and networks. We believe this is because they are difficult to interpret: while the models are extremely flexible, they have been criticized as “black box” techniques due to their difficulty in visualizing the effect of predictors on the outcome of interest. In order to make these models more useful for legislative scholars, we introduce a framework on integrating machine learning models with traditional parametric approaches. We then review three interpretative plotting strategies that scholars can use to bring a substantive interpretation to their machine learning models. For each, we explain the plotting strategy, when to use it, and how to interpret it. We then put these plots in action by revisiting two recent articles from *Legislative Studies Quarterly*.

This is a preliminary version[1] as it appears for the citation:

Jordan, Soren, Hannah L. Paul, and Andrew Q. Philips. *Forthcoming*. “How to Cautiously Uncover the ‘Black Box’ of Machine Learning Models for Legislative Scholars.” *Legislative Studies Quarterly*. DOI: 10.1111/lsq.12378

Corresponding author: Soren Jordan (sorenjordanpols@gmail.com).

[1] When the article is no longer *Forthcoming*, this will be replaced with the appropriate citation and consistent page numbers. The content will be identical.

Introduction

Machine learning (ML) techniques, especially tree-based ones, are growing in popularity in political science (Green and Kern, 2012; Streeter, 2019; Benoit, Munger, and Spirling, 2019; Kim, Alvarez, and Ramirez, 2020). Typically focused on maximizing predictive accuracy, supervised ML methods have often been criticized as a "black-box" approach since it is difficult to make inferences about the effect that a particular predictor has on the outcome of interest.[1] Despite recent articles in political science espousing the benefits of ML models (c.f., Montgomery and Olivella, 2018), ML model interpretation has typically focused on predictive accuracy rather than visualizing the relationship between the predictors and the outcome (Kaufman, Kraft, and Sen, 2019; Anastasopoulos and Bertelli, 2020) (with some exceptions, Green and Kern (2012); Suzuki (2015); Muchlinski et al. (2016); Kim, Alvarez, and Ramirez (2020)). This "black-box" complexity has kept machine learning models from more prevalent usage in legislative studies; we could only find 16 uses of machine learning models in *Legislative Studies Quarterly* from 2010 to 2020, despite their broader growth in political science. Moreover, these articles were almost entirely confined to network analysis (such as Bendix and MacKay, 2017; Metz and Jäckle, 2016; Bonvecchi, Calvo, and Stein, 2016) or text analysis (c.f., Baumann, Debus, and Müller, 2015; Proksch et al., 2019; Goet, Fleming, and Zubek, 2020).

We argue that, with proper guidance on how to visualize the relationship between a predictor and an outcome, machine learning models can be incredibly useful to legislative scholars. In particular, we recommend the use of machine learning models and visualizations as a complementary approach for fine-tuning parametric models and potentially revealing non-linear relationships. Extending a strategy first briefly discussed in Funk, Paul, and Philips (2021), we suggest that scholars engage in the following sequence:

1. Estimate a parametric model that tests theoretically-grounded hypotheses.
2. Use a machine learning approach on the same set of theoretical predictors to evaluate the robustness of the initial parametric tests.
3. Adjust the initial parametric model to account for any nuances revealed in the machine learning approach.

In this supplemental role—data-*informed* but not data-*dredging*—ML models can help scholars investigate potential misspecifications that traditional parametric models may overlook. For example, a ML approach can account for non-linear effects often involved in models of legislative activity (for instance, logarithmic or quadratic effects of age, ideology, or seniority: Bowler, McElroy, and Muller, 2020; Crosson et al., 2019). Such models can also offer a unifying framework for truncated variables with many zeroes, like fundraising (Bonica, 2020) and lobbyist spending (McKay, 2020). Finally, certain ML models allow legislative scholars to investigate a variety of potential interactions, as predictors are inherently conditional instead of interactions being specified parametrically (Anderson, Box-Steffensmeier, and Sinclair-Chapman, 2003; Osborn et al., 2019; Howard and Owens, 2020, among many others). Our goal is to encourage legislative scholars to consider ML models not as opaque "black-box" approaches that inhibit theoretical inference, but rather a powerful tool to uncover hidden, complex relationships that parametric models might miss.

To facilitate the use of machine learning models, we first provide a brief background on common tree-based machine learning approaches, as well as present a framework for integrating them with traditional parametric models. We then introduce three graphical interpretation tools: Variable Importance Plots, Partial Dependence Plots and Individual Conditional Expectation plots. For each, we define what the plot illustrates, how to interpret it, and when to use it, noting both the plots' respective benefits and constraints. We illustrate the utility of these models and plots for legislative scholars by revisiting two recent articles from *Legislative Studies Quarterly*. These examples offer substantive breadth from both US and non-US legislatures, and also illustrate the modeling flexibility of our ML approach (one illustrates a classification example for a binary variable, the other example is for a continuous outcome). Further, when we apply the machine learning approach as a supplement to the authors' original grounded theoretical parametric models, we uncover substantively interesting non-linearities and interactions that illustrate how ML tools can be used to reinforce and invigorate novel tests of theoretical expectations.

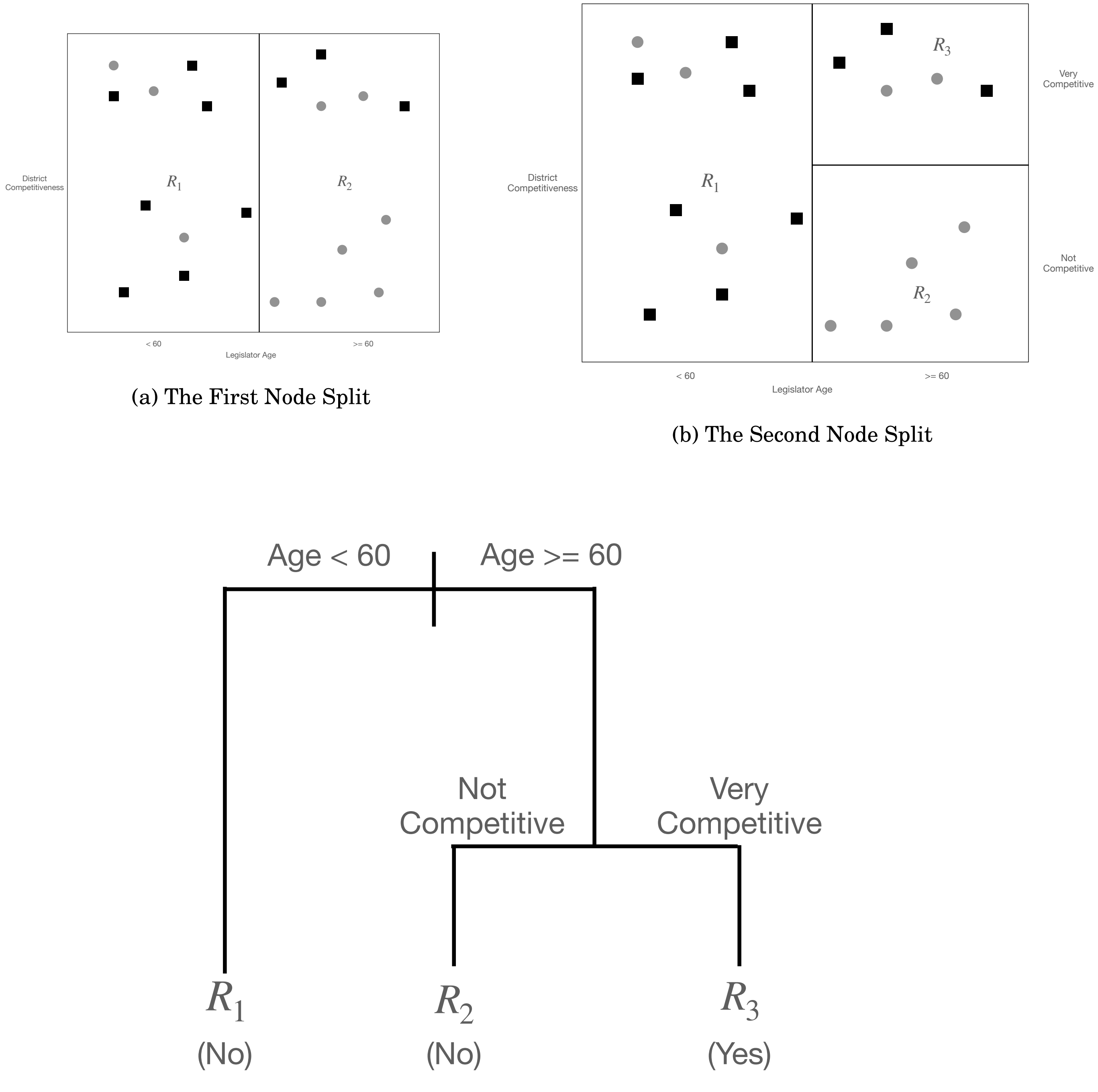
A Primer on Machine Learning Approaches

Generally, ML algorithms specify a continuous or categorical outcome of interest as a function of k predictors for i observations $f(\mathbf{x}_i)$. These are mapped onto the dependent variable. Typical ML applications focus on predictive accuracy. One extremely common ML approach are Random Forests, a tree-based non-parametric approach to classifying or predicting outcomes (Breiman, 2001). The core of tree-based algorithms start with a Classification and Regression Tree, or CART model, which works as follows. Along the range of one of our predictors, the algorithm selects a "cutpoint," or value that best partitions the data into two regions. Optimal partitions (sometimes called a greedy approach) are simply those which maximize our predictive ability—typically the proportion correctly predicted (if the dependent variable is categorical) or mean squared error (MSE) (if the dependent variable is continuous).

This is better illustrated through an example, which we present in Figure 1. Imagine we are trying to classify whether a legislator will vote yes (shown as gray circles) or no (black squares) on a bill, using two predictor variables, legislator age and whether their district is electorally competitive or not. The CART model first finds an optimal cutpoint; in Figure 1a this is a split along legislator age at 60, which partitions the dataspace into regions R_1 and R_2. Given this first split, our best guess is that any legislator in R_1 (i.e., less than 60) will vote no (leading to a 30% misclassification rate since 3 of the 10 observations in R_1 are gray circles), and in R_2 (greater than or equal to 60) will vote yes (this region also has a 30% missclassification rate). Successive splits can be made which further partition the dataspace in a given region. In Figure 1b, the CART model now partitions R_2 into two regions—R_3 contains legislators in highly competitive districts (for which we now predict black squares) while R_2 contains legislators in non-competitive districts (in which we perfectly predict gray circles).

We can improve on predictive ability by conducting additional splits. CART models can be succinctly depicted as trees, as done in Figure 1c, which shows each split, or "node," that was done in Figures 1a and 1b. In normal applications with many predictors, CART models would continue growing a tree—in effect continuing to partition subsequent data-spaces—until some stopping criterion is reached (e.g., stop when there are only 5 observations left in a node), leading to a "terminal node." With our newly-created CART model, we can feed new observations in to make predictions. For instance, a 75 year old legislator in a very competitive district is likely to vote yes on a bill since they would end up in the R_3 terminal node. It is also worthwhile to compare the CART approach to standard regression. While a linear regression would find that competitive districts are more likely to have legislators vote no, and would conclude that as age increases legislators are more likely to vote yes, it lacks the sharp (and inherently non-linear) cutpoints that CART models contain. This allows for more complicated functional forms than are often feasible in parametric models.

