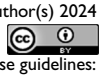
Original Article

Sociological Methodology
2024, Vol. 54(2) 217–268


Article reuse guidelines:
DOI: 10.1177/00811750231217734
http://sm.sagepub.com

Incorporating Machine Learning into Sociological Model-Building

Mark D. Verhagen[1,2,3]

Abstract

Quantitative sociologists frequently use simple linear functional forms to estimate associations among variables. However, there is little guidance on whether such simple functional forms correctly reflect the underlying data-generating process. Incorrect model specification can lead to misspecification bias, and a lack of scrutiny of functional forms fosters interference of researcher degrees of freedom in sociological work. In this article, I propose a framework that uses flexible machine learning (ML) methods to provide an indication of the fit potential in a dataset containing the exact same covariates as a researcher's hypothesized model. When this ML-based fit potential strongly outperforms the researcher's self-hypothesized functional form, it implies a lack of complexity in the latter. Advances in the field of explainable AI, like the increasingly popular Shapley values, can be used to generate understanding into the ML model such that the researcher's original functional form can be improved accordingly. The proposed framework aims to use ML beyond solely predictive questions, helping sociologists exploit the potential of ML to identify intricate patterns in data to specify better-fitting, interpretable models. I illustrate the proposed framework using a simulation and real-world examples.

Keywords

Machine learning, Misspecification, Explainable A.I., Computational methods

It is common knowledge that valid inference crucially depends on a correctly specified relationship between the outcome of interest, y, and the explanatory variables, X (Buja, Brown, et al. 2019; Cameron and Trivedi 2005; Long and Trivedi 1992). In practice, much more attention is typically paid to identifying relevant variables to include in a model rather than to making sure the functional relationship among these variables is correctly specified. This is evidenced by the fact that variables are often simply assumed to affect the outcome in a linear and additive way (Hindman 2015). However, there is little reason to believe linear additive models appropriately reflect the underlying data-generating process (DGP). At the same time, estimating incorrectly specified models can lead to biased findings; there are noteworthy examples throughout the social sciences—and likely many more that have gone unnoticed—where more complicated functional relationships, which might include nonlinearities

[1]Leverhulme Centre for Demographic Science, Oxford, UK
[2]Nuffield College, University of Oxford, Oxford, UK
[3]Department of Sociology, University of Oxford, Oxford, UK

Corresponding Author:
Mark D. Verhagen, University of Oxford, New Rd, Oxford, OX1 2JD, UK.
Email: mark.verhagen@nuffield.ox.ac.uk

or interactions, have led to reversed findings (Christensen and Christensen 2014; Dougherty et al. 2015; Freedman 2009; Heckman, Humphries, and Veramendi 2018; McClintock 2017; Muñoz and Young 2018). Despite this type of criticism of the standard linear additive model having been around for many decades, it remains the workhorse throughout most empirical sociology today (Abbott 1988; Berk 2004; Duncan 1984; Lundberg, Johnson, and Stewart 2021). In this article, I incorporate methods from machine learning (ML) and explainable A.I. (X-AI) into the standard empirical workflow to help sociologists (1) assess whether their hypothesized model fits the data well by comparing its fit against a flexible ML model, and (2) improve their model when it does not accurately represent the patterns in the data by unpacking the ML model using X-AI techniques.

In the past, simple models like the linear additive functional form were often a necessity due to computational constraints.[1] Yet these historical limitations are no longer a factor, giving rise to ML methods that instead exploit the exponential increase in available computational power. These methods let the data dictate the relationship among variables, rather than relying on the researcher to hypothesize the functional form between the two.[2] Often, this flexibility leads to better-fitting models that improve on researcher-hypothesized models in fitting the data (Brand et al. 2021; Grimmer, Roberts, and Stewart 2021); this has led to increased adoption of ML throughout industry and academia (Athey 2018; Rahal, Verhagen, and Kirk 2022).[3] However, ML methods are almost exclusively applied within a predictive context due to their "black box" nature (Mullainathan and Spiess 2017; Shmueli 2010); their inclusion into sociological inquiry, which has an overwhelmingly explanatory focus, remains limited.[4] This is unnecessary, as the fact that ML methods can identify associations among variables holds value for explanatory work as well.

Concretely, I propose to incorporate ML methods into the typical quantitative empirical workflow in the following way. Assume a researcher is interested in modeling the association between some interest variable and an outcome. To this effect, they apply a conditioning-on-observables approach to control for confounders and develop a hypothesized functional form, $\tilde{f}(\cdot)$. However, it is unclear a priori whether this model correctly represents the underlying relationship among variables in the data. In the first step of the proposed approach, $\tilde{f}(\cdot)$ is estimated and, in parallel, a flexible ML model, or preferably an ensemble method like a Super Learner that combines multiple ML methods, is estimated to the same data (Baćak and Kennedy 2019). The flexible model thus uses the same set of variables as identified by the researcher in their hypothesized model and follows the same inferential logic.[5] However, the functional form in which variables relate to one another is left to be decided by the data, effectively considering a much wider range of possible model specifications than a researcher would typically do during model-building. Second, the fit of both the hypothesized model and the flexible ML model is evaluated out-of-sample, either through cross-validation or a formal holdout set, to identify a possible difference in fit between the two models. Third, in case a lack of appropriate specification in $\tilde{f}(\cdot)$ is found, the flexible model is unpacked to better understand the lack of specification in $\tilde{f}(\cdot)$, such that it can be improved.

The framework relies on a number of concepts from ML or closely adjacent fields. First, that out-of-sample estimates of model fit can be used to compare the ability of widely varying modeling approaches to fit the data (Rose 2013; Stone 1977; Verhagen 2022). Second, that ML methods can be used as efficient function approximators that can uncover relevant patterns in data (Baćak and Kennedy 2019; Van der Laan, Polley, and Hubbard 2007). Third, that methods from the rapidly evolving field of X-AI—like the Shapley values approach explored in this article that decomposes model predictions along covariates—allow researchers to distill understanding from ML methods, which can then be used to improve hypothesized models (Samek et al. 2019). These elements, in particular the last, allow for an iterative process where ML is used in a supporting role in model-building. The novelty of the framework thus lies in using ML methods as a guide, in such a way that the overarching goal of the framework is still to generate interpretable models (Agrawal, Peterson, and Griffiths 2020; Rudin 2019). This complementary role should be juxtaposed with the overwhelmingly predictive focus of ML methods in sociology up until now (Molina and Garip 2019). The framework is also in line with increasing calls to use ML in a supportive role throughout the empirical pipeline, rather than completely replacing interpretable methods for ML equivalents (Agrawal et al. 2020; Rudin 2019; Rudin et al. 2010).[6,7]

Incorporating ML and X-AI methods into empirical work should not only improve model-building but also improve transparency into the research process. A lack of emphasis on functional form specification provides researchers with relative freedom in evaluating multiple functional forms and choosing which one to report (Muñoz and Young 2018; Sala-i Martin 1997; Simonsohn, Simmons, and Nelson 2020). Such "researcher degrees of freedom" are to blame for a large number of important empirical findings turning out to be un-reproducible and a more general "crisis in science," where results are often dependent on over-engineered models (Gelman and Loken 2013; Ioannidis 2005; Young 2018). Comparing a researcher's carefully tailored model to the fit of an ML method estimated to the same data can help identify such an over-engineered model, by providing insights into model fit that are not a result of a researcher's decision-making process but rather a feature of the data. More generally, overfitting is a central concern throughout ML, and the frequent re-estimating of models on subsets—standard practice throughout ML—can further protect against "*p*-hacking" (Gelman and Loken 2013).

More generally, the proposed framework provides a welcome realignment of empirical methods with today's computational reality (Efron and Hastie 2016) and is in line with increasing calls for a re-appreciation of the predictive power of social theories that can be observed more broadly throughout sociology (Hofman, Sharma, and Watts 2017; Verhagen 2022; Watts 2014, 2017).

The rest of this article is structured as follows. I first introduce a toy example to illustrate the risk of model misspecification to inference. Then, I introduce the three steps of the proposed framework and illustrate each by applying them to the same toy example. I next compare the proposed approach to other computational frameworks. In the empirical section, I apply the framework to one simulated and two real-world case studies. In doing so, I illustrate how the framework identifies (1) missing

nonlinearities and interactions in a simulated dataset on the association between schooling and earnings, (2) nonlinearities and spatial heterogeneity in a large dataset on London house sales, and (3) various intricate patterns in voting preferences among U.S. voters in the General Social Survey (GSS).

MISSPECIFICATION AND RISKS TO INFERENCE

Consider the following toy example. Our interest lies in whether having grandchildren affects elderly individuals' worries about climate change. We analyze a survey ($N = 900$) in which respondents were asked their level of concern with climate change on a seven-point scale. In addition, we have information on respondents' age, sex, whether they completed high school, and, our interest variable, whether they have grandchildren. We might propose the following model (Model 1) within a conditioning-on-observables framework:

$$y_i = \beta_0 + \beta_1 x_{\text{age},i} + \beta_2 x_{\text{sex, i}} + \beta_3 x_{\text{high\_school},i} + \beta_4 x_{\text{grandchildren, i}} + \epsilon_i, \quad (1)$$

which simply plugs all the relevant controls and our interest variable into a linear additive model. We might also consider including a nonlinearity in the effect of age by adding a square (Model 2):

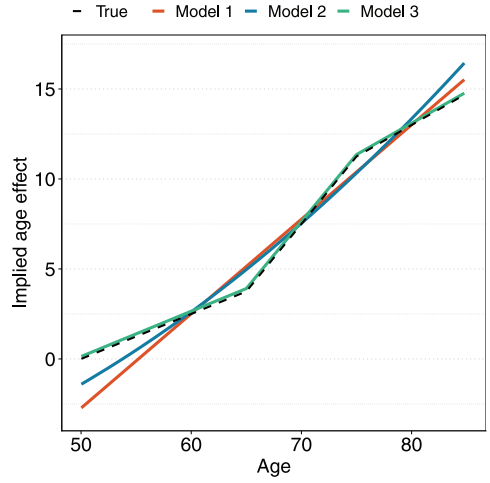
$$y_i = \beta_0 + \beta_1 x_{\text{age},i} + \beta_2 x_{\text{sex, i}} + \beta_3 x_{\text{high\_school},i} + \beta_4 x_{\text{grandchildren},i} + \beta_5 x^2_{\text{age},i} + \epsilon_i. \quad (2)$$

Results for estimating both models on a simulated dataset are presented in the first two columns in the left-hand panel of Table 1. We find a substantial difference between the two models in terms of the estimated association between having grandchildren and concerns with climate change. The effect is statistically insignificant for Model 1 with a point estimate of -0.02, whereas it is negative and significant ($p<0.01$) for Model 2 with a point estimate of -0.68. A Likelihood Ratio (LR) test strongly prefers the second model over the first ($p<0.001$). We might feel comfortable concluding our empirical analysis at this stage.

Imagine, however, that the true effect of age is indeed nonlinear, but piece-wise linear rather than a second-degree polynomial. Specifically, there is a stronger increase in concern by age among people age 65 to 75, with more modest increases at other ages. If we had estimated a functional form in line with this nonlinearity, we would find results as reported in the third column of the left panel of Table 1. These results show that the effect of having grandchildren is in fact associated with a higher level of concern. The estimated age effect per functional form is illustrated in the right panel of Table 1, with the dashed line representing the "true" effect in the DGP and the colored lines the best-fitting effect given the flexibility of the model. The problem is that both the linear and the quadratic function of age overshoots the true effect as the respondent's age increases. Coupled with the fact that grandchildren are more prevalent among the older population, this effect leads to incorrect inference.

Table 1. Table on the left shows regression results when estimating a functional form assuming linear age, quadratic age, and a step function for age; Plot on the right shows implied effect of age for the three model specifications.

	Model 1	Model 2	Model 3
(Intercept)	−19.02***	−2.16	9.97***
	(0.51)	(3.42)	(0.23)
Age: linear	0.52***	0.03	
	(0.01)	(0.10)	
Grandchildren	−0.02	−0.68**	0.57*
	(0.21)	(0.25)	(0.23)
Sex: female	−0.97***	−0.99***	−0.95***
	(0.10)	(0.10)	(0.08)
High school	0.41***	0.38***	0.36***
	(0.10)	(0.10)	(0.08)
Age: squared		0.00***	
		(0.00)	
Age: 50 +			0.26***
			(0.02)
Age: 65 +			0.49***
			(0.03)
Age: 75 +			−0.41***
			(0.04)
Age: 85 +			−0.17
			(0.13)
R^2	0.88	0.89	0.91
Adj. R^2	0.88	0.89	0.91
Num. obs.	900	900	900
RMSE	1.44	1.42	1.23



Note: Table shows OLS regression coefficients. Standard errors are in parentheses. The outcome variable of interest is a seven-point scale indicating respondent's concerns with climate change (7 = highest, 1 = lowest).
*$p < 0.05$. **$p < 0.01$. ***$p < 0.001$.