The above description describes the construction of a single tree, which tends to overfit data in real-world examples. Random Forests—the machine learning approach we use here—innovate



(a) The First Node Split

(b) The Second Node Split

(c) Tree Representation

Figure 1: Classification and Regression Tree (CART) Example

on CARTs in two ways. First, Random Forests bootstrap the data—given a dataset of size N, randomly draw N observations with replacement; do this many times, creating B bootstrapped datasets—which allows the construction of hundreds or even thousands of CARTs. This process is called *bagging*. Thus, instead of making a single prediction, Random Forests create an "ensemble" of predictions by averaging over all trees (hence the "Forest" in Random Forest). Second, by forcing the algorithm to choose from a random subset of predictors at each node, trees end up becoming decorrelated from one another.[2] Rather than having a series of trees that all essentially make the same sorting choices (which would offer little leverage over a single decision tree), we can further boost predictive accuracy by constructing a series of forcibly decorrelated trees (Hastie, Tibshirani, and Friedman, 2013).

The bagging procedure also produces a set of "out-of-bag" observations. These observations occur because of the nature of Random Forest models. Growing B trees requires taking B bootstrapped samples of the original dataset. As with standard bootstrapping procedures, N observations are drawn (i.e., randomly draw N observations with replacement, which is the same number of the total observations in the dataset). A bootstrapped sample will include roughly two-thirds of all unique observations in the original dataset, meaning that approximately one-third of the sample observations were *never* drawn in a particular bootstrap sample (Hastie, Tibshirani, and Friedman, 2013). It is this one-third that is the "out-of-bag" sample, since these observations were not used to train a particular regression tree (these out-of-bag samples of course differ across the B trees, just like the bootstrap samples randomly vary from tree to tree). Out-of-bag samples are quite useful; since they were not included in the bootstrap sample for tree b, they can be used as an out-of-sample prediction.

Finally, notice that the cutpoints illustrated in Figure 1 are generated based on predictive ability. While this limits the *hypothesis testing* ability of machine-learning models (i.e. we cannot create statistical tests of whether a linear relationship exists between two variables), they offer an extraordinary insight into potentially nuanced, non-linear or even interactive relationships between variables, which is why we advocate for their use. Consider a simple example of a quadratic relationship between a predictor x and some outcome. If using a parametric model, failure to include both x and x^2 can lead to omitted variable bias and a linear fit will poorly approximate the underlying relationship. With tree-based models, users do not need to specify any functional form; the model will automatically find the quadratic relationship on the basis of the predictive ability. The same intuition can be extended to interactions. In parametric models interaction terms must be specified and included, typically assuming linear interactive relationships (but see Hainmueller, Mummolo, and Xu, 2019). Given tree-based model's iterative approach to finding cutpoints, any interactions between variables that increase predictive ability will be included. This is one of the main sources of leverage in the machine-learning approach to *model building*, if not *hypothesis testing*.

We focus rather exclusively on Random Forests. While many other ML approaches exist, we think Random Forests are particularly useful to legislative scholars for several reasons. First, they allow us to model complex, potentially non-linear relationships as well as hidden interactions; as Montgomery and Olivella (2018, p. 729) put it, "standard models are often insufficiently flexible to capture nuances in the data—such as complex nonlinear functional forms and deep interactions—when no clear a priori expectations exist." Second, Random Forests allow the inclusion of a substantial number of covariates, unlike parametric approaches which run the risk of an overly-saturated model. Estimating a Random Forest with a additional potential predictors allows us to see how well the key predictors from the parametric model hold up against a large(r) group of potential covariates. If the Random Forest model shows that predictors of theoretical interest also have large explanatory power relative to an unconstrained set of predictors, this provides evidence of the importance of these explanations. Or, the Random Forest might help adjudicate between which additional predictors to include as controls, depending on which are indicated to have predictive power.[3] Last, Random Forests are popular ML tools that have been used elsewhere in political science (Suzuki, 2015; Muchlinski et al., 2016; Funk, Paul, and Philips, 2021), including legislative politics (Bonica, 2018), helping ensure Random Forest models employed by legislative scholars would find wider audiences within political science.

However, this approach does come with some limitations. First, if basic parametric assumptions can be met, then traditional approaches, like OLS or logistic regression, may have lower bias than Random Forests (Muchlinski et al., 2016).

Second, machine learning models do not allow for traditional null-hypothesis statistical significance testing. In other words, if we have specified the functional form correctly, not only are traditional parametric estimates easier to interpret, we can also create measures of uncertainty for them. However, in the cases in which researchers *cannot* meet regression assumptions, or when they have nonlinear or conditional expectations about their data generating process, then ML approaches may actually strengthen inferences.

Integrating Machine Learning Models with Parametric Models

As previewed in the introduction, we recommend using ML models as robustness checks. Users can estimate their original, theoretically grounded parametric model specification, then re-estimate the same set of predictors using a Random Forest model, which can then be used to gain insight into their original findings. Based on these new findings, users can fine-tune their parametric models, yielding better specified models along with measures of uncertainty.

Specifically, we propose leveraging ML's fully non-parametric approach after executing a standard theoretical parametric model, which is limited in its ability to capture complex data generating processes, such as conditional and nonlinear effects. The replication of a theoretically-grounded parametric model using ML models and visualizations allow researchers to leverage the greater flexibility in order to refine their understanding of theoretically interesting relationships under examination. ML findings may mirror the results from the initial linear models, serving as a traditional robustness check. Alternatively, this second set of findings could reveal interactions or non-linearities that users can approximate in their linear models using covariates such as interaction terms or splines. The shadow side of this data-driven technique is the potential for data-mining. That is, a researcher could hypothetically mine a large dataset for the best predictors rather than grounding their model specification in theory. Our three-step approach should help obviate this concern: as with *any* model, we should always start with theory before moving to estimation, allowing us to leverage the benefits of data-driven modeling that is still rooted in theory.

To make machine learning models effective, however, we need to resolve their most glaring deficiency: the difficulty in visualizing or interpreting the effects of predictors on the outcome. To better interpret these models, a series of graphical tools—the most popular of which we discuss below—have been created to assist users. All of the ones we cover, with the exception of Variable Importance Plots, can be used with *any* model, not just Random Forests. However, they are uncommon in parametric models since estimates tend to be easy to interpret (e.g., by examining coefficients, marginal effects, etc.). We explain these three common graphical interpretation tools essential to using Random Forests in legislative scholarship below.

Graphic Interpretation Tools

Variable Importance Plots

What it is: A Variable Importance Plot (VIP) helps indicate which variables of interest are strong predictors. While there are several different ways to create a measure of variable "importance" in Random Forest models, the basic idea is as follows (Breiman, 2001). After estimating all trees, take each of the out-of-bag observations (those observations for which the tree was *not* estimated on) and permute (i.e., reshuffle) a predictor. The justification is that permutation, "effectively voids the effect of a variable, much like setting a coefficient to zero in a linear model" (Hastie, Tibshirani, and Friedman, 2013, p. 593). The out-of-bag observations with the permuted predictor are then fed through their corresponding tree, from which we calculate the prediction error (either mean squared error for a continuous dependent variable, or the classification error rate for a categorical dependent variable). The overall decrease in accuracy is therefore the difference between this prediction error when the variable is permuted, and the prediction error on the out-of-bag observations when the variable is not permuted. This procedure is done for each predictor. More formally, given some variable of interest, say x_s, we can first obtain the average out-of-bag prediction error for all trees for which x_s was included for at least one split. This average prediction error is compared to another average where x_s is randomly permuted.[4] VIPs show the scaled difference between the original prediction and the permutation.[5]

How to interpret it: VIPs are typically shown with all predictors lined up from most to least important, where each value shows the average decrease in accuracy, scaled by its standard deviation in order to normalize the measures.[6] The more "important" predictors have a much larger decrease in accuracy (i.e. permuting results in large decreases in predictive accuracy); the larger the values, the better the predictor does at reducing prediction error. Typically, the y-axis of a VIP is percent reduction in prediction error, so we can interpret individual values as the percent of prediction error reduced by the inclusion of any particular variable. When interpreting the y-axis of a VIP, users should note the *relative* differences in prediction error reductions.[7] Thus, while VIPs are mostly used to show *relative* importance between predictors, implying that their scales are not constant from model to model, they do have one very intuitive value. A value of zero indicates that a variable is no better at predictive accuracy than random noise.

When to use it: VIPs are among the first plots that a scholar should produce after estimating a machine learning model. They are often the best initial insight into which predictors are doing the most "work" in predicting or classifying the dependent variable. With a VIP in hand, then, scholars can argue for the relative importance of a particular variable in predicting an outcome (as a robustness check to their theoretical specification), using the VIP to identify whether the scholar's predictor is among the most important in a large set of predictors. Since Random Forests are not at risk of over-saturation like parametric models, VIPs allow researchers to evaluate how well their theoretically important covariate(s) predict their outcome of interest compared to the covariates included in their original model as well other possible predictors that a parametric model would be over-saturated by (if included). Take for instance Funk, Paul, and Philips (2021), who examine the relationship between women's legislative representation and government spending. They use VIPs to discern whether women's representation even appears to be associated with spending, and find that, relative to other common determinants of government spending, women's representation is very important. Additionally, we can use VIPs to ensure that no important control variables are omitted from future models. If a parametric model is to be used in future analyses (which may run the risk of being overparameterized), the user might take the top 10, 15 or 20 most important predictors, as shown from a VIP, since they tend to be highly predictive of the outcome of interest. However, we do not believe that just because a theoretically-important variable is not the *most* important predictor means it should be thrown out of future parametric models. Instead, VIPs help show the explanatory power of each predictor relative to one another, as well as relative to a meaningful "null" of zero predictive power.

Partial Dependence Plots

What it is: One common way to visualize the results of ML models is thorough Partial Dependence Plots (PDPs), which show the average marginal effects on the prediction using training data (Friedman, 2001). Let x_{is} be our independent variable of interest for observation i, and $\mathbf{x}_{ic} = \{x_{i1}, x_{i2}, \cdots, x_{ic}\}$ be all predictors not including x_{is}. Given our prediction function, $\hat{f}(\mathbf{x}_i)$ (where $\mathbf{x}_i \in \{x_{is}, \mathbf{x}_{ic}\}$), then the partial dependence function is (Greenwell, 2017):

$$f_{PDP}(x_s) = \mathrm{E}_{x_s}[\hat{f}(x_s, \mathbf{x}_c)] = \int \hat{f}(x_s, \mathbf{x}_c) p_c(\mathbf{x}_c) \mathrm{d}\mathbf{x}_c \tag{1}$$

where:

- $f_{PDP}(x_s)$ is our PDP prediction for variable x_s
- $\mathrm{E}_{x_s}[\hat{f}(x_s, \mathbf{x}_c)]$ is the expected value/probability of y_i, given our prediction function $\hat{f}(x_s, \mathbf{x}_c)$
- $\int \hat{f}(x_s, \mathbf{x}_c) p_c(\mathbf{x}_c) \mathrm{d}\mathbf{x}_c$ is the integral across our prediction function ($\hat{f}(x_s, \mathbf{x}_c)$) times the marginal probability density, $p_c(\mathbf{x}_c)$, of all control variables $\mathbf{x}_c$, which is given as: $p_c(\mathbf{x}_c) = \int p(\mathbf{x}) \mathrm{d}x_s$

In practice, this is estimated as:

$$\bar{f}_{PDP}(x_s) = \frac{\sum_{i=1}^{N} \hat{f}^{(i)}(x_s, \mathbf{x}_{ic})}{N} \tag{2}$$

The intuitive explanation here is that we hold all control variables, $\mathbf{x}_{ic}$, at their actual value for each observation i. We then set x_s to a single value of x_s for all observations (e.g., fix $x_s = 1$ for all observations). Next, this fixed value of x_s (as well as the corresponding $\mathbf{x}_{ic}$ values specific to each observation) is run through the prediction function, $\hat{f}^{(i)}(x_s, \mathbf{x}_{ic})$, in order to obtain a prediction for every observation i. We then fix x_s to the next value (e.g., $x_s = 2$), calculate new predictions, and continue this process of iterating across all values of x_s for which we want to create the PDP. As a final step, we average over all resulting predictions in order to obtain the average predicted value of the outcome, $\bar{f}_{PDP}(x_s)$, given each value of x_s. This forms the PDP. This same logic can be extended to multiple independent variables, allowing us to create PDPs that investigate interactive relationships between predictors.