The issue with doing inference into the first two models is that the assumption of exogeneity, that is, $E(\epsilon|X) = 0$, is violated. The error term for the first two models will be small for observations around ages 60, 70, and 80, as the estimated effect lies close to the true effect, but it will be larger at the ends of the age distribution. The coefficient estimates still reflect the best-fitting model given the functional form, but clearly lead to incorrect inference. The more general phenomenon of exogeneity is further illustrated by the four bivariate relationships plotted in Figure 1.[8] Three of the four models include a nonlinear relationship, although a linear coefficient estimated to the data ($N = 80$) is statistically significant at the conventional 5 percent level. These examples illustrate two typical risks of misspecification: that we wrongfully assume an effect to be linear when it is not, and that correlation of our interest variable with a misspecified control can lead to omitted variable bias (Cameron and Trivedi 2005; Long and Trivedi 1992).[9]

Various statistical tests have been proposed to combat misspecification. The most popular ones are White's test for functional misspecification (White 1980, 1981) and Ramsey's RESET test (Ramsey 1969). White's test statistic is based on a comparison of the estimated coefficients from a hypothesized model $\hat{\beta}$ with those from a weighted regression $\hat{\beta}_{\text{WLS}}$, whereas the Ramsey test statistic compares the hypothesized model with a model that includes higher-order versions of the explanatory variables and

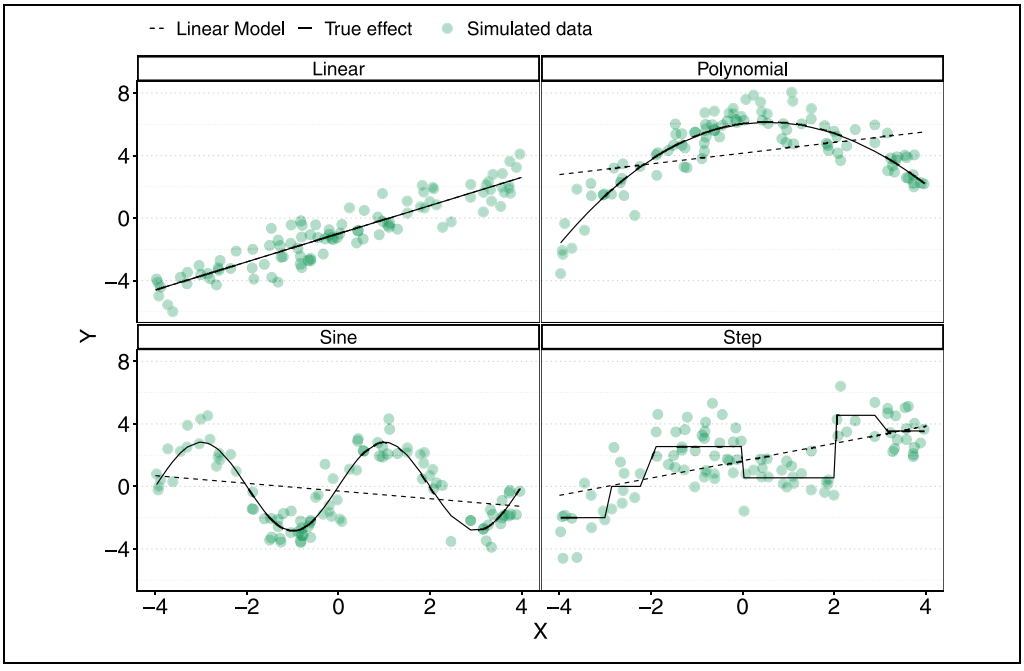


Figure 1. Implied association from a linear model (dashed line) relative to the true effect (solid line).

Note: The linear models are estimated based on 80 uniformly distributed observations for the X variable, plugged into one of four functional forms (see the Appendix) with an additive normally distributed error. Data points are shown in green.

evaluates their statistical significance. Both provide computationally efficient statistics to assess misspecification of the functional form, although both assume the true model lies within a specific class of functions $\mathcal{M}$ with possibly restrictive assumptions (Golden et al. 2016).[10] Since the introduction of the White and Ramsey tests, more computationally intensive (semi-)parametric methods have been developed (Robinson 1988a, 1988b; Yatchew 1997), and substantial follow-up research has built on their principles (Golden et al. 2016). Unfortunately, misspecification tests have a number of well-known theoretical and practical problems (Buja, Brown, et al. 2019; Long and Trivedi 1992).

First, they are generally under-powered and can struggle to identify misspecification in multivariate settings (Buja, Kuchibhotla, et al. 2019:616).[11] Second, they provide limited insight into what to do next, if a test is rejected. Third, and perhaps most important, the actual implementation of misspecification tests is limited in published empirical work (Long and Trivedi 1992; Open Science Collaboration 2015). As a case in point, the two flagship journals in sociology—the *American Sociological Review* and *American Journal of Sociology*—totalled 70 research articles in 2022. Of these, 40 included quantitative analyses, of which 32 implemented a conditioning-on-observables approach. None of these articles implement any of the standard misspecification tests like the White or RESET tests.[12] In practice, researchers enjoy relative freedom in specifying their functional form, as well as the robustness and specification

tests they use to verify its appropriateness. The crisis in reproducibility and the effect of researcher degrees of freedom on empirical results thus extends to the (lack of) specification checks chosen by researchers to validate their model assumptions (Simonsohn et al. 2020; Young 2018).

Even without malicious intent, many functional forms are from an outdated age of computational limitations (Buja, Kuchibhotla, et al. 2019:615; Efron and Hastie 2016; Muñoz and Young 2018). The linear additive functional forms plugged into exponential family probability distributions were historically preferred due to their ease of estimation and interpretation, not their de facto appropriateness to study social life. This pragmatic preference for parsimony has led to the continued prevalence of simplistic functional forms, even though the computational constraints under which they were developed are no longer present, and the wealth of qualitative research in the social sciences consistently implies that social life is in fact highly complex and probably not appropriately modeled using such simple functional forms (Abbott 1988). As the toy example above illustrates, inference can easily go astray given the often blind acceptance of simplistic functional forms in empirical work.

A COMPUTATIONAL FRAMEWORK TO IMPROVE MODEL-BUILDING

I propose a computational framework that exploits ML methods to obtain an indication of the potential model fit in a dataset. This potential can then be compared to the fit of a researcher's own hypothesized model, and thus used to diagnose a possible lack of specification in the latter. This estimate provides researchers with an intuitive assessment of how well the data *could* be modeled versus how well the data *is* modeled by their hypothesized model. ML thus takes a guiding role in model-building. Whenever it is found that the ML model improves on the researcher-hypothesized model, methods from the X-AI domain can be implemented to unpack and better understand the ML model and subsequently improve the hypothesized model.

Given a dataset $\mathcal{D}$ and a researcher-hypothesized functional form $\tilde{f}(\cdot)$, relating outcome of interest y with independent variables X, the proposed framework consists of the following three simple steps (see Figure 2 for a schematic illustration):

1. Estimate hypothesized model $\tilde{f}(\cdot)$ to the data, as well as an ML model $\bar{f}(\cdot)$.
2. Evaluate the model fit of both $\tilde{f}(\cdot)$ and $\bar{f}(\cdot)$ and assess a possible lack of fit in $\tilde{f}(\cdot)$.
3. Diagnose why $\bar{f}(\cdot)$ improves on the hypothesized model and improve $\tilde{f}(\cdot)$ accordingly.

Crucial to the framework is that the ML method is estimated using the same variables as present in the model originally hypothesized by the researcher. The framework is not designed for data-mining, where a large number of possible explanatory variables are included in the model. In such a case, one risks including variables that might improve model fit but could harm understanding of the underlying processes—typical

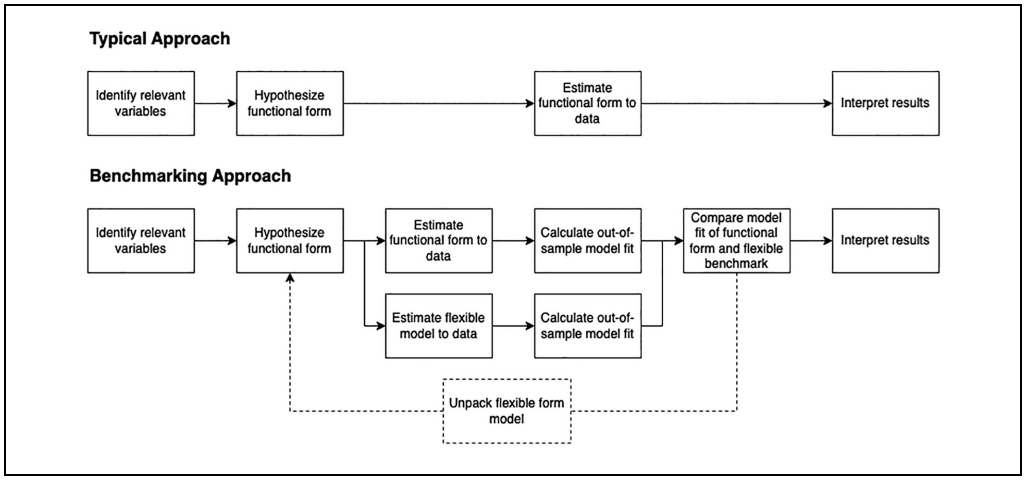


Figure 2. Schematic illustration of the proposed framework.

examples would be (accidentally) including post-treatment variables or colliders. Instead, the functional form is scrutinized given the exact same inferential curation of variables and logic as present in the researcher's originally hypothesized model.[13] The following sections describe the three components of the proposed framework in more detail and illustrate them using the toy example.

Step I: The benchmarking model

The first component of the proposed framework is a flexible form model $\bar{f}(\cdot)$, which serves as a benchmark of the possible model fit in the data. This model uses the same variables and thus follows the same inferential logic as the hypothesized model $\tilde{f}(\cdot)$, but it evaluates patterns not necessarily considered by the researcher's functional form. The ML methods do not rely on prespecified functional forms, but distill the functional form from the observed data (Baćak and Kennedy 2019; Grimmer et al. 2021). However, they retain more structure than do non-parametric approaches like local regression and often suffer less from the "curse of dimensionality," where the number of data points required scales exponentially with the covariate space (Bishop 2006).

Many different ML methods can be applied to datasets commonly encountered in the social sciences (Athey 2018; Hastie, Tibshirani, and Friedman 2009). In fact, limiting the benchmarking step to a single researcher-curated ML model to serve as $\bar{f}(\cdot)$ would invite some of the very risks this framework is designed to address in terms of researcher degrees of freedom in model-building. Tuning parameters in ML models can easily be calibrated to diminish model fit and thus imply appropriate fit of $\tilde{f}(\cdot)$, when this is not the case. Relatedly, some methods fit certain patterns in data better than others with often limited a priori guidance (Berk and Bleich 2013; Rose 2013). A principled approach would thus estimate not a single but an ensemble of flexible form models during the benchmarking phase (Baćak and Kennedy 2019).

The Super Learner is an example of such an ensemble method, consisting of multiple ML methods or the same method with different parameter settings. Among these, the most appropriate model is identified using cross-validation. The oracle result by Van Der Laan and Dudoit (2003) shows that a Super Learner is indeed optimal to identify the best-fitting model to approximate an unknown DGP among evaluated models, and the price of including large numbers of models into the Super Learner is minor in terms of performance (Baćak and Kennedy 2019; Van der Laan et al. 2007; Van der Vaart, Dudoit, and van der Laan 2006). Therefore, many models can and should be evaluated, and ensembles of various methods are typically preferred as function approximators over single models (Agrawal et al. 2020).[14] A Super Learner can also be specified prior to model estimation and preregistered, improving transparency in the research process (Open Science Collaboration 2015). Naturally, the full set of models and the resulting fit of each should be part of the research output.[15]

Among the broader set of ML methods, some approaches lend themselves better than others to modeling social science data (Chen and Guestrin 2016; Lundberg et al. 2020). In particular, tree-based methods deal well with various types of data often encountered in the social sciences (e.g., categorical and numerical variables) and can handle missing data. They also do not require exponential increases in sample size as the feature space increases. Furthermore, the most popular tree-based methods, like the Random Forest (RF) or Gradient Boosting (GB) model, require relatively little tuning on the part of the researcher and can be applied out-of-the-box or with limited effort to most social science datasets (Breiman 1996; Freund and Schapire 1996).[16] These advantages make tree-based approaches an attractive class of ML methods to include in the benchmarking step. As a case in point, consider the nonlinear associations in Figure 3, which were introduced earlier. As before, the black lines illustrate true associations and the green diamonds are a hypothetical dataset generated by the true association and some white noise. The blue dots and red squares are out-of-sample predicted values based on a GB and an RF model trained to 80 observed data points and fed with 100 uniformly distributed X values to generate $\hat{y}$-predictions. Both models learn the nonlinearity in the data well and do so without requiring any pre-hypothesized functional relationship.