How to interpret it: A PDP shows the expected value of the outcome—given a particular value of one variable—after accounting for the average effects of all other predictors (Hastie, Tibshirani, and Friedman, 2013). A PDP will show the observed range of an independent variable along an x-axis with the predicted value (or probability of classification, if the dependent variable is categorical) of the outcome along the y-axis. It should be interpreted, then, as the predicted value of the dependent variable across the range of an independent variable, after accounting for the average effects of all other predictors. There is no universal standard for what constitutes a "large" effect in a PDP, so users should be careful to place the effects on the scale of the dependent variable. Another option is to compare the size of the PDP effect to the effects estimated through the parametric model.

When to use it: Any time a scholar has a particular variable of interest—i.e., a theoretically important variable—they should consider investigating the PDP for the relevant independent variable. This is the first, most informative plot on the nature of the relationship between some variable of interest and the dependent variable. As an aside, since the PDP shows an expected value while averaging over other predictors, the plot contains only a single prediction line. For a discipline increasingly accustomed to errors in predictions (Kastellac and Leoni, 2007), a lack of uncertainty may be unsatisfying. However, given the non-parametric nature of machine learning models, it is not possible (and arguably inconsistent with the approach) to specify or calculate traditional parametric tools of inference, like standard errors or confidence intervals. But, since the PDP is averaging over the N predictions in Equation 2, it is possible to emulate a “confidence interval” (though it should not be referred to as such) by additionally calculating the standard deviation (or some other quantity) of the N individual predictions and plotting it, in addition to the PDP line. While this approach does not provide the same hypothesis testing leverage as a traditional standard error or confidence interval, it can be used to investigate other interesting questions, such as where the PDP diverges considerably or where there is more or less heterogeneity in the predictions.[8] The more commonly accepted method of addressing this heterogeneity in the individual predictions, though, is another plot entirely, which we discuss below.

Individual Conditional Expectation Plots

What it is: One critique of PDPs is that by only showing the average prediction across observations, they might overlook heterogeneity in the effects. Individual Conditional Expectation (ICE) plots (Goldstein et al., 2015) can account for this by showing individual predictions, $\hat{f}^{(i)}(x_s)$, rather than the average, as done with PDPs. The average/PDP can even be indicated in the ICE plot by overlaying it with a distinctive color or line type. As it is common to show the actual observed x_i value for each ICE line, only a percentage of all the individual ICE curves are typically shown when there is a high number of observations for the sake of clarity.[9] Two extensions of ICE plots involve centering all curves around a fixed point (a “c-ICE” plot) as well as taking the first derivative of the curves (“d-ICE”); for brevity we discuss these variants in the Online Appendix.

How to interpret it: Like a PDP, the ICE plot will show the the observed range of an independent variable along an x-axis with the predicted value/probability of the outcome along the y-axis. Recall that it differs from the PDP by showing individual conditional expectations, so there will be multiple lines. Each line illustrates an individual expectation. The user decides how many individual expectations to show (driven largely by how many will fit). If the user is randomly sampling lines to draw, as with any randomly generated plot, a seed should be set to ensure reproducibility.

When to use it: Researchers should use an ICE plot if they anticipate a potential relationship to be heterogeneous, such that the PDP obscures individual relationships. As a corollary, if the average relationship describes the fit well, and there is no evidence of substantial individual-level heterogeneity, the PDP should suffice. Users can evaluate this by simply comparing the ICE plot to the PDP plot for the same variable: if most lines in the ICE plot look similar to the average (the PDP), the PDP would present the same information with more clarity. Last, ICE plots can be used to show the relationship between x_s and a second variable through the use of shading. Let x_z be some other predictor. If x_z is categorical, $\hat{f}^{(i)}$ can be shaded by the value of x_{iz}. If x_z is continuous, shades can be assigned according to whether x_{iz} is on either side of some threshold—such as the median, mean, or certain percentiles—or a shade gradient can be used, making these thresholds more continuous-feeling.

Taken together, these three visualization tools form the core toolkit for an analyst looking to use ML models as a robustness check of a theoretical specification. With them, the analyst can evaluate which predictors are most important in explaining a variable as well as evaluate the nature of the relationship between an independent variable and the dependent variable. We next illustrate how these tools can be applied in practice through replicating two recent articles from *Legislative Studies Quarterly*.

Example 1: Howard and Owens (2020)

Howard and Owens (2020) (henceforth HO) examine the conditions under which bills bypass US Senate committee proceedings. The authors use a cross-sectional time-series of Senate “S" bills from 1985 to 2014 and find that bills introduced by ideologically extreme minority-party members of the Senate are more likely to bypass Senate committee proceedings. HO estimate a logistic regression model for their dichotomous bypass dependent variable, measured so that “1” represents a bill bypassing committee and “0” represents taking a bill to committee.[10] A central piece to HO’s argument is that “institutionally weak" bill sponsors have incentives to take a more individualistic approach to introducing a bill to the Senate floor. We were able to replicate their logistic model exactly, as shown in Table 1.

Table 1: Replicating Howard and Owens (2020): Predicting Bypass

	Coefficient	Standard Error
Polarization	-0.130	(0.015)*
Cosponsors	-0.007	(0.004)*
Bills Introduced	-0.002	(0.002)
Extremity	0.004	(0.004)
Time Remaining	-0.025	(0.003)*
Sponsor Seniority	0.006	(0.005)
Duplicate Bill	0.486	(0.088)*
Committee Chair	0.665	(0.124)*
Floor Leader	1.318	(0.339)*
Minority Sponsor	-0.434	(0.221)*
Up for Election	0.127	(0.050)*
Nontrivial Bill	0.784	(0.114)*
Party Bill	0.911	(0.247)*
Extremity * Minority	0.007	(0.006)
Minority * Cosponsors	0.012	(0.006)*
(Intercept)	6.076	(0.933)*
Log Likelihood	-13837.45	
N	48006	

Note: Dependent variable is if the bill bypassed Senate committee, replicating Howard and Owens (2020), Table 2, Model 1. Includes fixed effects for Congresses and policy area. Two-tailed tests. * $p < 0.05$.

We then use a Random Forest modeling approach, expecting that the machine learning model will replicate the authors' main findings while additionally uncovering hidden interactions between the authors' primary variables of interest.[11] Notice, for instance, in Table 1 that the theoretical interaction between minority status and member extremity needs to be explicitly parameterized in order to be estimated (Hainmueller, Mummolo, and Xu, 2019). Notice also that other variables emerge as conventionally statistically significant, but the original authors focus on interpreting the explicit interactions. To help illustrate the value of machine learning models and to put the plots above into action, we create VIP, PDP and ICE plots from the Random Forest model.

Variable Importance Plot

Figure 2 shows the VIP. This plot orders the predictors by the amount of error they reduce in the model, thus indicating their

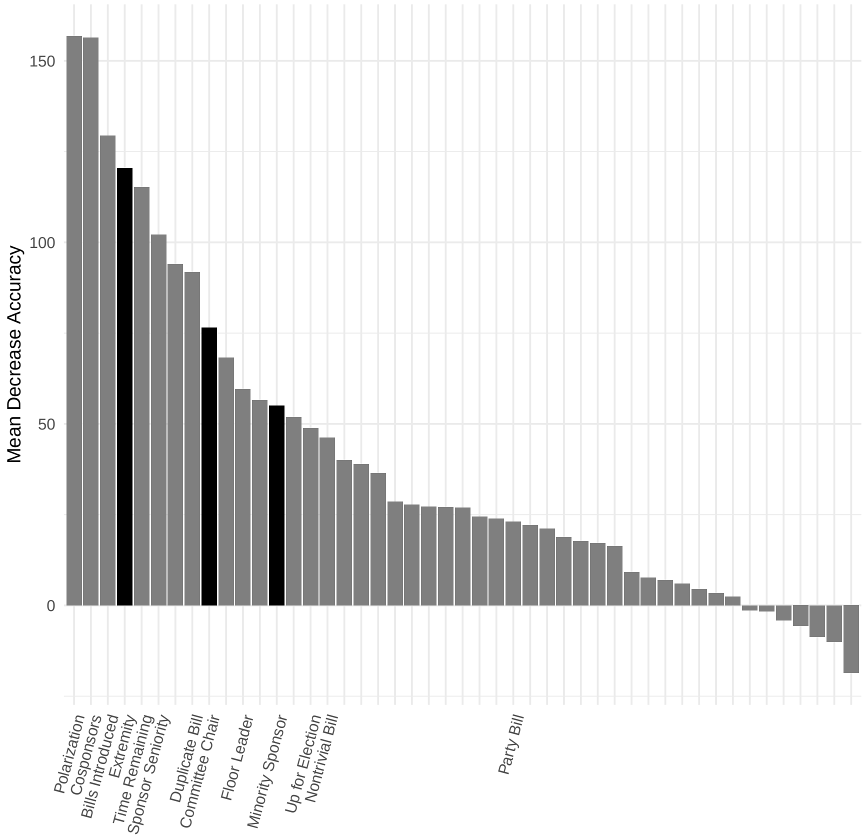


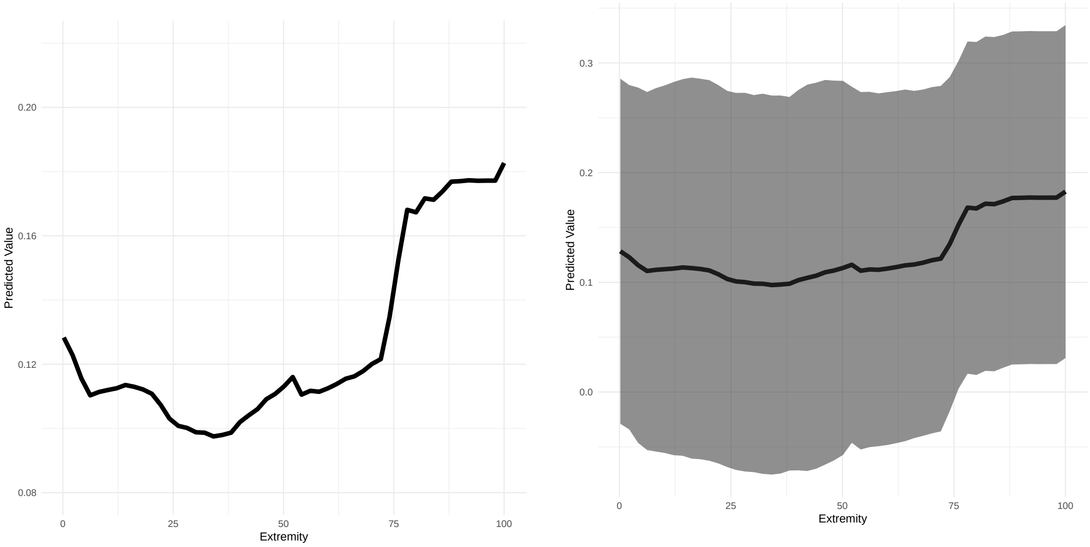
Figure 2: Variable Importance Plot. HO's Key Theoretical Variables Are Important Predictors (Omitting Labels for Fixed Effects)

"importance." The greater the value of mean decrease accuracy on the y-axis, the better the predictor does at reducing prediction error. Recall that a value of zero on the y-axis indicates that a variable's predictive accuracy is no better than random noise. HO include a variety of fixed effects (for Congresses, policy areas, and years): we omit their labels on the plot to help provide clarity about which substantive predictors provide the most predictive ability. The leverage of the plot is that the VIP provides not only a check of the theory but also encourages us to investigate the variables we were not expecting to exert influence.