I illustrate the first step of the framework by estimating a Super Learner including a number of GB and RF models with various parameter estimates to the toy example introduced earlier. I also include the two researcher-hypothesized models to the Super Learner, reflecting the model including linear (`SL.GLM_Linear`) and quadratic (`SL.GLM_Quadratic`) specifications of age. For illustrative purposes, I add the true model (`SL.GLM_PieceLinear`), even though this model was not hypothesized originally. Table 2 shows a truncated output from the Super Learner and the root mean squared error (RMSE)[17] of various models included in the Super Learner, estimated based on 10 folds using cross-validation (for the full output, see Appendix Table A1). The best-performing model is, unsurprisingly, the true model, but its performance is closely tracked by a GB model that strongly outperforms the linear and the quadratic specifications. Including the GB model with different parameters (in this case a slow learning rate combined with high or low tree depth) leads to overfitting and a strong

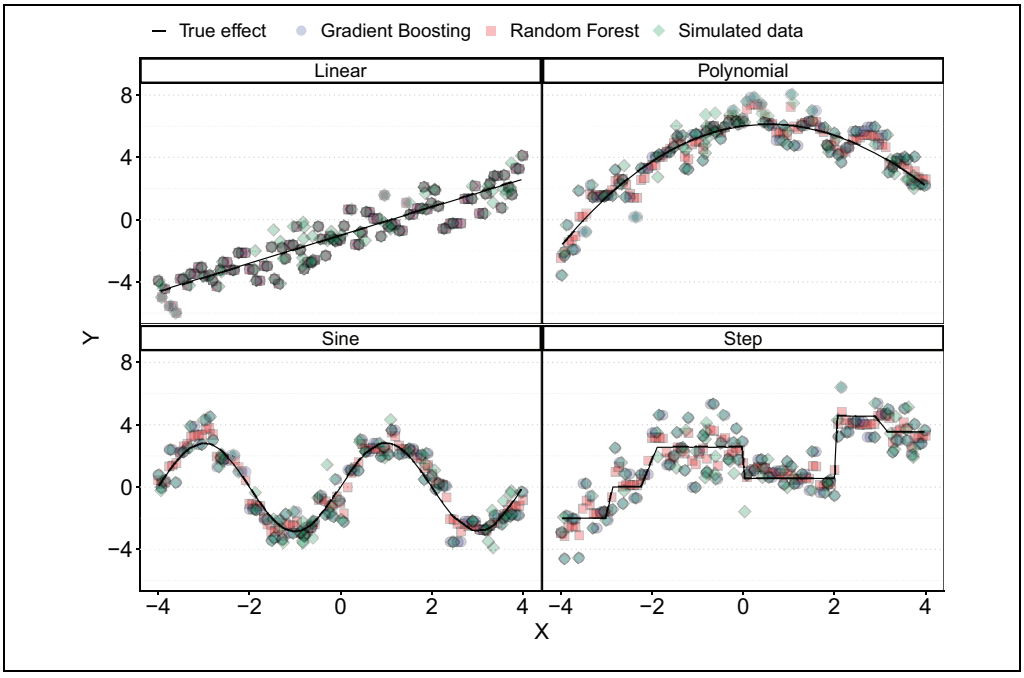


Figure 3. Predicted outcome using a GB (blue dots) and RF (red squares) model fit to noisy data (green diamonds).
Note: True effect is depicted with the black line. Each model is estimated based on 80 uniformly distributed observations for the X variable, plugged into the specified functional forms (see the Appendix). Predictions are made using another 100 uniformly distributed observations. The default parameter values for the RF and GB (with `nrounds` set to 200) models are used from the `randomForest` and `xgboost` packages in `R`.

Table 2. Truncated Super Learner Performance Using the Toy Example Data.

Model	Ave.	SD	Min.	Max.
SL.GLM_PieceLinear	1.218	0.060	1.046	1.320
SL.GB_200_1_0.1	1.229	0.059	1.107	1.327
SL.GB_500_1_0.1	1.231	0.059	1.104	1.323
SL.GB_100_1_0.1	1.241	0.059	1.129	1.351
RF_200_2	1.383	0.064	1.229	1.544
SL.GLM_Quadratic	1.384	0.063	1.174	1.536
SL.GLM_Linear	1.413	0.065	1.321	1.569
SL.GB_200_5_0.01	2.698	0.088	2.484	2.926
SL.GB_200_6_0.01	2.698	0.088	2.486	2.927

Note: The models included in the Super Learner are a GB model with parameter grid: `ntrees` = [100, 200, 500], max depth = [1, 2, 3, 4, 5, 6], `shrinkage` = [0.01, 0.1], and a Random Forest model with parameter grid: `mtry` = $[\sqrt{n},\ 2\sqrt{n}]$, `ntree` = [100, 200, 500]. Super Learner performance based on RMSE & 10-fold CV. GB parameters: `n_rounds_max_depth_eta`, RF parameters: `n_trees_mtry`. Output is truncated for brevity, see Appendix Table A1 for the full output.

underperformance in terms of fit. This further illustrates the necessity of including a large set of models into an ensemble like the Super Learner.

Step II: Estimating and comparing model fit

The second component of the proposed framework is a metric to compare the fit of benchmarking model $\bar{f}(\cdot)$ from the previous step with that of the hypothesized model $\tilde{f}(\cdot)$. Typically, comparative statistics like the Akaike Information Criterion (AIC) or LR tests (in case of nested models) are used to compare two or more explicitly hypothesized models. These statistics are not easily transferred to ML methods, mainly because they rely on in-sample diagnostics and specified functional forms or require estimation of degrees of freedom, which are challenging for ML models (Janson, Fithian, and Hastie 2015). As a result, ML methods are typically evaluated on their out-of-sample performance (Hastie et al. 2009; Shmueli 2010). The broad comparability of out-of-sample predictions across modeling domains makes them an attractive fit metric for the framework (Stone 1977; Van Der Laan and Dudoit 2003; Verhagen 2022).

Formally, the out-of-sample estimation of model fit would require splitting the total dataset $\mathcal{D}$ into two partitions, a training set, $\mathcal{D}_{\text{train}}$, used to estimate the model, and a test set, $\mathcal{D}_{\text{test}}$, used to evaluate the model's fit. By making predictions with the model estimated based on $\mathcal{D}_{\text{train}}$ but using data from $\mathcal{D}_{\text{test}}$ to make predictions, the latter predictions can be compared with the actually observed outcomes, and summary metrics of fit like the RMSE can be calculated (Hastie et al. 2009). Separating off a testing set is generally preferred for a truly out-of-sample estimate of fit, but it does sacrifice part of the sample available for estimation (usually 20–30 percent).[18] A common alternative is K-fold cross-validation, where the data are split into K equal-sized folds. The model is estimated K times, each time omitting one fold and using the omitted fold to generate predictions (Kohavi 1995).[19] The Super Learner approach discussed in Step 1 similarly relies on K-fold cross-validation. A closely related approach is Monte Carlo cross-validation, where M random splits of $\mathcal{D}$ into training and testing sets are made rather than mutually exclusive folds. The latter can be helpful when the data have additional structure that complicates separation of $\mathcal{D}$ into distinct subsets. The two approaches have been shown to be similar in practice (Yousef 2020). Another benefit of implementing cross-validation to obtain a metric of model fit is that bootstrapped estimates of the coefficients for the researcher-hypothesized model $\tilde{f}(\cdot)$ are obtained, which can further help identify a lack of robustness in an estimated coefficient.[20]

As others have noted, fit need not automatically equate with the best model from an inferential perspective (Grimmer et al. 2021; Muñoz and Young 2018). This discussion is often concerned with including colliders or post-treatment variables in a model, or simply including as many variables as a researcher can possibly think of. Such practices typically improve fit, but they can invalidate inference. This is the central reason why the proposed framework uses the same inferential reasoning in terms of which explanatory variables are included in the flexible model as the originally hypothesized model. Much of the criticism regarding blind "fit-hunting," where all causal or

Table 3. Three linear additive models and a GB model estimated on a train set consisting of 80 percent of the total dataset.

Variable	*N*	Mean	SD	Min.	Pctl. 5	Pctl. 95	Max.
RMSE							
GB	1000	1.241	0.06	1.062	1.145	1.34	1.414
Linear	1000	1.417	0.064	1.195	1.314	1.522	1.643
Quadratic	1000	1.391	0.065	1.181	1.284	1.495	1.602
PieceLinear	1000	1.222	0.06	1.049	1.124	1.316	1.414
GB vs. Linear	1000	–0.176	0.049	–0.315	–0.254	–0.097	–0.024
GB vs. Quadratic	1000	–0.149	0.046	–0.284	–0.222	–0.073	0.008
GB vs. PieceLinear	1000	0.019	0.02	–0.042	–0.013	0.051	0.109
R^2							
GB	1000	0.656	0.005	0.638	0.647	0.664	0.671
Linear	1000	0.626	0.005	0.609	0.617	0.635	0.64
Quadratic	1000	0.630	0.005	0.613	0.621	0.639	0.645
PieceLinear	1000	0.656	0.005	0.639	0.648	0.665	0.672
GB vs. Linear	1000	0.030	0.002	0.024	0.027	0.033	0.037
GB vs. Quadratic	1000	0.025	0.002	0.019	0.022	0.028	0.032
GB vs. PieceLinear	1000	–0.001	0.000	–0.002	–0.001	0.000	0.000

Note: The three linear additive models assume the effect of age to be linear, quadratic, or piece-wise linear, respectively. A GB model is also fit using the `xgboost` package in R with the following parameters: `depth` = 1, `nrounds` = 200, `eta` = 0.1, which were shown to be optimal in the Super Learner routine (see Table 2). RMSE and OOS R^2 is calculated for each model based on the remaining 20 percent of the data. The last three rows of the RMSE and R^2 parts show bootstrapped results of the difference between the optimal GB model and each of the linear additive models. 1,000 splits of the total dataset into train and test sets are evaluated.

inferential logic is effectively abandoned in favor of model fit, is thus not applicable, and variable selection is driven by the substantive question rather than fit (Grimmer et al. 2021:412).[21] However, it can be challenging to identify when a difference in fit between the flexible and researcher-hypothesized model warrants a re-evaluation of the latter. For example, improvements in fit for a covariate uncorrelated with the interest variable may have little bearing on the consistency of the interest variable.[22] Ideally, we would like clear statistical guidance on the extent of bias in an underspecified model. In practice, the decision to unpack the flexible form model will likely depend on the willingness of both the researcher to defend a specification with an apparent lack of fit, and the research community to accept it. This is not a consequence of the framework, but a level of uncertainty resulting from embracing a state of the world where we acknowledge that we have very little knowledge about the true DGP and are unwilling to make stringent assumptions on it a priori.

Putting the second step of the framework into practice, I compare the model fit of the GB model identified in Step 1 with the two hypothesized alternatives for the toy example. I include two fit metrics: the *R*-squared (R^2) and the RMSE, and evaluate both using 1,000 splits of the dataset into train and test sets. The results are summarized in Table 3. For illustration's sake, the true piece-wise model is included in this step as well. The bottom three rows for both metrics show the difference in the fit metric between the flexible model and the hypothesized models. The flexible model

strongly outperforms the linear and quadratic models, but it does not improve on the true model, as should be expected.[23]

Step III: Unpacking the ML model

The third and final component of the proposed framework is a method to generate understanding of the ML model whenever it outperforms the hypothesized model in a way that warrants re-evaluation of the originally hypothesized functional form. The increasing application of ML models in everyday life has led to pressure to improve understanding of the inner workings of ML models.[24] These developments have led to the emergence of X-AI, which focuses on understanding the patterns underlying ML models. Two general approaches can be identified within X-AI: global and local explainability. The former attempts to describe the mechanics of a model in general terms, that is, which variables tend to be important for the model's overall performance. The reporting of variable importance measures are an example of this approach (Baćak and Kennedy 2019; Brand et al. 2021). In local explainability, the goal is to explain the drivers of singular predictions made by a model. Local explanation methods are more appropriate for the framework, as they provide insights into the actual functional patterns between covariates and the outcome.

Various local explanation methods have been developed for different substantive questions that may be asked of a model (Doshi-Velez and Kim 2017; Lipton 2018; Zhou et al. 2021).[25] For example, much explainability research is focused on providing additional ad-hoc context into a model's predictions for practitioners making (high-stakes) decisions. In such cases, the ability to *quickly* calculate and extract information underlying a model's prediction could be more relevant than emphasizing fidelity. Conversely, an ethical reviewer of a system-in-action might require more fine-grained and detailed explanations. Within the proposed framework, the onus is on understanding the underlying patterns found by the ML model, rather than any practical or regulatory concerns. This motivates the use of Shapley values, which have a number of unique and attractive properties with extensive theoretical grounding (discussed in more detail below).[26] Importantly, recent advances have made their estimation computationally feasible and rekindled a general interest in their use (Aas, Jullum, and Løland 2021; Heskes et al. 2020; Samek et al. 2019). Shapley values also align closely with human intuitions of model interpretation, making them appropriate for the iterative process envisioned in the framework (Doshi-Velez and Kim 2017; Lundberg and Lee 2017; Rudin 2019). Shapley values have been applied in various fields, notably medicine (Lundberg et al. 2020; Tang et al. 2021), but have also found applications in sociological research, for instance, as a way to understand complex network dynamics (van der Laan et al. 2022) and to resolve path dependencies in decomposition metrics like segregation indices (Elbers 2023).

Shapley values for model explanation

Shapley values were originally developed within cooperative game theory to assist in the task of distributing a game's overall payout to its participants. Because a

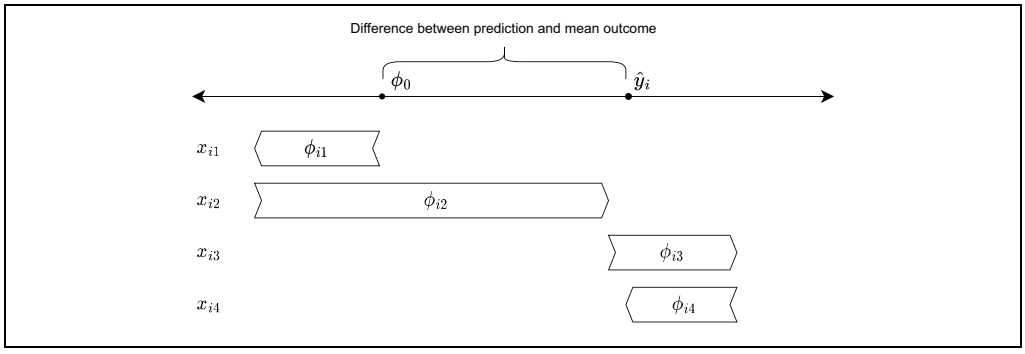


Figure 4. Prediction $\hat{y}$ and mean value ϕ_0 based on a model with four explanatory variables. *Note*: The difference between the prediction and overall mean is decomposed along the Shapley values of each of the four variables.

game's payout can depend on its participants' actions in a potentially complex manner, such an attribution function is not trivial to determine—for instance, some players' participation may have no bearing on the outcome at all. Shapley (1953) developed a perturbation-based approach to determine such an allocation mechanism for games of arbitrary complexity that possesses various attractive theoretical guarantees.

When used for model explanation, every single prediction $\hat{y}_i$ that is made by a model is viewed as the payout of a potentially complex game, where the covariate values x_{ik} underlying the prediction are viewed as the game's K participants. The goal of Shapley values is to attribute the value of the prediction $\hat{y}_i$ among its K covariates x_{ik}:

$$\hat{y}_i = \phi_0 + \sum_{k=1}^{K} \phi_{ik}. \tag{3}$$

Here, $\hat{y}_i$ is a single prediction made by some potentially complex model and ϕ_0 resembles the overall mean across predictions. The Shapley values thus attempt to decompose the deviance of a specific prediction made by the model with respect to the mean prediction across all observations. The additive decomposition of a prediction $\hat{y}_i$ amongst its covariates leads to the typical waterfall plots associated with Shapley values (see Figure 4). In this example, the difference between $\hat{y}_i$ and the overall mean ϕ_0 is decomposed among the four covariates. The first and fourth covariates, x_{i1} and x_{i4}, have a negative impact on $\hat{y}_i$ as their associated Shapley values drive the prediction to the left. The second and third covariates, x_{i2} and x_{i3}, have a positive impact. Taken together, they add up to the total difference between $\hat{y}_i$ and the mean. Each prediction $\hat{y}_i$ thus has its own K Shapley values ϕ_{ik} that precisely determine the difference between $\hat{y}_i$ and ϕ_0.

As every single prediction $\hat{y}_i$ is decomposed into K Shapley values, one for each covariate, the approach effectively leads to a matrix of $N \times K$ Shapley values:

$$\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1K} \\ x_{21} & x_{22} & \dots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NK} \end{bmatrix} \xrightarrow{\text{Shapley value estimation}} \begin{bmatrix} \phi_{11} & \phi_{12} & \dots & \phi_{1K} \\ \phi_{21} & \phi_{22} & \dots & \phi_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{N1} & \phi_{N2} & \dots & \phi_{NK} \end{bmatrix}, \tag{4}$$

where all K Shapley values per observation exactly add up to the model's prediction for that observation:

$$\phi_o \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} \sum_{j=1}^{K} \phi_{1j} \\ \sum_{j=1}^{K} \phi_{2j} \\ \vdots \\ \sum_{j=1}^{K} \phi_{Nj} \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix}. \tag{5}$$

Note that if the model was a standard linear additive model without an intercept (i.e., $\hat{y}_i = \sum_{k=1}^{K} \hat{\beta}_k x_{ik}$) then calculating Shapley values that satisfy the additive function $\hat{y}_i = \phi_0 + \sum_{k=1}^{K} \phi_{ik}$ is simple. Every Shapley value should simply correspond to the estimated coefficient $\hat{\beta}_k$ times the covariate value x_{ki} for that observation: $\phi_{ki} := \hat{\beta}_k x_{ki}$, as shown by Aas et al. (2021). If we were to plot the N tuples consisting of the Shapley values ϕ_{ik} and the covariates x_{ik} for a single variable k, this would result in a perfect linear relationship between the covariates and the Shapley values with slope $\hat{\beta}_k$. When more complex predictive models are used, plotting the Shapley values ϕ_{ik} and the covariate values x_{ik} in such a joint manner provides a graphical way to study the implied association between a covariate and the outcome (Lundberg et al. 2020:59).

The process of calculating the K Shapley values for a single prediction $\hat{y}_i$ relies on a perturbation-based approach, where we assess the effect of omitting information on a covariate x_{ik} on that model's prediction.[27] This is done by defining all possible "information sets" consisting of a set of s out of the total K covariates. Define $\mathcal{M}_i^k$ to be some information set including covariate k, and $\mathcal{M}_i^k/k$ to be its complement, excluding information on covariate k. For each $\mathcal{M}_i^k$, we also make a prediction for its equivalent in $\mathcal{M}_i^k/k$. This leads to two predictions—one including information on variable k and one without—which are then differenced. The Shapley value ϕ_{ik} for some prediction $\hat{y}_i$ and covariate x_{ik} is defined as a weighted mean over all information sets $\mathcal{M}^k$:

$$\phi_{ik} = \overbrace{\sum_{\mathcal{I} \in \mathcal{M}^k}}^{\text{Sum over all information sets}} \underbrace{\frac{|\mathcal{I}|!(|\mathcal{M}| - |\mathcal{I}| - 1)!}{\mathcal{M}!}}_{\text{Weighting function}} \overbrace{(\bar{f}(\mathcal{I} \cup k) - \bar{f}(\mathcal{I}))}^{\text{Difference in prediction}}. \tag{6}$$

This process is then repeated for every variable k and for every prediction $\hat{y}_i$, leading to the $N \times K$ matrix in Equation 4.

Shapley values have a number of unique theoretical properties that make them attractive as a method to generate understanding of models. First, the sum of all

Shapley values ϕ_{ik} and the mean prediction, ϕ_0, match the actually observed prediction $\hat{y}_i$. Second, whenever the inclusion of variable k into the information set has no effect on the model's prediction—$\bar{f}(\mathcal{I}) = \bar{f}(\mathcal{I} \cup k)$ for all $\mathcal{I}$— its Shapley value is zero. Third, whenever two covariates j and k contribute equally to every prediction—$\bar{f}(\mathcal{I} \cup j) - \bar{f}(\mathcal{I}) = \bar{f}(\mathcal{I} \cup k) - \bar{f}(\mathcal{I})$ for all $\mathcal{I}$—their Shapley values are the same: $\phi_{ij} = \phi_{ik}$. Fourth, they are consistent with respect to addition and multiplication across models. Prior work shows Shapley values are the only local explanation method to possess these properties (for detailed discussions and proofs, see Lundberg and Lee 2017; Lundberg et al. 2020; Shapley 1953; Young 1985).

The computational burden of calculating Shapley values stems from three elements. First, the number of possible information sets is 2^K and scales exponentially with the number of covariates. Second, calculating the prediction $\bar{f}(\mathcal{I})$ when the information set $\mathcal{I}$ consists of a subset $\mathcal{S} \subset \mathcal{K}$ requires an expectation $\mathrm{E}[x|x^* = x_s]$ to be evaluated for the variables $\bar{\mathcal{S}}$ not in the information set. Typically, the empirical density of the missing variables is used, but this assumes independence of the feature space. Joint Gaussian and copula-based methods have been proposed as an alternative, which further increase computation time (Aas et al. 2021; Heskes et al. 2020). Finally, many predictions have to be assessed to ensure sufficient tuples $[\phi_{ik}, x_{ik}]$ to infer patterns in the model of interest, thus further scaling the computational requirements linearly. Fortunately, computationally efficient methods have been developed to calculate Shapley values. Here, I use the Shapley value estimation method optimized for tree-based models, Tree SHAP, following Lundberg et al. (2020).

Tree SHAP exploits the structure of the estimated tree, allowing for a much smaller set of relevant information sets to be assessed. Specifically, information sets that only differ from one another in terms of variables that do not feature in the branches of the tree can be ignored, as they will not lead to different predictions.[28] This has made estimation efficient to the point that exact Shapley values can be calculated, rather than having to rely on numerical approximations (Lundberg et al. 2020:64–65), and even renders pairwise Shapley values feasible to calculate in reasonable time (Lapuschkin et al. 2019; Lundberg et al. 2020). Pairwise Shapley values assess the joint omission of two variables from the model and the subsequent effect on the outcome, allowing interactions to be explicated. In the standard Shapley decomposition in Equation 3, interactions would be divided among the relevant interacting variables.[29] A final benefit of the Tree SHAP approach is that the independence assumption when making expectations for those variables outside of the information set, $\bar{\mathcal{S}}$, is relaxed. A more detailed discussion of Shapley values, pairwise Shapley values, and the Tree SHAP algorithm is provided in the Appendix.

To illustrate how Shapley values can be used to generate understanding into an ML model, I calculate all N Shapley values for the age variable based on predictions using the GB model identified in the toy example above. Figure 5 presents a joint scatterplot of the N Shapley values $\phi_{i,\,\mathrm{age}}$ together with the N values of the covariate $x_{i,\,\mathrm{age}}$. I normalize the Shapley values as well as the true underlying effect such that they can be visualized jointly.[30] The Shapley values accurately recover the piecewise linear association in the underlying data, and would provide concrete insights into how the

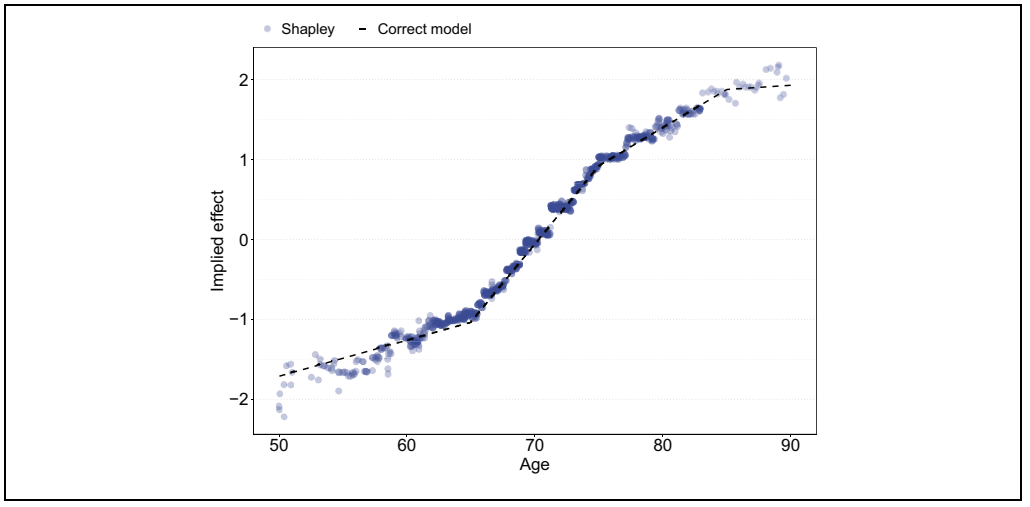


Figure 5. Implied effect size of age through Shapley values versus the correctly specified model.
Note: Both implied effects have been normalized to allow for comparisons with the true functional form.

hypothesized functional forms could be improved from the linear and quadratic specifications of age.