The VIP reveals that the predictive ability of HO's key predictors, denoted by black bars, are far from zero and relatively better at reducing prediction error compared to other covariates in the model. HO's main explanatory variables fall within the top 13 most important predictors. The ideological extremity of sponsors is the fourth most important predictor for the bypass outcome; the mean squared error of a randomly permuted extremity variable increases by over 100 percent relative to the actual variable. Thus, we would conclude that including the extremity variable is important if we are trying to predict bypass outcomes. In addition, we find that the committee chair variable is the ninth most important and minority-party member is the thirteenth. However, we also discover interesting nuance to HO's theoretical story, as the most important predictor is chamber polarization (which HO discuss very little). Overall, the VIPs confirm the importance of HO's explanatory variables of interest, but indicate an opportunity to explore the nature of the effects of chamber polarization (an increasingly important predictor of Senate behavior: e.g. Basinger and Mak (2020)), which we examine in further detail below.

Partial Dependence Plot

Using the VIP plot in Figure 2 as a guide, we can investigate HO's original theoretical variables of interest, in particular being attentive to their relationship with polarization. The first PDP is shown in Figure 3 and shows the predicted value that a bill will bypass a Senate committee for each value member extremity, after accounting for the average effects of all other predictors in the model. Figure 3a displays a traditional PDP with a single line for the expected value of the dependent variable. Figure 3b adds a shaded region, which represents the standard deviation of the N individual predictions. This shaded region is not a confidence interval; however, it does indicate some heterogeneity in the predictions, which we can further explore in an ICE plot. The PDP reveals a more nuanced picture of the positive association between ideological extremity and the likelihood to bypass committee than the one we find in HO's analysis. The relationship between these two variables is non-linear. The distribution of the predicted values is convex with the lowest inflection point around 35 on the ideological extremity scale, which ranges from 0 to 100. A positive relationship between ideological extremity and the probability of



(a) Traditional PDP

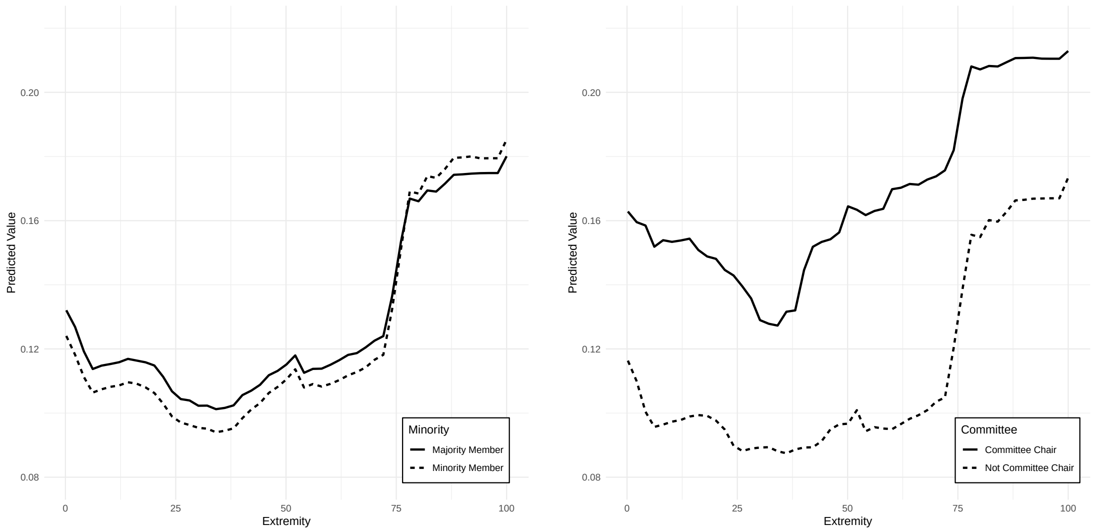
(b) PDP, with One Standard Deviation Shaded

Figure 3: Partial Dependence Plots: Member Extremity

bypassing committee does not occur until after this point. Once ideological extremity surpasses the midway point (50), the predicted values steeply increase, suggesting that as sponsors become more ideologically extreme, they become much more likely to bypass committee to introduce their bill to the floor of the Senate. We also note that predicted value of bypassing committee increases the most sharply after 75 on the ideological extremity scale after which point, the number of observed values for the predictor decreases. This means that relatively few observations are explaining the highest predictions. Taken together, the PDP suggests that both moderate and extreme members are the most likely to have their bills bypass committee: suggesting potentially interesting nuances with regard to (bi)partisanship.

To test the nonlinear relationship observed in Figure 3 within a parametric framework, we could create a knot based on the inflection point observed at an ideological extremity score of 35 using a linear spline approach. For example, if the regression coefficient for the effect of ideological extremity on bypassing below the 35 extremity score knot were not statistically significant and the coefficient above the 35 threshold point were statistically significant, this would suggest that ideological extremity only matters for bypassing behavior after a certain threshold level. We could recommend that researchers use AIC and BIC tests to assess whether or not a piecewise estimator is a better fit compared to a traditional linear estimator, which specifies the extremity covariate in its original form. The smaller the AIC/BIC test statistic, the better the relative fit of the model given the data. So, if the AIC/BIC test statistics are smaller for the piecewise regression compared to the original linear model specification, then this provides further support for a nonlinear data generating process. The relationship displayed in Figure 3 could also be interpreted as quadratic. We could include a quadratic term in a traditional logistic regression by squaring the ideological extremity independent variable. Squaring the extremity variable is a more blunt approach to capturing the nonlinear relationship between ideological extremity and bypassing committee; however it is simpler than a linear spline approach. If the squared variable were statistically significant, this would provide support for a nonlinear relationship between ideological extremity and bypassing committee. Our larger point, though, is that the ML model encourages us to examine more closely whether there is a potential nonlinearity, complemented by theory, that is not well captured in the parametric approach.

Yet another benefit of ML models is that they are inherently interactive as well as being inherently non-parametric (as shown above). In the next two PDPs, we interact ideological extremity with the two other key explanatory variables to see if and how they condition this fundamental finding. In Figure 4a, we find



(a) Conditioned by Minority Status. (b) Conditioned by Committee Chair Status.

Figure 4: Interactive Partial Dependence Plots: Member Extremity.

partial support for HO's key interactive hypothesis that minority-party membership conditions how ideological extremity affects likelihood to bypass. Minority sponsors that fall under the most extreme ideologically quadrant (75-100) appear to be 0.02-0.03 more likely to bypass committee than majority sponsors. Prior to this cutpoint at 75 on ideological extremity scale, majority-party members are more likely by approximately 0.02 to bypass committee compared to minority-party members. Figure 4a not only portrays the conditional nature of relationship between ideological extremity and minority-party membership on the outcome of interest, but it also uncovers non-linear relationships, which cannot be uncovered in HO's original analysis due to their logistic regression model's parametric linearity assumption. In Figure 4b, we show that the nature of the relationship between ideological extremity and bypassing committee is the same for committee and non-committee chairs (the form of the PDP line is pretty similar for both party statuses). Ideological extremity does not appear to be conditional on committee chairmanship, though being a committee chair does produce a large, positive intercept shift. Observing this relationship might lead us to test further features of the authors' original theory: especially given recent research about the continued deference to committee leaders, even in an era of high polarization (Curry, 2019).

We can use PDPs to investigate interactions among any two predictors without needing to specify it formally in the model. For instance, HO treat chamber polarization as a control variable and give it little attention in their discussion of the results, but the VIP suggests it is one of the most important predictors. As such, we might want to explore the effects of polarization with a set of PDPs that explore chamber polarization's independent and conditional effects on the outcome of interest. As shown in Figure 5, chamber polarization has a negative effect on the likelihood to bypass committee, affirming HO's general finding. Interestingly, the plot suggests that when the distance between the two parties' median members falls approximately between 60 and 67, the decline in likelihood to bypass is the most steep. As we move along the x-axis after this range, the predicted values for bypassing begin to level out, such that bypassing is extremely unlikely (less than 5%) after polarization reaches 70. Figure 5 suggests that there are non-linearities in chamber polarization's effects, impossible to observe in the logistic regression framework but possible under the machine learning framework.