RESEMBLANCES TO OTHER COMPUTATIONAL FRAMEWORKS

Before applying the framework to three empirical cases, I briefly discuss a number of methods and frameworks that share a similar intuition and computational appetite to the proposed framework. In spirit, the first two steps of the framework are indebted to modern types of specification tests that rely on comparing residuals from classic non-parametric regression to those provided by a researcher's own hypothesized model.[31] Such approaches can identify a broader range of nonlinear functional forms than those considered by the RESET test, for example, although formal tests of improved fit remain challenging to compute due to the dependence of local estimates to the data at hand. More generally, the amount of data required to compute such local estimates scales exponentially with the number of independent variables (Bishop 2006; Yatchew 2003). The ML methods proposed for the framework presented here retain considerably more structure and are more efficient from a data perspective than non-parametric tests are, improving efficiency and allowing for better post-estimation analysis.[32]

The proposed framework also shares similar goals to the field of model robustness. The emphasis in model robustness lies in exposing the potential role of researcher degrees of freedom during the model-building process (Sala-i Martin 1997; Simonsohn et al. 2020; Young 2019). A large number of "plausible" models are estimated that slightly differ to the hypothesized model to guard against researchers cherrypicking a specification. The key difference in model robustness is that variation is typically induced by excluding variables in the model, rather than assessing different functional

relationships among variables (Young and Holsteen 2017).[33] The functional complexity and the fit of the models is of limited importance (Slez 2019; Young 2019). The proposed framework instead focuses on the functional relationship among a set number of variables and the emphasis is on specification.[34] Both approaches could be combined by applying the proposed framework to different sets of explanatory variables, as I will illustrate in two empirical case studies. Despite their differences, both frameworks are similarly indebted to the growth in computational power that allows researchers to consider many different functional forms and to use these computational riches to be more critical of modeling assumptions and improve transparency into the research process (Muñoz and Young 2018:4).[35]

Closest in spirit to the proposed framework is an emerging literature in behavioral economics and psychology using ML models to find the predictive limit of behavioral theories (Agrawal et al. 2020; Kleinberg, Liang, and Mullainathan 2017; Peterson et al. 2021). The typical approach is to compare the performance of some behavioral theory to predict the outcomes of a stylized experiment (e.g., the fairness perception of two vignettes [Agrawal et al. 2020]) with a flexible ML model trained to the same data. If the ML method leads to better predictions than a behavioral theory, cases where the behavioral theory underperformed relative to the ML benchmark are studied. This research focuses solely on large-scale experimental data and has not been extended to empirical work as broadly as I propose here. The onus in this literature is on theory completeness rather than correct functional form specification. Post-estimation interpretation of the ML method—the third step in the proposed framework—is mostly ad hoc, if present at all.[36] However, the central premise of using ML to find patterns in data not necessarily hypothesized by a researcher is a key similarity between the two approaches.

A final strand of research similar to the proposed framework is the field of "autometrics," which also aims to find better-fitting models than a researcher's own hypothesizing might yield. The autometrics approach includes a large number of base transformations of the explanatory variables into a linear additive functional form. This "stacked" model is then iteratively trimmed to reach an optimal model using in-sample fit metrics and specification tests (Doornik and Hendry 2015). Autometrics is a more classical approach to flexible model-building compared to the ML methods proposed here. Specifically, the autometrics setup can be subsumed by the set of ML methods called "generalized additive models" and spline-based methods, which can be folded into Step I of the proposed framework (Hastie and Tibshirani 1987; Hastie et al. 2009).[37] Related, scholars have put forward similarly stacked models that include interactions but then apply variable selection through least absolute shrinkage and selection operator regression (Blackwell and Olson 2022; Beiser-McGrath and Beiser-McGrath 2020). The proposed framework shares the same general intuition of the above approaches but uses a broader set of methods such that a wider range of patterns can be identified, and it emphasizes out-of-sample evaluation to reduce the risk of overfitting and path dependence in eliminating regressors from the model. As with all frameworks mentioned here, the inclusion of the third step in the proposed framework is another key difference with the above approach.

APPLYING THE FRAMEWORK IN PRACTICE

I apply the proposed framework to three case studies. The first is a simulation based on the Mincerian wage equation, a classic field of research investigating the returns of additional years of schooling on a person's income (Lemieux 2006). The second is a hedonic regression of house prices applied to a large dataset of transactions in the London retail market (Malpezzi 2003). The third is a study of the demographic determinants of voting preferences in the United States using the General Social Survey (GSS) (Davis and Smith 1991). The Mincerian wage application is relevant because it provides an example of a functional form that has been actively innovated upon over the past decades and illustrates how the proposed framework can speed up model-building. I chose the hedonic regression and voting examples because both enjoy considerable academic interest, and models typically include a number of standard control variables with a (novel) interest variable. Because the latter is often correlated with the control variables, a correct functional relationship is crucial for inference. For these latter two case studies, I evaluate whether the typical complexity in which control variables feature in the functional form is sufficient, and improve them where necessary. Across these cases, I identify various nonlinearities and interaction effects using the framework.[38]

Application I: Mincerian wage simulation

The Mincerian wage equation is a classic economic tool used to estimate the effect of an additional year of education on an individual's wages—the "return to education" (Lemieux 2006). As mentioned above, the Mincerian wage equation has been actively innovated upon over the past decades. In the original functional form, log yearly wages was related to years of education and work experience in a linear additive fashion:

$$\ln(\text{wages}_i) = \beta_0 + \beta_1 x_{\text{educ},i} + \beta_2 x_{\text{exp},i} + \epsilon_i. \tag{7}$$

Subsequently, a square term was added to allow for a level of nonlinearity in the effect of work experience:

$$\ln(\text{wages}_i) = \beta_0 + \beta_1 x_{\text{educ},i} + \beta_2 x_{\text{exp},i} + \beta_3 x^2_{\text{exp},i} + \epsilon_i. \tag{8}$$

More recently, a step function was added in the effect of education to allow for different linear effects by level of education:

$$\begin{aligned}\ln(\text{wages}_i) = \beta_0 + \beta_1 x_{\text{educ\_0\_8},i} + \beta_2 x_{\text{educ\_9\_10},i} + \beta_3 x_{\text{educ\_11\_12},i} + \\ \beta_4 x_{\text{educ\_13\_14},i} + \beta_5 x_{\text{educ\_15+},i} + \beta_6 x_{\text{exp},i} + \beta_7 x^2_{\text{exp},i} + \epsilon_i.\end{aligned} \tag{9}$$

For illustrative purposes, I also study a fourth, hypothetical functional form where each coefficient slightly differs by sex:

$$\begin{aligned}\ln(\text{wages}_i) = {} & \mathrm{I}(x_{\text{sex},i} = \text{Female})[\beta_0^* + \beta_1^* x_{\text{educ\_0\_8},i} + \beta_2^* x_{\text{educ\_9\_10},i} + \beta_3^* x_{\text{educ\_11\_12},i} + \\ & \beta_4^* x_{\text{educ\_13\_14},i} + \beta_5^* x_{\text{educ\_15+},i} + \beta_6^* x_{\text{exp},i} + \beta_7^* x_{\text{exp},i}^2 + \epsilon_i] + \\ & \mathrm{I}(x_{\text{sex},i} = \text{Male})[\beta_0 + \beta_1 x_{\text{educ\_0\_8},i} + \beta_2 x_{\text{educ\_9\_10},i} + \beta_3 x_{\text{educ\_11\_12},i} + \\ & \beta_4 x_{\text{educ\_13\_14},i} + \beta_5 x_{\text{educ\_15+},i} + \beta_6 x_{\text{exp},i} + \beta_7 x_{\text{exp},i}^2 + \epsilon_i].\end{aligned} \tag{10}$$

To illustrate the proposed framework, I simulate data using the four specifications above as DGPs, plugging coefficient estimates as found in the recent empirical literature into each specification. For the final specification, I vary the coefficients of sex to fall within a standard error of the coefficients found in the literature (see the Appendix for the DGPs) (Heckman et al. 2018; Lemieux 2006).

Based on these four DGPs and a synthetic sample of 50,000 individuals' age, years of education, years of work experience, and sex, I generate four outcomes: one using each DGP. The synthetic sample is based on the GSS 2018, such that the distribution of the explanatory variables is representative of an actual working population. For illustrative purposes, I calibrate the error term in each functional form such that the proportion of explainable variance is constant in every dataset irrespective of the DGP used to generate the outcome variable. Descriptives for the dataset can be found in Appendix Table A2. Linear-I refers to the first functional form above provided by Equation 7, Linear-II refers to Equation 8, and so on.

The above leads to four datasets consisting of a distinct vector of outcomes y and the same matrix of explanatory variables X, where each outcome vector follows from one of the four underlying functional forms in Equations 7 to 10. The first is based on a DGP where both education and work experience affect the outcome linearly, the second where work experience follows a second-degree polynomial relationship, and so on. I next propose four hypothetical functional forms, $\tilde{f}(\cdot)$, to estimate to each of the four datasets. The first three functional forms are equal to those defined in Equations 7 to 9, as well as a fourth functional form that is equal to Equation 9 but includes a dummy for sex. This means none of the four hypothesized models aligns with the true DGP in the fourth dataset, which follows the DGP provided in Equation 10. I include this fourth form for illustrative purposes, as it is common practice to "control" for group differences by including a dummy variable, although this might simplify the true pattern in the data, which in this case concerns an interactive effect rather than a level difference. Estimating these four functional forms means that for the first dataset, all four hypothesized models should have the appropriate flexibility to model the data. For the second dataset, only the second, third, and fourth functional forms should, and for the third dataset, only the third and fourth functional forms should be able to fit the data well. None of the proposed functional forms have sufficient flexibility to estimate the underlying DGP in the fourth dataset appropriately.

In addition to the four hypothesized functional forms, I estimate a Super Learner including various tree-based ML methods—the first step of the framework. A GB model performs best for each of the four datasets (see Appendix Tables A3 and A4). In the second step of the framework, I compare the model fit of the flexible model with the four hypothesized models using Monte Carlo cross-validation. The results show that the flexible model is able to match the true functional form's performance in the first three datasets and strongly outperforms the most flexible functional form in

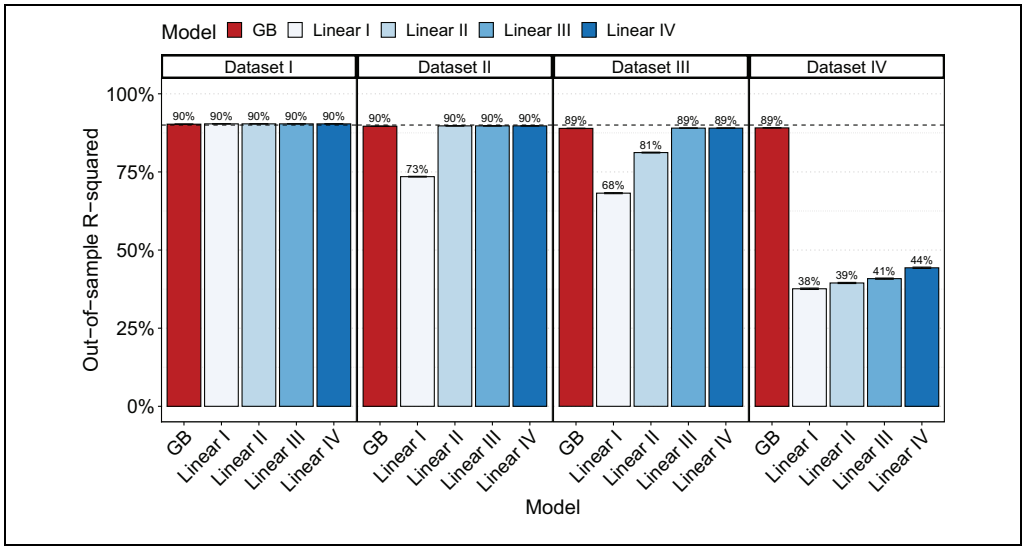


Figure 6. Out-of-sample R^2 for the four datasets with varying DGPs, using four functional forms and the GB model.

Note: The GB model converges on the correct functional form's model fit in each dataset, whereas the prespecified models only reach the maximum explanatory power when the functional form's complexity equals or outperforms that of the DGP. I used 100 splits into an estimation set of 80 percent and evaluation set of 20 percent.

the fourth dataset (see Figure 6). The framework thus identifies the underlying model without requiring the researcher to hypothesize a functional form in all four datasets.

Finally, the flexible model can be unpacked to infer why it improved on the hypothesized functional forms—the third step in the framework. We know the true DGP in this case, making it obvious why the flexible model outperformed the underspecified functional forms, but the Shapley values easily identify the correct association of the independent variables with the outcome (Figure 7). The Shapley values capture the linearity of both explanatory variables in the first dataset, the nonlinearity in years of work experience in the second dataset, and the step-wise function in the third dataset. The Shapley values also pick up the interaction with sex in the fourth dataset, as illustrated by varying the Shapley values by sex. Clearly, assuming linearity where none is present leads to a misrepresentation of the true returns to education; incorporating the correct flexibility is crucial for inference.

Application II: London house prices

As a second case study, I use a large dataset on transactions in the London housing market. The typical approach to modeling house prices is through a hedonic pricing model that assumes each house is composed of various traits for which buyers have certain preferences. Buyers effectively combine the value of individual traits to determine the (monetary) value of an entire house. In hedonic regression, interest typically

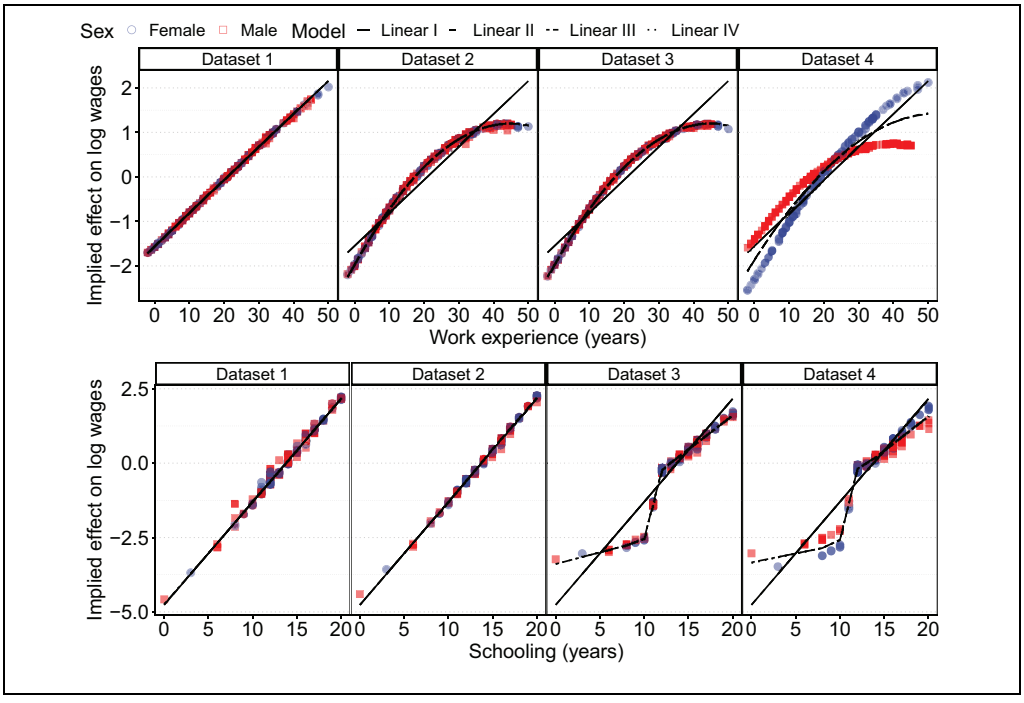


Figure 7. Effect of work experience (red squares) and schooling (blue circles) as predicted by the four estimated functional forms, and as implied by Shapley values.
Note: The correct functional form is approximated well by the Shapley values for each dataset and clearly shows an interaction effect between sex and both effects in the fourth dataset. Shapley values and implied effects from the estimated models were scaled such that they can be easily compared visually.

lies in the price elasticity of certain traits, for example, how much the house price increases with an additional square meter of living space or the inclusion of a garden. In practice, linear additive models are typically estimated relating observed house characteristics with log prices, although many authors have suggested that the assumptions of additive linearity implicit in this functional form might be unreasonable (Fan, Ong, and Koh 2006; Malpezzi 2003).

In this application I use typical house characteristics, like house size, number of rooms, and property type as explanatory variables. I also include a number of neighborhood characteristics. In many applications, researchers attempt to address spatial heterogeneity by including neighborhood-level observables like crime indices, travel distances to local centers, or deprivation scores. Again, in the typical model these variables are added in a linear additive framework, although it is often argued that spatial heterogeneity is considerably more complex (Elhorst 2010). To estimate the hedonic regression, I use a dataset of nearly 630,000 house sales containing basic house characteristics and a neighborhood identifier. The transaction data were collected and merged by the Reshare project hosted at the UK Data Service (Chi et al. 2021). I include neighborhood-level data from the Department of Transport and Communities and Ministry of Housing, Communities and Local Government.[39] I also include the

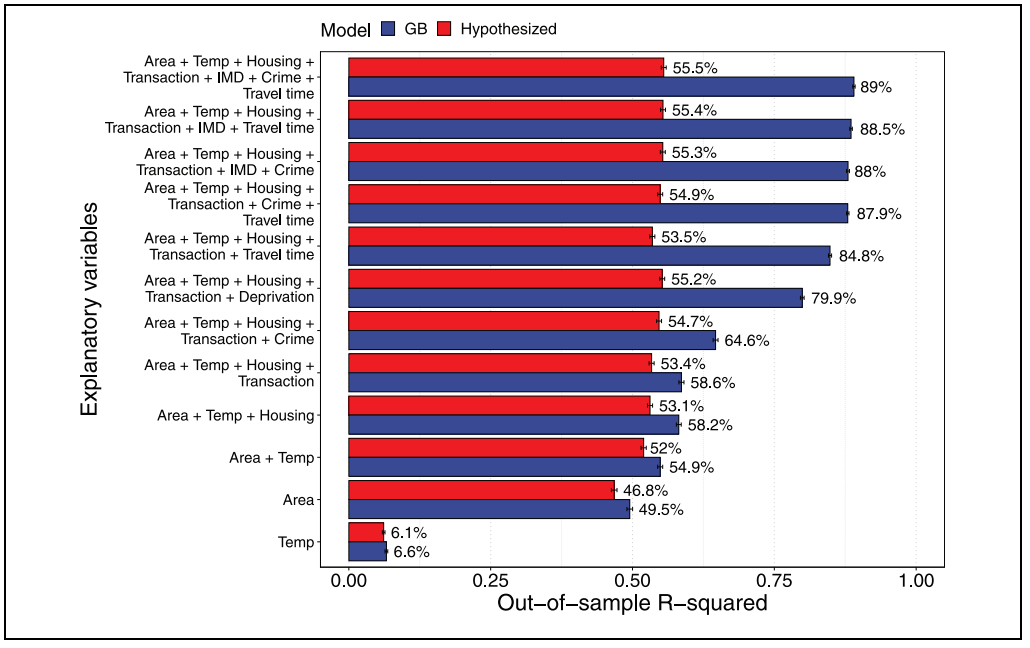


Figure 8. Out-of-sample R^2 for the hypothesized model and the flexible model.
Note: "Temp" contains yearly and monthly time trends. "Area" contains the house size and number of rooms. "Housing" includes the property type, and whether the house was newly built. "Transaction" reflects the type of transaction. "Crime," "deprivation," and "travel time" reflect the crime index, the deprivation index, and the travel time from the house to the nearest local hub, respectively. 100 splits into an estimation set of 80 percent and evaluation set of 20 percent are used.

year and month of the transaction. Descriptive statistics of the dataset are shown in Appendix Table A5.

For the hypothesized model, I estimate a linear additive functional form relating log house prices to the variables depicted in Appendix Table A5. As is customary in the literature, I assume linear trends for all continuous variables, including the temporal variables. The exact functional form is as follows:

$$\begin{aligned}\ln(\text{House price}_i) = {} & \beta_0 + \beta_1 x_{\text{area},i} + \beta_2 x_{\text{rooms},i} + \beta_{3,\ldots,5} x_{\text{propertytype},i} + \beta_6 x_{\text{new},i} + \\ & \beta_7 x_{\text{travel\_time},i} + \beta_8 x_{\text{crime},i} + \beta_9 x_{\text{deprivation},i} + \beta_{10} x_{\text{year},i} + \\ & \beta_{11} x_{\text{month},i} + \beta_{12,\ldots,16} x_{\text{transaction\_type},i} + \varepsilon_i .\end{aligned} \tag{11}$$

Following the framework, I start by estimating both the hypothesized model and a Super Learner to the data. In addition to the model in Equation 11, I also apply the framework to subsets of the full model. The best-fitting flexible model is again a GB approach, which will be used as the flexible model (see Appendix Tables A6 and A7 for the Super Learner output for the full set of covariates, and the selected models for the covariate subsets, respectively). In the next step, the fit of the flexible model is compared to the hypothesized model. The results are summarized in Figure 8 and are striking. Already among the simplest housing characteristics, like size and property

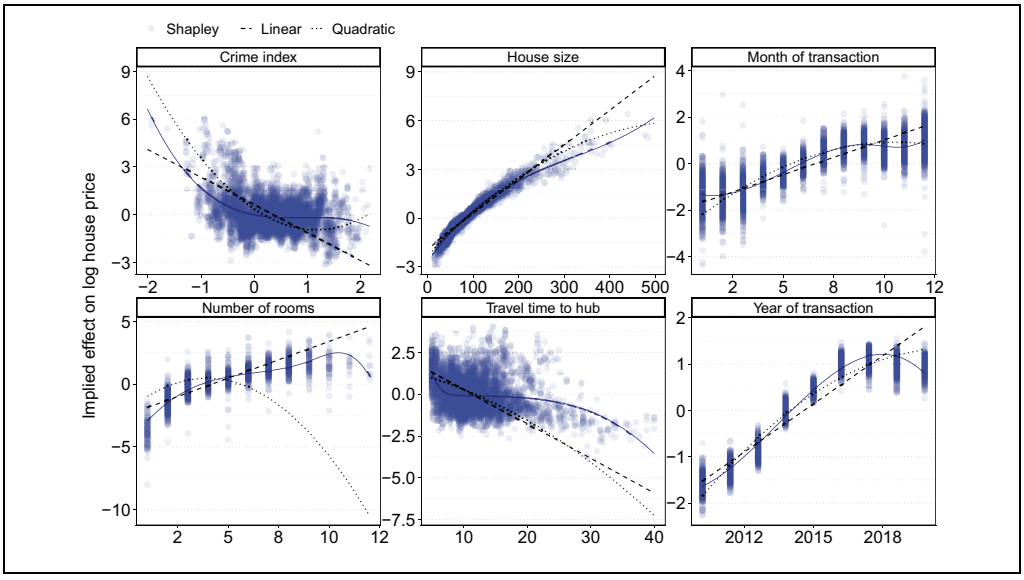


Figure 9. Plots show implied effect from a typical linear additive model (dashed line), when adding a squared term (dotted) and as implied by Shapley values (scatter).
Note: A LOESS fit is included in the Shapley plots to indicate the association between each variable and the outcome. Implied effects are scaled to allow for visual comparisons.

type, the flexible model improves on the linear additive functional form by about 5 pp in terms of the out-of-sample R^2. We see the largest improvement when adding neighborhood variables like "travel time to hub," with the difference in R^2 more than 30 pp. This strongly implies that spatial heterogeneity is poorly addressed by the functional form in Equation 11, a point to which I will return.

To evaluate why the flexible model outperforms the linear additive framework, I calculate Shapley values for the flexible model. These are visualized for six explanatory variables: the size of the house, the travel time to the nearest local hub, the crime index, number of rooms in the house, and the two temporal variables (year and month of sale). Figure 9 shows the results. We see nonlinearities for most of these variables, ranging from a slightly decreasing elasticity for the size of the house to piece-wise linearity in the effect for travel time and the crime index. It is also clear that the number of rooms, year, and month variables should all be modeled in a nonlinear way. When including these nonlinearities into the linear additive model, the fit improves and is strictly preferred according to an LR test,[40] although a remaining gap in model fit still points at further interactions and nonlinearities among variables, possibly across time. Specifically, the considerable noise in the Shapley values of the neighborhood-level variables implies that these variables do not follow a very precise pattern with respect to the neighborhood characteristics, and including nonlinearities may not suffice to capture the underlying patterns well.