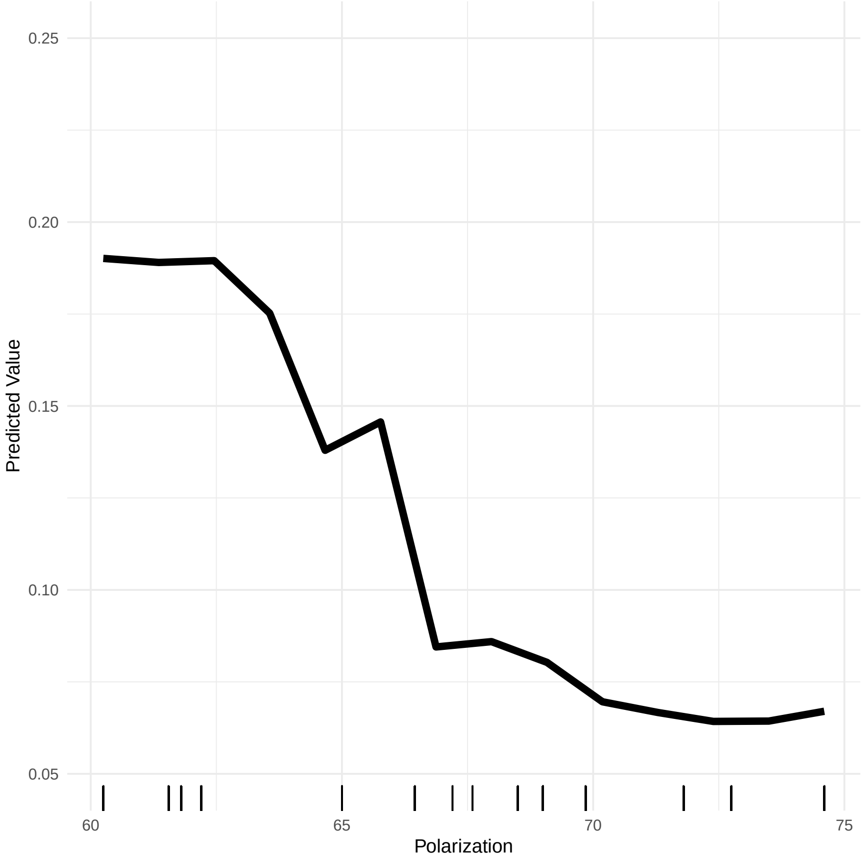
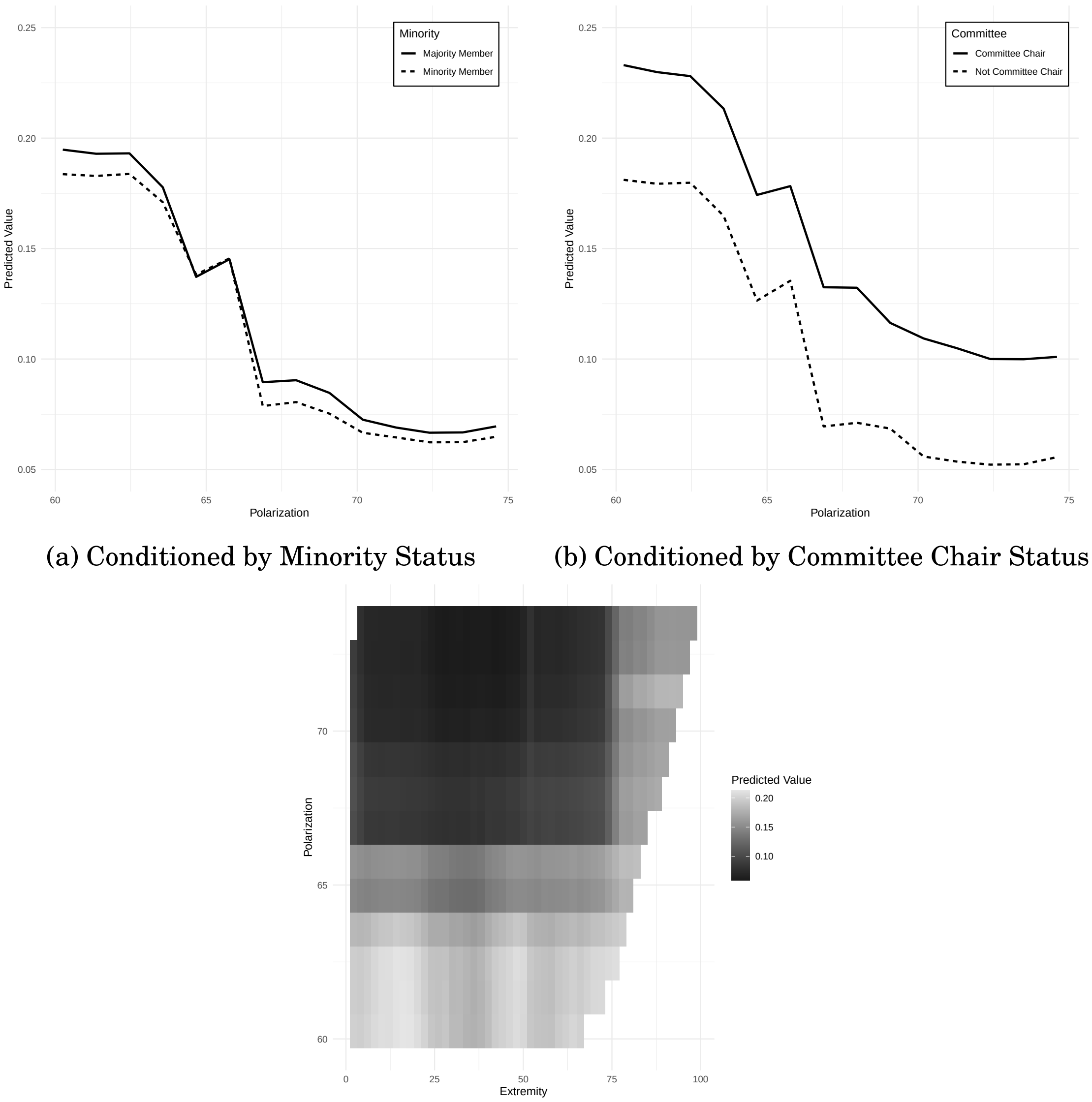


Figure 5: Partial Dependence Plot: Polarization

Next, we explore possible interactions between chamber polarization and HO's three key explanatory variables. While Figure 6a shows that there may be some overlap in how party member status affects likelihood to bypass for the middle range of polarization, we see that increases in chamber polarization lead to less bypassing for both majority- and minority-party members. We find no evidence of an interaction between committee chairmanship and chamber polarization in Figure 6b, but we do see a somewhat large, positive intercept shift for committee chairs.

Finally, we illustrate a strategy for PDPs between continuous variables. Figure 6c that shows an interaction between two



(a) Conditioned by Minority Status

(b) Conditioned by Committee Chair Status

(c) Conditioned by Member Extremity

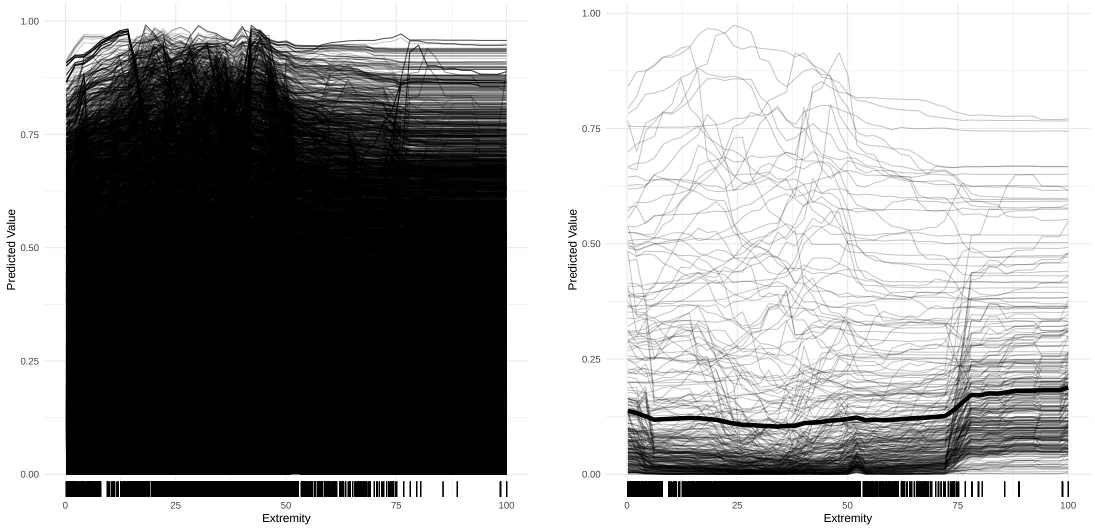
Figure 6: Interactive Partial Dependence Plots: Polarization

continuous variables: ideological extremity and chamber polarization. This time, the area of the plotted space is colored by the predicted value of the dependent variable, similar to a heat map, with lighter shades representing higher predicted values.[12] We find evidence of a conditional relationship in the lower half of the plot. At lower levels of chamber polarization, bill sponsors of across the ideological extremity spectrum become more like to bypass committee. However, as chamber polarization increases, we see that only the most ideologically extreme sponsors remain likely to bypass committee. At high levels of chamber polarization, bill sponsors below the 75th percentile of ideological extremity become much less likely to engage in bypassing. This interesting nuance echoes the increasing importance of ideological members and partisanship in a polarized Congress (e.g. Finocchiaro and Rohde, 2008).

Individual Conditional Expectation Plot

Readers might be skeptical of a single line's ability to average over possible predictions, like the PDP presumes. The ICE plot addresses this skepticism: it shows many of the individual conditional expectations that the PDP is averaging over. It would be impractical to show *every* individual expectation, since we have many observations. This impracticality is illustrated in 7a: the space is so dense that no single line is observable. As a remedy, we randomly sample 500 sets of predictions to show in Figure 7b.

To explore possible heterogeneity, which might be overlooked by PDPs, we show the ICE plot for HO's main finding that ideological extreme sponsors are more likely to bypass legislative committees to introduce a bill. This ICE plot is shown in Figure 7. The average of all of the lines is the thick, dark black line which represents the PDP (as seen in Figure 7b). The thinner lines



(a) All Individual Expectations

(b) Subset of Individual Expectations

Figure 7: Individual Conditional Expectation Plots: Member Extremity

are the individual conditional expectations: the more divergence exists between these and the PDP, the more evidence there is for heterogeneity.

The greatest heterogeneity is observed in the region of Extremity from 30 to 75, or members who are moderately ideological but not extremely so. In this region, the same increase in Extremity would have *different* effects on the probability of bypass, leaving the other characteristics of the model unchanged. When considering moderately ideological members, then, there's evidence that changes in extremity could either increase or decrease the overall probability of bypass. We can compare this to the original PDP in Figure 3a, where increase in extremity are assumed to uniformly increase the probability of bypass. The ICE plot suggests the relationship is more nuanced. This suggestion is further illustrated in 6: changes in the probability of bypass, given an increase in extremity (moving from left to right) for moderate members (from 30 âĂŞ 50 on the x-axis) *increases* (moves from gray to white) in low polarized environments (lower on the y-axis) but *decreases* (moves from gray to black) in more polarized environments (higher on the y-axis). It's for this reason that ICE plots are sometimes said to "reveal interactions and individual differences" (Wright, 2018, p. 8).

Here, it seems like the heterogeneity is concentrated at levels of extremity under 75. There are many observations that diverge considerably, both in functional form and in predicted probability, suggesting that averaging over the expectations in the PDP might not be desirable. This provides additional evidence that the linear functional form assumed by the original model masks considerable non-linearity.

None of this discussion is meant to critique HO's original theory or findings. Instead, we simply seek to illustrate that the machine learning approach suggests some interesting functional forms that may not be well captured by the parametric model. Marrying these machine learning insights with HO's original theory could produce substantively interesting parametric tests of an enriched theoretical story.

Example 2: Poyet and Raunio (2020)

For our second example, we revisit Poyet and Raunio (2020) (henceforth PR), who examine the impact of electoral vulnerability on legislative speechmaking. Their main findings are that as intra-party vulnerability ("1 minus the margin between the number of votes separating the MP and the first nonelected challenger and the total number of votes", p. 13) increases, speeches by members of parliament (MP) decline. Moreover, when intra-party competition is high, the less vulnerable to an intra-party defeat an MP is, and the more speeches the MP will deliver. In contrast, they do not find evidence for two of their other hypotheses: (1) that the higher the inter-party vulnerability (the rank of the election within the party list), the greater number of speeches, and (2) that at high levels of intra-party vulnerability, opposition MPs will deliver more speeches than government MPs. PR find these results using a negative binomial regression on data from the Finnish Parliament between 1995 and 2019. The dependent variable is the number of speeches that an MP delivers over a single term. The model is strictly replicated in Table 2.