Given the stark increase in model fit when adding neighborhood-level variables, I include random intercepts on the neighborhood level to a simple model including the house size and temporal explanatory variables. This effectively provides a fully non-linear association for each neighborhood's characteristics and the outcome variable, which seems reasonable based on the large increases in model fit when adding any of the neighborhood characteristics to the flexible model, combined with the variation in their Shapley values. Based on this modification of the functional form, the out-of-sample R^2 of the model strongly improves to 84 percent, which is much closer to that of the flexible model. In other words, the heterogeneity among neighborhoods could not be captured by the three variables when included in the model in either a linear or polynomial form—although this strategy is often encountered in the literature (Malpezzi 2003)—and requires more intricate modeling. In this case, the model's ability to estimate flexible patterns led to the identification of individual neighborhoods, thus mimicking a random intercept approach, and the framework illustrates that spatial heterogeneity is highly predictive, but poorly accounted for by the available neighborhood characteristics (Elhorst 2010). As in the previous example, assuming linearity would have led to incorrect inferences in the elasticities of most house characteristics in the data. It would be particularly problematic for inference to ignore the considerable variation at the neighborhood level, as this clearly points at important omitted variables that could bias inference.

Application III: Party identification in the United States

As a third and final case study, I evaluate the demographic determinants of party identification in the United States using the GSS (Davis and Smith 1991). Party identification is of substantive interest in the social sciences (Freeden, Sargent, and Stears 2013), and the GSS is an often-used resource for this purpose. Throughout this literature, a number of demographic variables are used as typical controls, including respondent's age, sex, race, educational attainment, and income. Another variable of substantive interest is usually included in the analysis, like cognitive ability (Meisenberg 2015) or social class (Morgan and Lee 2017). These interest variables naturally tend to correlate with the control variables, making correct specification critical.

Appendix Table A8 shows descriptive statistics of the GSS containing information on voting preferences and demographic characteristics between 1974 and 2018. The outcome of interest is a seven-point scale indicating whether the respondent identifies strongly with the Democratic party (value of 1) or the Republican party (value of 7). The GSS also includes the respondent's age, years of schooling, income level across 12 brackets, sex, and race, as well as the year of the survey wave. As the hypothesized model, I use a standard linear additive model as often encountered in the literature (Freeden et al. 2013). Most functional forms include a linear time trend for the year of the survey, and add most demographic variables as a linear determinant or dummy. The hypothesized functional form is as follows:

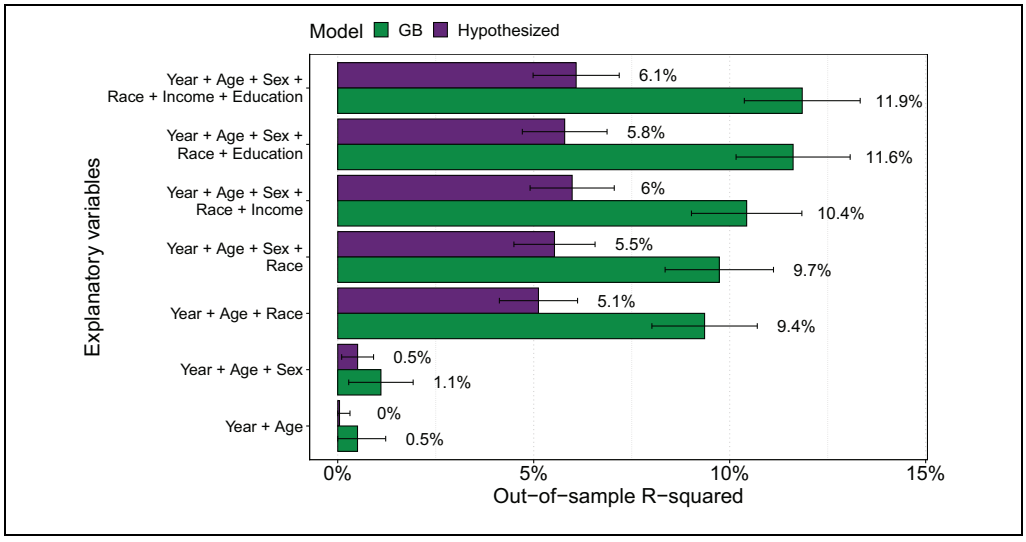


Figure 10. Out-of-sample R^2 for the hypothesized model compared to the flexible model.
Note: The "year" set reflects the survey wave. The "age" set reflects the respondent's age, "sex" their sex, "race" their race, "income" their income, and "education" their years of education. The optimal models from the Super Learner routine and the hypothesized models were estimated to 80 percent of the data, and the out-of-sample R^2 was calculated on the remaining 20 percent for each of the models. The data were split 1,000 times into estimation- and evaluation-sets.

$$y_i = \beta_0 + \beta_1 x_{\text{age, i}} + \beta_2 x_{\text{female, i}} + \beta_3 x_{\text{black, i}} + \beta_4 x_{\text{other, i}} + \beta_7 x_{\text{education, i}} + \beta_8 x_{\text{income, i}} + \beta_9 x_{\text{survey\_year, i}} + \epsilon_i, \tag{12}$$

and resembles that found in Meisenberg (2015).

I start by estimating Equation 12 as well as a Super Learner containing various tree-based ML models using both the full set of covariates and subsets. The best-performing model is again a GB model, although the RF also performs well (see Appendix Tables A9 and A10 for the Super Learner output for the full set of covariates, and the selected models for the covariate subsets, respectively). Figure 10 shows results from benchmarking the out-of-sample R^2 with the hypothesized models. The explanatory power of the flexible model is almost double that of the hypothesized model across subsets, indicating a considerable lack of appropriate specification in Equation 12. The flexible model improves most when race is added to the model, although the hypothesized model already underperforms when simply including temporal, age, and sex variables. It is also noteworthy that overall fit is comparatively low.

Using the full set of covariates, Figure 11 illustrates Shapley values and Shapley interaction values for the age, schooling, and income variables. These Shapley values show there are clear nonlinearities in the associations of age, income, and years of schooling with the outcome. Assuming these associations to be linear does not fit the implied effect well, and adding square terms improves the model fit considerably (Figure 11A).[41] Ignoring these complexities and assuming a linear functional form would wrongfully imply a decreasing effect by age, even though there is a clear

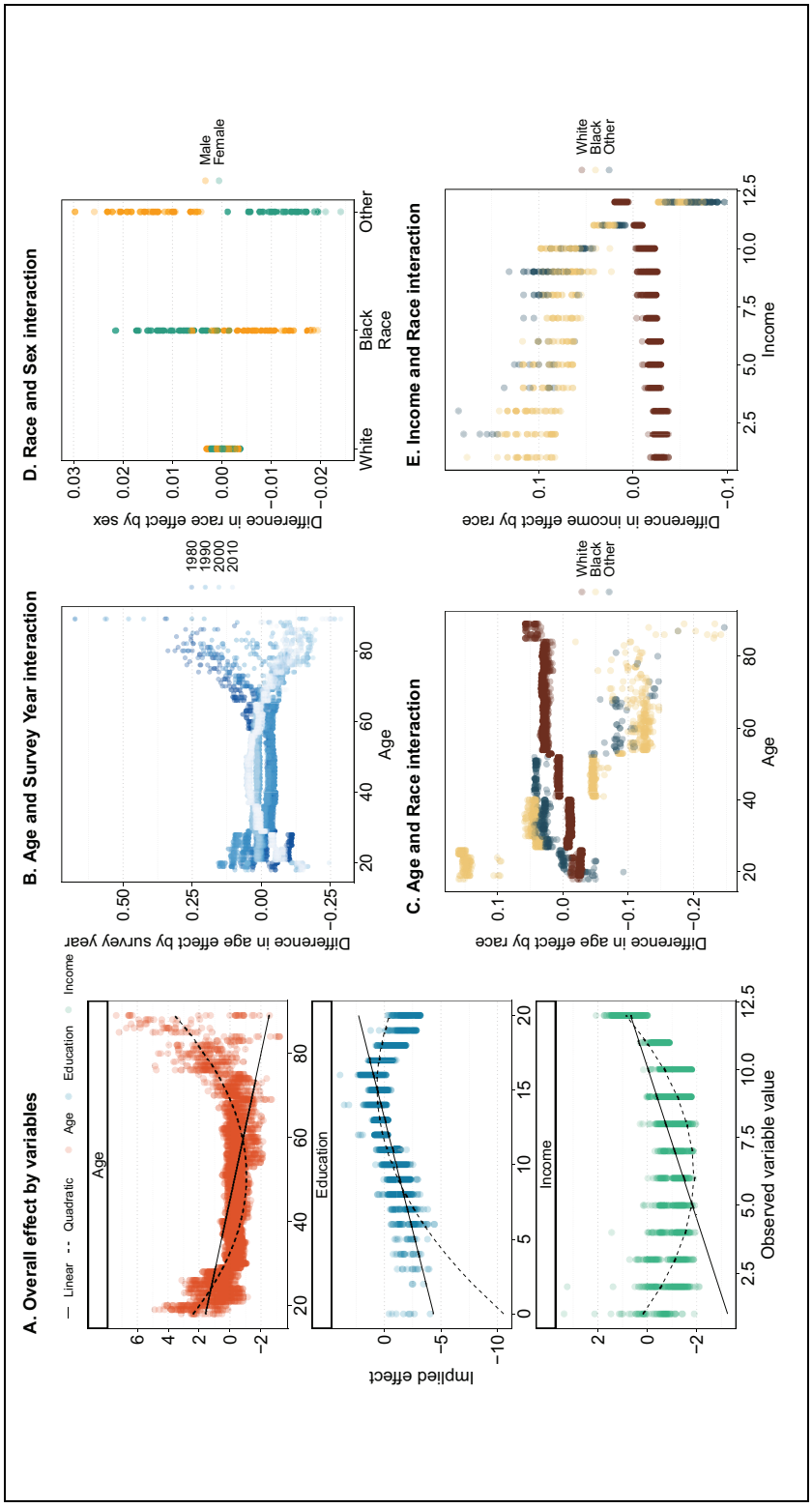


Figure 11. Overall implied effect as estimated by Shapley values for age, education, and income variables: (A) overall effect by variables, (B) age and survey year interaction, (C) age and race interaction, (D) race and sex interaction, and (E) income and race interaction.
Note: Panel A shows including effects as implied by a linear (solid) and quadratic (dashed) functional form. Panel B shows differences as implied by pairwise Shapley values in the effect per age relative to the overall effect by survey year, and by race in Panel C. Panel D shows differences as implied by pairwise Shapley values in the effect per race for females and males. Panel E shows differences as implied by pairwise Shapley values in the effect per race for income.

upward association at later ages. Similar incorrect conclusions would be drawn for both the education and income associations when assuming a linear additive model.

These direct effects can be further decomposed into a number of interactions by estimating pairwise Shapley values, allowing each Shapley value to be decomposed into a direct and indirect effect. The indirect effects indicate how Shapley values for certain observed characteristics deviate from the overall effect of the variable, conditional on a second variable. The estimates in panels B to E all show indirect effects and illustrate that younger respondents associated less strongly with Democrats in earlier waves than in later waves (Figure 11B), but also that White individuals have a stronger positive age effect than do non-White individuals (Figure 11C). We see similarly interesting dynamics when interacting sex and race, which show that the sex effect is more pronounced for Black respondents but is less pronounced for others (Figure 11D). The effect of income shows similarly complicated dynamics, where higher income is associated more strongly with Republican identification for White individuals than for non-White individuals (Figure 11E). Adding race interactions for sex, income, and age as implied by the Shapley values improves the fit of the hypothesized model using conventional in-sample fit statistics.[42]

DISCUSSION

This article set out to address a key problem in quantitative sociology: that we do not know what the appropriate functional form might be to model a dataset. In addition, a historic preference for parsimonious and easy-to-estimate models due to computational limitations has led to researchers hypothesizing relatively simplistic functional forms with little scrutiny. Not only does this lead to risks of misspecification bias and incorrect inference, but a lack of emphasis on appropriate specification can lead to researcher degrees of freedom muddying empirical findings. I argue that instead of trusting researchers to conjure up the correct functional form, we should use methods that embrace the fundamental uncertainty regarding the underlying patterns in a dataset to help sociologists improve their model-building.

I proposed a framework using ML methods to generate a data-driven estimate of the fit potential in a dataset. This fit potential indicates how well an outcome could be modeled when the functional relationship among variables is dictated by the data. Such an estimate provides an indication of whether a researcher's own functional form might miss important nuances like interactions or nonlinearities. Crucially, the fit potential is a feature of the data and not a result of the researcher's choices, improving transparency in the empirical process. Whenever the ML method finds more intricate patterns in the data than our own models do, we can unpack the former to provide guidance on how to improve the latter. This is contrary to popular belief that ML models are fundamentally black boxes that cannot convey any intuition into the patterns they identify. More generally, the proposed framework provides a bridge between the ability of ML methods to identify intricate patterns in data and a desire for interpretable models. By incorporating existing methods into the standard empirical workflow, ML models

become complementary tools in the sociologist's empirical toolkit, as opposed to the near exclusive use of ML for predictive questions, as is the status quo in sociology.