Like in Example 1, we use a Random Forest modeling approach to replicate the findings from their negative binomial regression and interpret it using our graphical tools. PR include an interaction between intra-party vulnerability and party score in the district, but, like Example 1, our machine learning approach and graphical interpretation allows us to uncover potential

Table 2: Replicating Poyet and Raunio (2020): Predicting Speeches

	Coefficient	Standard Error
Female	-0.167	(0.043)*
Intra-party Vulnerability	-0.030	(0.283)
Inter-party Vulnerability	0.055	(0.078)
Party Score	0.086	(0.022)*
Intra-party Vulnerability * Party Score	-0.090	(0.023)*
Group Leader	-0.005	(0.091)
Party Leader	0.010	(0.105)
Committee Chair	0.071	(0.064)
Minister	0.240	(0.072)*
Opposition Major	-0.314	(0.051)*
Size of Party	0.002	(0.003)
Seniority	0.014	(0.013)
Exposure	1.025	(0.070)*
(Intercept)	5.438	(0.276)*
Log Likelihood	-7488.31	
N	1228	

Note: Dependent variable is if the number of speeches, replicating Poyet and Raunio (2020), Table 2, Model 3, in the Supplemental Appendix. Two-tailed tests. Includes fixed effects for terms and parties. * $p < 0.05$.

non-linearity and possible interactions without needing to introduce them parametrically.

Variable Importance Plot

We examine VIPs to see how prominent PR's four key independent variables—intra-party vulnerability, inter-party vulnerability, a dichotomous variable for whether the MP is in government, and the share of votes going to an MP's party in their district (their "party score")—are in terms of predictive ability. As shown in Figure 8, all four variables—depicted here by black bars—are important predictors, illustrated by their non-zero reductions in prediction error *and* by that the fact that their inclusion in the model

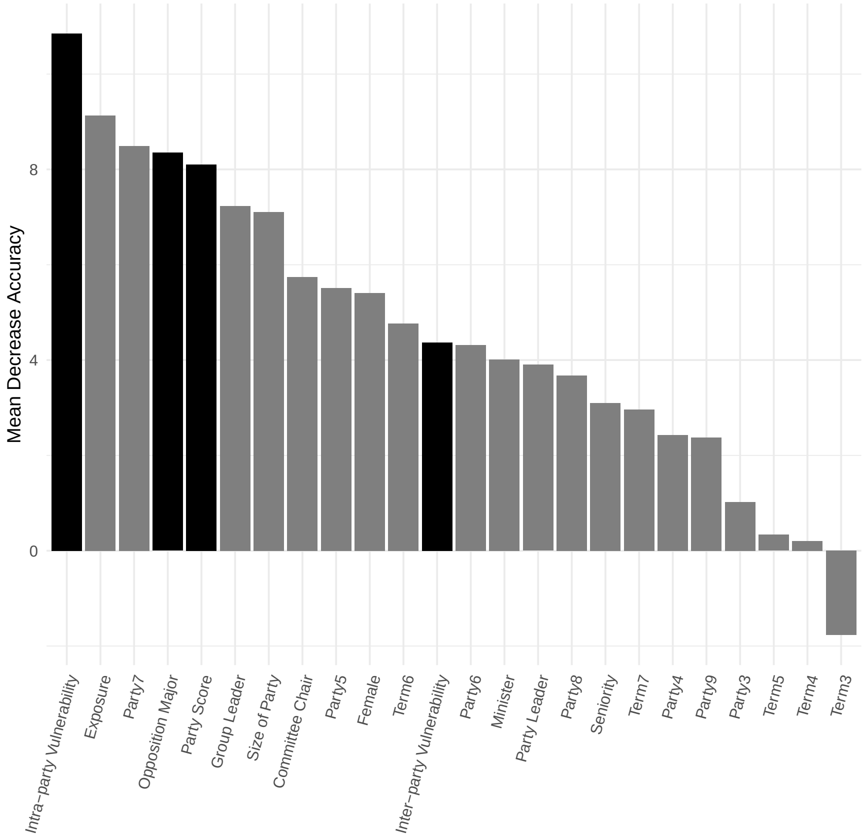


Figure 8: Variable Importance Plot. PR's Key Theoretical Variables are Important Predictors

reduces more prediction error than many of the other predictors. In fact, the most important predictor is intra-party vulnerability. This contrasts with many of the fixed effects PR include, among them legislative terms and party fixed effects. The VIP suggests that PR's key variables go a long way towards explaining differences in speechmaking activity between MPs. It also shows variables which likely should be included in future models of speechmaking due to their relative importance in Figure 8, among them exposure (number of days an MP spent in parliament) as well as whether the MP is a party leader or a minister. To be clear, just because some theoretically-important predictors, such as inter-party vulnerability, are in the middle in terms of relative predictive importance does not mean they should be excluded from the model. The VIP simply indicates that particular theoretical variable is not the most important in predicting the dependent variable.

Partial Dependence Plot

Given that all four of PR's key theoretical variables were some of the most important predictors in their model, we next move to Partial Dependence Plots in order to add nuance to PR's findings. Take for instance their finding of the negative relationship between intra-party vulnerability and speechmaking. As shown in Figure 9, we reach a similar conclusion using a PDP. Recall that the PDP is showing us the predicted value of the number of legislative speeches across the range of intra-party vulnerability, averaging over the effects of all other predictors. We make an additional insight to PR's initial finding in that at very low levels of intra-party vulnerability (intra-party vulnerability ranges from zero to one, with one being the most vulnerable) there appears to be no relationship (or perhaps even a slight positive one) with speechmaking, as shown by the flat PDP line on the left side of Figure 9. We also note a peculiar spike in the number of speeches given when intra-party vulnerability is about 0.8. To account for the seemingly null effect at low levels of intra-party vulnerability within a parametric framework, we could create a knot at 0.35 in intra-party vulnerability to see if there is a difference in the slopes and statistical significance below and above the 0.35 breakpoint.

While we find support for PR's conclusions regarding intra-party vulnerability, we reach a slightly different conclusion regarding their finding of no relationship between inter-party vulnerability

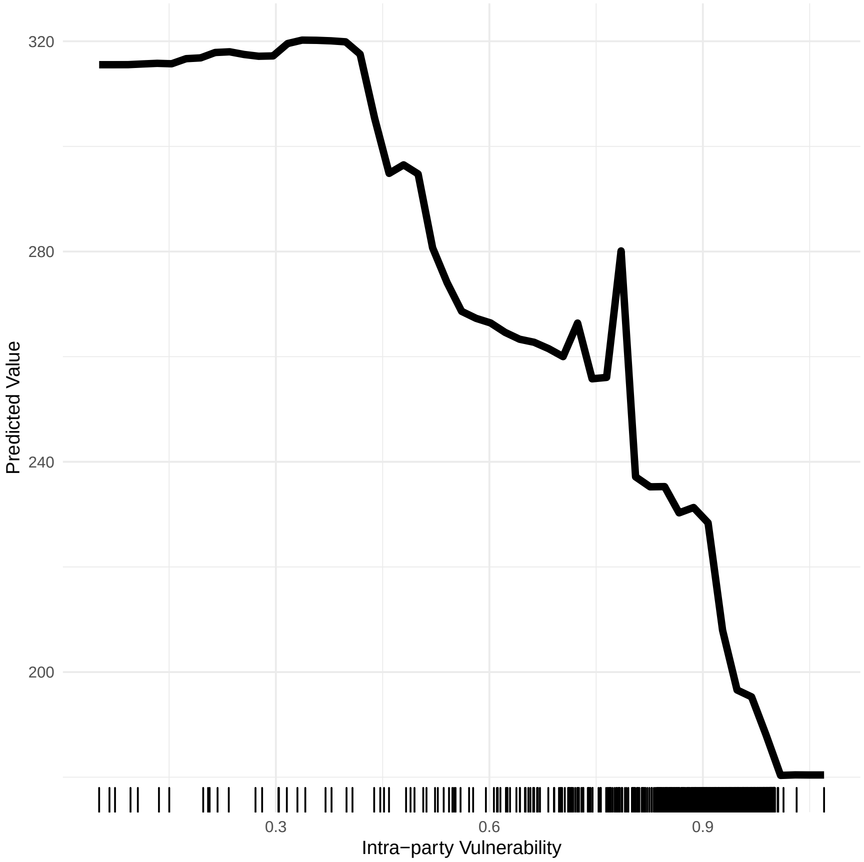


Figure 9: Partial Dependence Plot: Intra-party Vulnerability

and speechmaking. As shown in Figure 10, there does appear to be a negative relationship when inter-party vulnerability lies between one and two. While these effects in Figure 10 are much smaller than those shown in Figure 9, they are still substantial; an MP with an inter-party vulnerability of two makes about 30 less speeches than one with a vulnerability value of one. Therefore, Figure 10 suggests that there might be something different about very low or very high inter-party vulnerability MPs (where there is no relationship with speechmaking) and middle-vulnerability MPs (where there is a strong negative relationship), something which we cannot see using standard estimation techniques. Additionally, we note the considerable non-linearity in the relationship between inter-party vulnerability

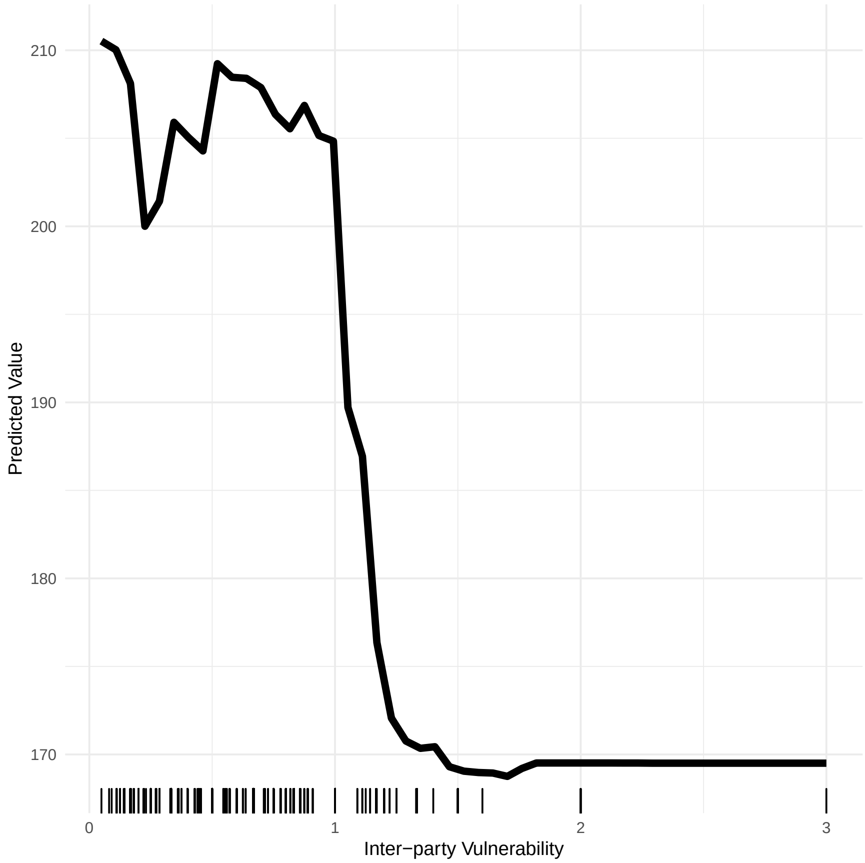


Figure 10: Partial Dependence Plot: Inter-party Vulnerability

and speechmaking. However, a parametric model would restrict this relationship to being linear across its entire range: in essence drawing a straight line from the top-left corner of Figure 10 to the bottom-right corner. At low levels of inter-party vulnerability, the predicted values of speeches fluctuate between a 10-point range. After a score of 1 in inter-party vulnerability, we observe a steep decline in the predicted values of the number speeches, which then levels off around an inter-party vulnerability score of 1.75. These findings suggest that there may be a small interval—between 1 and 1.75 in inter-party vulnerability—when we observe large changes in the predicted values of speeches. While we caution readers and authors not to become overly-reliant on these figures (recall that there are no satisfactory measures of uncertainty when using PDPs), we could further probe this unexplored feature of the authors' original theory by creating a piecewise regression by creating knots at 1 and 1.75 in the inter-party vulnerability variable. We would find support for this result within a parametric framework if the 1-1.75 region of inter-party vulnerability were negative and statistically significant, while the below-1 and above-1.75 regions were not statistically significant, although this latter finding may not be surprising given the paucity of observations above 1.75.

In Figure 11 we plot an interactive PDP between party score (on the vertical axis) and intra-party vulnerability (horizontal). This is a PDP depiction of PR's hypothesis that intra-party

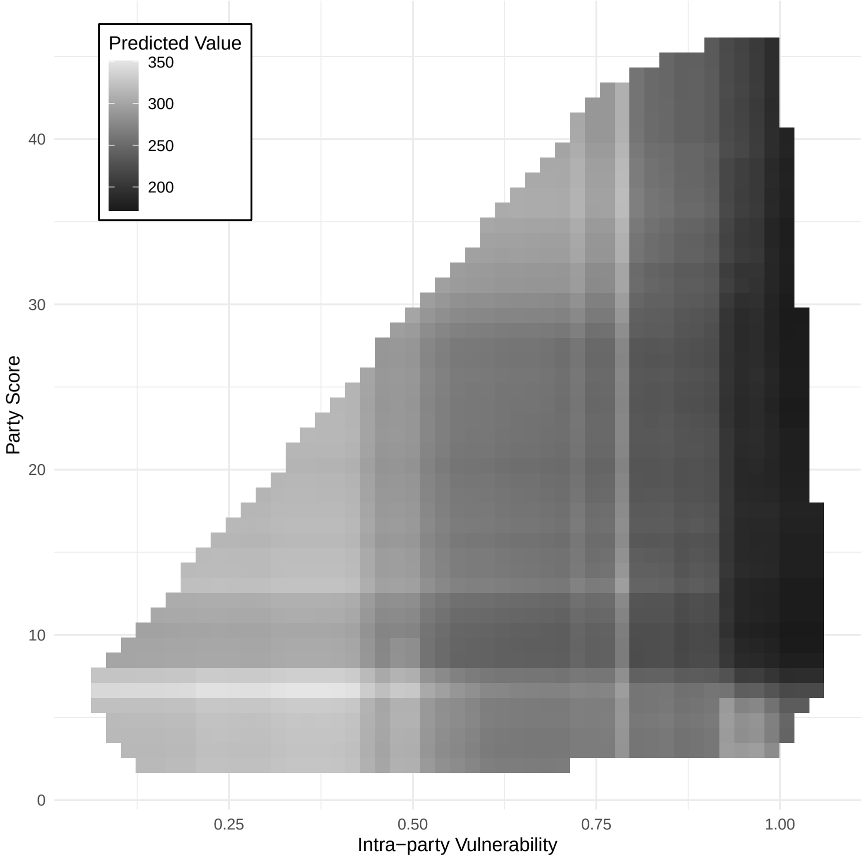


Figure 11: Partial Dependence Plot: Intra-party Vulnerability and Party Score.

competition and the party's electoral district performance are conditional on one another. Our findings align with with PR's in that it is only at very high levels of intra-party vulnerability (between 0.85 and 1) that the predicted speeches decrease rapidly as party score increases. However, this effect appears to be conditional on party score, since the prediction declines from over 300 speeches when the party score is less than 5 to around 200 when the party score is around 10. While this effect can be seen at other levels of intra-party vulnerability in Figure 11, the decline in speeches appears to be far less drastic. For party score, we find that the predicted number of speeches generally decline, regardless of party score, as the level of intra-party vulnerability declines.

Again, though, the virtue of the machine learning and visual interpretation approach is that it allows the analyst to examine interesting non-linear functional forms without having to introduce them specifically in the model. For instance, observe the PDP of party score in the district in Figure 12. As a reminder, PR's initial negative binomial regression found a positive relationship between party score and number of speeches. The PDP, though, suggests this effect might actually be quadratic. When the party score in the district is very low, speechmaking is actually the highest. Then there is a steep decline in speechmaking through the first quartile of party score. At that point, speechmaking begins to rise as party score rises. The machine learning approach allows for us to uncover the initial non-linearities, which we can then test in a more traditional, parametric framework with the inclusion of an interaction term of party score squared with intra-party vulnerability or a linear spline that creates a breakpoint at a party score of 10. Either approach would allow for the formal test of a non-linear hypothesis.

Individual Conditional Expectation Plot

Finally, we illustrate how an ICE plot can be used to verify the averaged relationships observed in the PDP. For instance, if we didn't have a theoretical reason to expect a quadratic relationship as observed in Figure 12, we could check whether the PDP was averaging over individual observations faithfully to the underlying data. Accordingly, we show the ICE plot for party score in Figure 13. Since there are much fewer observations in Example 2,

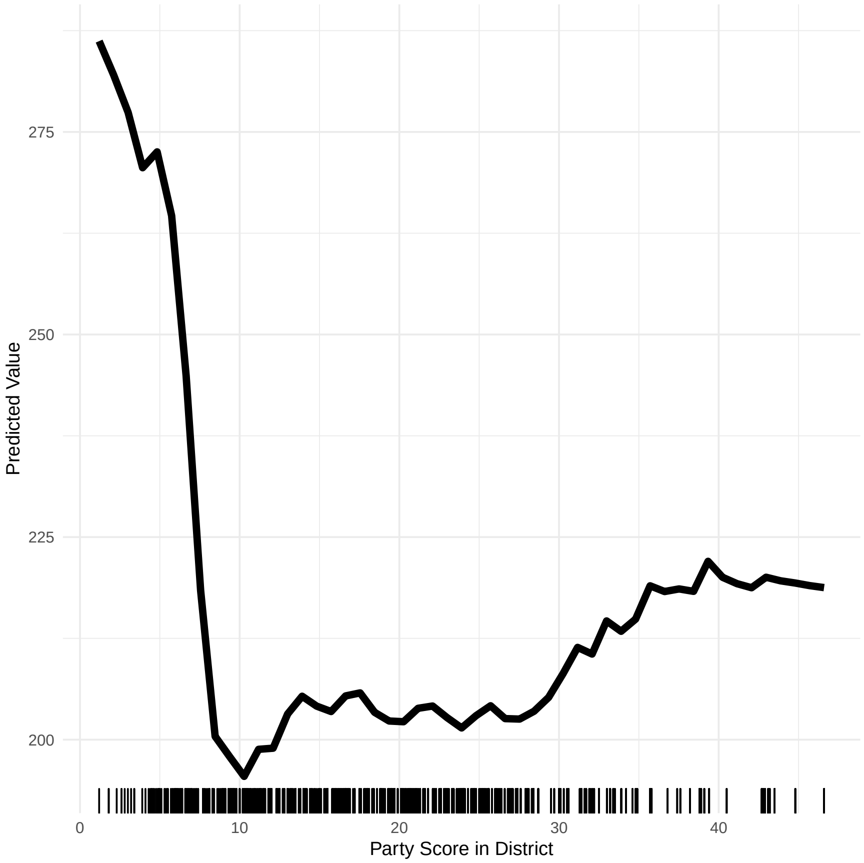


Figure 12: Partial Dependence Plot: Party Score

relative to Example 1, we plot the full set of individual conditional expectations.

Figure 13 shows us similar results to the PDP in Figure 12. Where the individual relationships are the most dense, we can see a distribution of predicted values that follow that single prediction line in the PDP. There is steep decline in the predicted value of number of speeches that occurs until a party score of 10. From a party score of 0 to 10, the densest predictions in Figure 13 fall from approximately 250-300 to 125-200. The greatest amount of heterogeneity—where the individuals lines are less concentrated and similar—occurs on left half of the x-axis, specifically to the left of

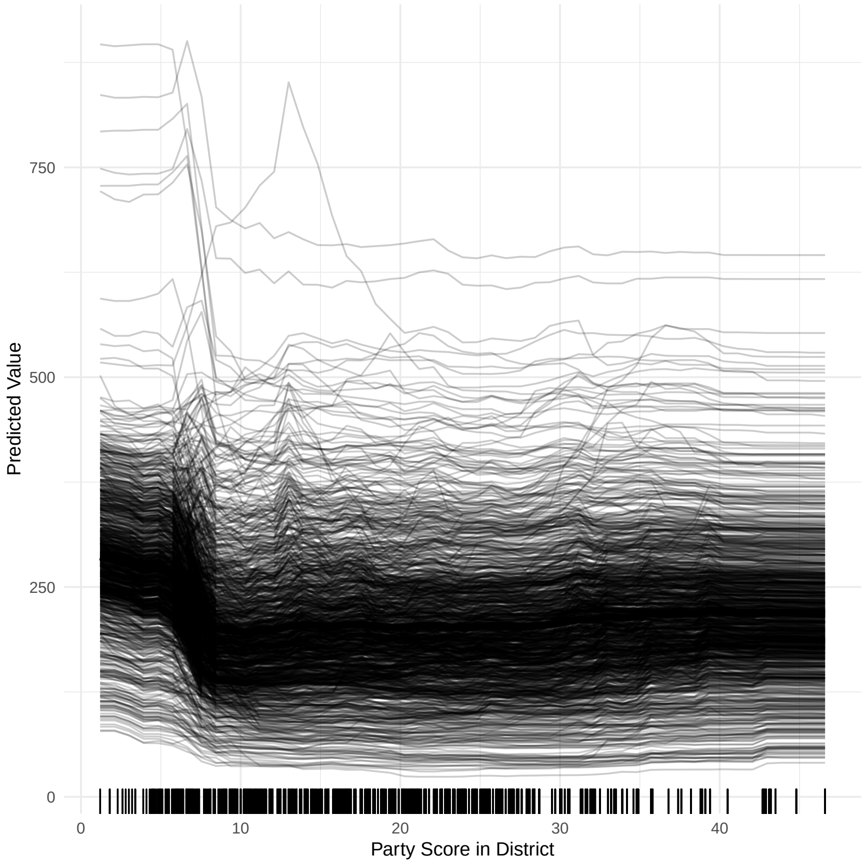


Figure 13: ICE Plot: Party Score.

a party score of 20. Notice too that some of the individual expectations are predicted to be at 700 at low levels of party score, which might drive the averaged PDP much higher in this range. Overall, the pattern looks similar to the PDP, but the individual relationships indicate some heterogeneity around the averaged value in the initial range of party score, where the observed effect is the largest.

Conclusion

While ML models are complicated, they are an accessible tool for verifying the robustness of traditional, parametric models, as demonstrated by their common use in other disciplines. We believe that machine learning models remain a major untapped resource for legislative scholars in particular whose theories often bear the non-linear characteristics that would benefit from more flexible models. We push back against the expectation that machine learning models are too "black-box" to be of any use by showing how researchers can use a suite of plots—among them VIP, PDP and ICE plots—to uncover important covariates and nuanced, potentially non-linear relationships in their data.[13] Using examples from existing legislative research, we provide a template for how scholars can use tree-based models and plots to improve the cogency and robustness of their work.[14]

Our primary advice to political scientists and other interested scholars is to use the machine learning visualization techniques discussed in this study as a robustness check for the findings derived from theoretically-driven parametric models, which constitute the bulk of traditional, statistical inference. While parametric models are relatively simple to interpret for a wide audience, they are constrained in their ability to infer nuanced, non-linear, and even hidden relationships in the data. Users can compare their parametric models—which are highly interpretable yet subject to substantial model assumptions—with their machine learning model to check that the former is properly specified. Using VIPs, users can observe the extent to which their most theoretically interesting variables are the best at reducing prediction error. PDP and ICE plots provide visual insight into the black box of machine learning, providing clear and nuanced depictions of the relationships in the data under analysis. From these visualizations, users can adjust their initial parametric tests to more closely account for nonlinearities and interactions with covariates such as cubic terms or linear splines. After estimating these updated parametric models, researchers can derive measures of uncertainty for these nonlinear functional forms.[15] Or, if the machine learning approach suggests many predictors are important, and a parametric model would be a poor fit, users might consider switching entirely to a machine learning approach that can handle many predictors. In sum, tree-based models, along with VIP, PDP, and ICE plots, allow for a more flexible assessment of the relationships under examination. These tools reveal important insights into potentially uncovered findings.

Soren Jordan is an Associate Professor at Auburn University. He received his Ph.D. in Political Science from Texas A&M University. His research interests are American polarization and quantitative methods.

Hannah L. Paul is an Assistant Professor at the University of Missouri. She received her Ph.D. in Political Science from the University of Colorado, Boulder. Her research interests are American political behavior, representation, and quantitative methods.

Andrew Q. Philips is an Assistant Professor at the University of Colorado, Boulder. He received his Ph.D. in Political Science from Texas A&M University. His research interests are political economy and quantitative methods.

Notes

[1] For instance, no coefficients are estimated nor are hypothesis tests conducted, unlike traditional parametric models like regression. For more on this, see the Online Appendix.

[2] The number of trees, stopping criterion, and number of subset predictors to choose at each node are "hyperparameters" typically chosen through cross-validation in order to avoid any subjective user-specific decisions. The use and optimization of hyperparameters is also detailed in the Online Appendix.

[3] As we note elsewhere, if a predictor of theoretical interest has relatively low predictive importance, researchers should not "throw away" their theory, but rather reconsider and possibly revise their theory.

[4] We describe VIPs for Random Forest models, although similar importance measures exist for other types of tree-based ensemble models, such as Gradient Boosting Machines.

[5] This measure is often scaled by the standard deviation of these differences.

[6] As an alternative, users might choose to show the average total decrease in node impurities (using the residual sum of squares for a continuous dependent variable or the Gini index for categorical) after splitting, whereby variables can be considered important if, when they are used at a node split, they do a good job in partitioning the data into correctly classified groups (categorical dependent variable) or improve predictive accuracy (continuous dependent variable) (Hastie, Tibshirani, and Friedman, 2013).

[7] Users should only include *theoretically* relevant variables in their VIP to begin with in order to avoid data mining. The benefit of a VIP is that users are not restricted in their selection of the number of theoretically interesting covariates.

[8] There is no consensus on whether emulating "confidence intervals" in this way is desirable; we elaborate on this point in the Online Appendix.

[9] Moreover, it is common to distinguish the actual observed value of x_s for each individual i, typically by using a dot or some other marker.

[10] The authors also model the method by which the bill bypasses committee proceedings (unanimious consent versus the Rule XIV procedure) with a multinomial logistic regression, but we focus on the binary bypass indicator.

[11] We include HO's Congress and policy area fixed effects as a series of dummy variables: the solution the most closely emulates the authors' original modeling choice. However, fixed effects in machine learning models might take other forms: for a full discussion, see the Online Appendix.

[12] The bottom right of the plot is not predicted as it is not observed: there are no observations with extreme members but low polarization, so the model is unable to classify these cases.

[13] Of course, there are other additional graphical strategies one might consider that we did not cover here, such as local interpretable model-agnostic explanations (LIME) (Molnar, 2020).

[14] In the Online Appendix, we step through model estimation in exacting detail.

[15] For an example of how to implement this parametric-machine learning-parametric estimation process, see (Funk, Paul, and Philips, 2020).

References

Anastasopoulos, L Jason, and Anthony M Bertelli. 2020. “Understanding delegation through machine learning: A method and application to the European Union.” *American Political Science Review* 114 (1): 291–301.

Anderson, William D, Janet M Box-Steffensmeier, and Valeria Sinclair-Chapman. 2003. “The keys to legislative success in the US House of Representatives.” *Legislative Studies Quarterly* 28 (3): 357–386.

Basinger, Scott J., and Maxwell Mak. 2020. “The ‘New Normal’ in Supreme Court Confirmation Voting: Hyper-Partisanship in the Trump Era.” *Congress & the Presidency* 47 (3): 365–386.

Baumann, Markus, Marc Debus, and Jochen Müller. 2015. “Personal characteristics of MPs and legislative behavior in moral policymaking.” *Legislative Studies Quarterly* 40 (2): 179–210.

Bendix, William, and Jon MacKay. 2017. “Partisan Infighting Among House Republicans: Leaders, Factions, and Networks of Interests.” *Legislative Studies Quarterly* 42 (4): 549–577.

Benoit, Kenneth, Kevin Munger, and Arthur Spirling. 2019. “Measuring and explaining political sophistication through textual complexity.” *American Journal of Political Science* 63 (2): 491–508.

Bonica, Adam. 2018. “Inferring Roll-Call Scores from Campaign Contributions Using Supervised Machine Learning.” *American Journal of Political Science* 62 (4): 830–848.

Bonica, Adam. 2020. “Why Are There So Many Lawyers in Congress?” *Legislative Studies Quarterly* 45: 253–289.

Bonvecchi, Alejandro, Ernesto Calvo, and Ernesto Stein. 2016. “Legislative Knowledge Networks, Status Quo Complexity, and the Approval of Law Initiatives.” *Legislative Studies Quarterly* 41 (1): 89–117.

Bowler, Shaun, Gail McElroy, and Stefan Muller. 2020. “Campaigns and the Selection of PolicyâĂŘSeeking Representatives.” *Legislative Studies Quarterly* 45: 397–431.

Breiman, Leo. 2001. “Random forests.” *Machine learning* 45 (1): 5–32.

Crosson, Jesse M, Alexander C Furnas, Timothy Lapira, and Casey Burgat. 2019. “Partisan competition and the decline in legislative capacity among congressional offices.” *Legislative Studies Quarterly* .

Curry, James M. 2019. "Knowledge, Expertise, and Committee Power in the Contemporary Congress." *Legislative Studies Quarterly* 44 (2): 203–237.

Finocchiaro, Charles J., and David W. Rohde. 2008. "War for the Floor: Partisan Theory and Agenda Control in the U.S. House of Representatives." *Legislative Studies Quarterly* XXXIII (1): 35–61.

Friedman, Jerome H. 2001. "Greedy function approximation: a gradient boosting machine." *Annals of statistics* pp. 1189–1232.

Funk, Kendall D., Hannah L. Paul, and Andrew Q. Philips. 2020. Point break: Using machine learning to uncover a critical mass in women's representation. Working paper.

Funk, Kendall D, Hannah L Paul, and Andrew Q Philips. 2021. "Point break: using machine learning to uncover a critical mass in women's representation." *Political Science Research and Methods* pp. 1–19.

Goet, Niels D, Thomas G Fleming, and Radoslaw Zubek. 2020. "Procedural Change in the UK House of Commons, 1811–2015." *Legislative Studies Quarterly* 45 (1): 35–67.

Goldstein, Alex, Adam Kapelner, Justin Bleich, and Emil Pitkin. 2015. "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation." *Journal of Computational and Graphical Statistics* 24 (1): 44–65.

Green, Donald P, and Holger L Kern. 2012. "Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees." *Public opinion quarterly* 76 (3): 491–511.

Greenwell, Brandon M. 2017. "pdp: an R Package for constructing partial dependence plots." *The R Journal* 9 (1): 421–436.

Hainmueller, Jens, Jonathan Mummolo, and Yiqing Xu. 2019. "How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice." *Political Analysis* 27 (2): 163–192.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2013. *The elements of statistical learning: Data mining, inference, and prediction*. Second ed. Springer Science & Business Media.

Howard, Nicholas O., and Mark E. Owens. 2020. "Circumventing Legislative Committees: The US Senate." *Legislative Studies Quarterly* 45: 495–526.

Kastellac, Jonathan P., and Eduardo L. Leoni. 2007. "Using Graphs Instead of Tables in Political Science." *Perspectives on Politics* 5: 755–771.

Kaufman, Aaron Russell, Peter Kraft, and Maya Sen. 2019. "Improving Supreme Court Forecasting Using Boosted Decision Trees." *Political Analysis* 27 (3): 381–387.

Kim, Seo-young Silvia, R Michael Alvarez, and Christina M Ramirez. 2020. “Who Voted in 2016? Using Fuzzy Forests to Understand Voter Turnout.” *Social Science Quarterly* 101 (2): 978–988.

McKay, Amy Melissa. 2020. “Buying Amendments? Lobbyists’ Campaign Contributions and Microlegislation in the Creation of the Affordable Care Act.” *Legislative Studies Quarterly* 45 (2): 327–360.

Metz, Thomas, and Sebastian Jäckle. 2016. “Hierarchical, decentralized, or something else? Opposition networks in the German bundestag.” *Legislative Studies Quarterly* 41 (2): 501–542.

Molnar, Christoph. 2020. *Interpretable Machine Learning*. Lulu. com.

Montgomery, Jacob M, and Santiago Olivella. 2018. “Tree-Based Models for Political Science Data.” *American Journal of Political Science* 62 (3): 729–744.

Muchlinski, David, David Siroky, Jingrui He, and Matthew Kocher. 2016. “Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data.” *Political Analysis* 24 (1): 87–103.

Osborn, Tracy, Rebecca J Kreitzer, Emily U Schilling, and Jennifer Hayes Clark. 2019. “Ideology and Polarization Among Women State Legislators.” *Legislative Studies Quarterly* 44 (4): 647–680.

Poyet, Corentin, and Tapio Raunio. 2020. “Reconsidering the Electoral Connection of Speeches: The Impact of Electoral Vulnerability on Legislative Speechmaking in a Preferential Voting System.” *Legislative Studies Quarterly* .

Proksch, Sven-Oliver, Will Lowe, Jens Wäckerle, and Stuart Soroka. 2019. “Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches.” *Legislative Studies Quarterly* 44 (1): 97–131.

Streeter, Shea. 2019. “Lethal force in black and white: Assessing racial disparities in the circumstances of police killings.” *The Journal of Politics* 81 (3): 1124–1132.

Suzuki, Akisato. 2015. “Is more better or worse? New empirics on nuclear proliferation and interstate conflict by random forests.” *Research & Politics* 2 (2): 2053168015589625.

Wright, Ray. 2018. “Interpreting Black-Box Machine Learning Models Using Partial Dependence and Individual Conditional Expectation Plots.” *SAS1950-2018* .