Illustrating the framework, I showed how the historic process of model-development could have been sped up considerably, using the example of the Mincerian wage equation. The framework effortlessly identified underspecification, and subsequent analysis of the ML models identified the necessary improvements to the functional form. The other two empirical examples, a hedonic regression of house prices and a model explaining party identification, both illustrate how often-encountered modeling strategies lead to considerably lower fit than does a flexible ML model. In the case of house prices in London, a number of simple nonlinearities were identified. Most importantly, the fundamental inability to address spatial heterogeneity through the inclusion of standard neighborhood characteristics became clear through the considerable differences in fit between the hypothesized and ML method when including neighborhood data. For party identification, a lack of complexity was similarly evident from applying the framework, and unpacking the flexible models showed important interactions and nonlinearities between the explanatory variables and outcome, which are not commonly implemented in the functional forms used to study party identification.

Existing misspecification tests have known limitations, but appropriate use would have identified a lack of specification in some of the examples presented in this article.[43] However, classic misspecification tests are known to be underpowered, especially in multivariate settings, and limited in the types of misspecification they consider. Fully non-parametric tests are more flexible, but suffer heavily from the curse of dimensionality. In addition, misspecification tests do not provide clear guidance on how to improve a functional form after misspecification is identified. Perhaps most importantly, use of even the most basic of misspecification tests is practically nonexistent in sociological work, and its selective implementation can suffer from the very same researcher degrees of freedom they are meant to address. Conversely, the only source of selectivity in the proposed framework is what models to include in a Super Learner to provide an estimate of the fit potential. As the price of considering a large amount of models is small, this risk is minimal and we can simply include a large amount of different types of methods. In contrast to classic misspecification tests, the proposed framework also provides researchers with concrete guidance into the type of patterns that may have been missed.

The approach presented here also has limitations. First, uncovering intricate patterns will still depend on the available data and might provide limited guidance in low N settings, although many of the ML methods proposed here can be applied to datasets typically encountered in sociology. Second, although considerable progress has been made in recent years, unpacking flexible ML methods remains challenging. This is in part because patterns can become complex to the degree that even local explanation methods like Shapley values will not yield easily digestible insights into the underlying functional form, especially when multiple interactions may be at play. More generally, active debate regarding the practical and philosophical aspects of explanation methods' accuracy to reflect how ML methods operate remains ongoing. However, the constant developments in the X-AI field are reassuring and exciting. Finally, when causal

questions rather than optimal fit are of interest, there is limited guidance on whether a difference in model fit between a hypothesized model and ML alternative is actually problematic. However, as the framework follows the same inferential curation of variables as a researcher's own model, missed functional relationships among variables should be expected to affect inference.

Perhaps the most challenging part of the framework is determining how much improvement in fit is enough to warrant a re-evaluation of the functional form. Unfortunately, this question will fundamentally depend on a combination of the substantive research question being asked and the true underlying DGP. As a result, the choice to accept a functional form will likely remain a debate among the academic community. This is not so much a consequence of the framework, but rather one of embracing the fact that we know very little about the true DGP and are unwilling to make stringent assumptions on it, nor blindly trust that a researcher-hypothesized model includes all the relevant intricacies in the data. A result of increasingly letting go of assumptions regarding the underlying DGP will be that we are left more frequently with questions like the one posed above, where we can rely less on statistical guidance and will instead have to rely on academic discussion.

At the root of most empirical sociological findings lies a functional form that is assumed to be correctly specified. Limited evaluation of whether this functional form is appropriate for the data leads to a number of serious risks. The model might not accurately reflect the patterns in the data, affecting the validity of statistical inference. Simply allowing researchers to report a single or curated number of functional forms without much scrutiny further exposes sociology to *p*-hacking. These practices stem from a time when limitations on computational power necessitated parsimony. These constraints are no longer applicable, yet the models we estimate retain a simplicity that likely belies the intricacies of the social mechanisms we are interested in. This is evidenced in the work of our qualitative colleagues, as well as the ever-increasing number of empirical examples where ML models outperform the linear additive models usually estimated by sociologists. I proposed a framework to address this issue by exploiting the benefits of ML methods—to find intricate patterns in data—to address this key issue throughout empirical work. This symbiosis of quantitative sociological work and the computational riches of today is long overdue, and it will only bear more fruit as the goals of the ML community increasingly align with the explanatory focus of sociologists.

APPENDIX

Shapley values

In game theory, Shapley values are defined as follows. Consider a game $\mathcal{G}$ consisting of $\mathcal{M}$ possible players. Further assume the game has some outcome $v(\mathcal{S})$ where $\mathcal{S} \subseteq \mathcal{M}$. Thus, there might be a subset of the total set of players who play the game. Shapley values were developed as a way to allocate each of the players in $\mathcal{S}$ a part of the game's outcome $v(\mathcal{S})$. Clearly, there are many ways to divide $v(\mathcal{S})$ among $\mathcal{S}$, but Shapley values are the only division that satisfy a number of properties discussed below. Shapley values are estimated as follows, where each player j gets their part of the payoff, equal to ϕ_j:

$$\phi_j(v(\mathcal{S})) = \phi_j = \sum_{\mathcal{S} \subseteq \mathcal{M}/j} \frac{|\mathcal{S}|!(M-|\mathcal{S}|-1)!}{M!}(v(\mathcal{S} \cup j) - v(\mathcal{S})), \quad j = 1, ..., M. \tag{A1}$$

The Shapley value ϕ_j for player j should be seen as a weighted sum of the differences in the outcome of the game $v(\cdot)$ given some set of players S when including player j to the set of players, versus the payoff without including player j.

Take as an example $\mathcal{M} = [1, 2, 3, 4]$, then the exact Shapley values are:

$$\begin{aligned}
\phi_1 &= \tfrac{1}{4}(v([1,2,3,4]) - v([2,3,4])) + \tfrac{1}{12}[(v([1,3,4]) - v([3,4])) + \\
&\quad (v([1,2,4]) - v([2,4])) + (v([1,2,3]) - v([2,3])) + (v([1,4]) - v([4])) + \\
&\quad (v([1,2]) - v([2])) + (v([1,3]) - v([3]))] + \tfrac{1}{4}(v([1]) - v([\oslash])) \\
\phi_2 &= \tfrac{1}{4}(v([1,2,3,4]) - v([1,3,4])) + \tfrac{1}{12}[(v([2,3,4]) - v([3,4])) + \\
&\quad (v([1,2,4]) - v([1,4])) + (v([1,2,3]) - v([1,3])) + (v([1,2]) - v([1])) + \\
&\quad (v([2,4]) - v([4])) + (v([2,3]) - v([3]))] + \tfrac{1}{4}(v([2]) - v([\oslash])) \\
\phi_3 &= \tfrac{1}{4}(v([1,2,3,4]) - v([1,2,4])) + \tfrac{1}{12}[(v([2,3,4]) - v([2,4])) + \\
&\quad (v([1,3,4]) - v([1,4])) + (v([1,2,3]) - v([1,2])) + (v([1,3]) - v([1])) + \\
&\quad (v([2,3]) - v([2]))] + \tfrac{1}{4}(v([3]) - v([\oslash])) \\
\phi_4 &= \tfrac{1}{4}(v([1,2,3,4]) - v([1,2,3])) + \tfrac{1}{12}[(v([2,3,4]) - v([2,3])) + \\
&\quad (v([1,3]) - v([1,3,4])) + (v([1,2,4]) - v([1,2])) + (v([1,4]) - v([1])) + \\
&\quad (v([2,4]) - v([2]))] + \tfrac{1}{4}(v([4]) - v([\oslash])).
\end{aligned} \tag{A2}$$

By adding all four Shapley values together and defining $\phi_0 = v(\oslash)$, we see that $\phi_0 + \phi_1 + \phi_2 + \phi_3 + \phi_4 = v([1, 2, 3, 4])$. In other words, the Shapley values sum to the outcome of the game. This is the "completeness" property. Three other properties of Shapley values are:

1. If $v(\mathcal{S} \cup i) = v(\mathcal{S} \cup j)$ for all coalitions $\mathcal{S}$ then $\phi_i = \phi_j$.
2. If including player j to the coalition $\mathcal{S}$ never changes $v(\mathcal{S})$ then $\phi_j = 0$.
3. If two games with outcomes $v(\cdot)$ and $w(\cdot)$ are combined then $\phi_j(v + w) = \phi_j(v) + \phi_j(w)$ and for any real number a, $\phi_j(av) = a\phi_j$.

As shown by Shapley (1953) and Young (1985), the Shapley values as defined in Equation A1 are the only distribution function for which these properties hold.

To adapt Shapley values as a method to explain a prediction $\hat{y}_i$ made by some model $\hat{f}(\cdot)$, we exchange $v(\mathcal{S})$ with $\hat{f}(x_i)$ and define:

$$\hat{f}(x_i) = \phi_0 + \sum_{k=1}^{K} \phi_{ik}, \tag{A3}$$

where $\phi_0 = \mathrm{E}(\hat{f}(x))$. In other words, the Shapley values ϕ_{ik} capture the difference between an individual prediction $\hat{f}(x_i)$ and the overall mean prediction ϕ_0. As was the case for the game theoretic setup, Shapley values for predictions are the only additive method that satisfy the above-mentioned properties (Lundberg and Lee 2017).

To calculate Shapley values for the prediction case, we substitute $\hat{f}(\cdot)$ for $v(\cdot)$ where we consider subsets $\mathcal{S} \subseteq \mathcal{K}$ of covariates instead of players. In other words, each player's contribution to the outcome in the game theoretical setup now represents each covariate's contribution to the prediction of a model. This requires evaluating the function $\hat{f}(\cdot)$ while excluding some covariates from the model. That is, replacing $v(\mathcal{S})$ in Equation A1 with $\mathrm{E}[\hat{f}(x|x = x_{\mathcal{S}})]$.

Generally speaking, calculating exact Shapley values is computationally challenging as the number of sets $\mathcal{S}$ is $2^{|\mathcal{K}|}$.

Kernel SHAP

Lundberg and Lee (2017) developed a computationally efficient way of estimating Shapley values. Their approach, Kernel SHAP, relies on reformulating the problem of estimating ϕ_j for $j = 0, \ldots, K$ as the following weighted least squares (WLS) problem:

$$\sum_{\mathcal{S} \subseteq \mathcal{K}} (\mathrm{E}[\hat{f}(x_i | x_i = x_{i,\mathcal{S}})] - (\phi_0 + \sum_{k \in \mathcal{S}} \phi_{ik}))^2 \mathrm{k}(\mathcal{K}, \mathcal{S}). \tag{A4}$$

In this setup, the function $\mathrm{k}(\cdot)$ reflects the Shapley kernel weights: $\mathrm{k}(\mathcal{K}, \mathcal{S}) = \frac{|\mathcal{K}|-1}{\binom{|\mathcal{K}|}{|\mathcal{S}|} |\mathcal{S}|(|\mathcal{K}|-|\mathcal{S}|)}$.[44] If we define $\mathbf{Z}$ to be a $2^{\mathcal{K}} \times (\mathcal{K}+1)$ matrix consisting of zeroes and ones indicating the inclusion of covariate k into the set $\mathcal{K}$ with a leading column of ones for ϕ_0, and if we then define $\mathbf{W}$ to be a $2^{\mathcal{K}} \times 2^{\mathcal{K}}$ matrix with the Shapley kernel weights defined above, and finally $\mathbf{f}$ to be a vector containing $\hat{f}_i(\mathcal{K}) = \mathrm{E}[\hat{f}(x_i | x_i = x_{i,\mathcal{S}})]$ for all $\mathcal{S}$ and ordered in line with $\mathbf{Z}$, the setup in Equation A4 can be rewritten as:

$$(\mathbf{f} - \mathbf{Z}\phi)^{\mathrm{T}} \mathbf{W} (\mathbf{f} - \mathbf{Z}\phi), \tag{A5}$$

which leads to the standard WLS solution $\phi = (\mathbf{Z}^{\mathrm{T}}\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{W}\mathbf{f}$. The benefit of this formulation is that the matrix multiplication $\mathbf{R} = (\mathbf{Z}^{\mathrm{T}}\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{W}$ only has to be calculated. Calculating $\mathbf{R}$ still requires a considerable matrix inversion, but by sampling a subset $\mathcal{K}'$ of $\mathcal{K}$ according to the Shapley kernel weights and including only the associated elements in $\mathbf{f}'$ and $\mathbf{Z}'$ allows for a considerably more computationally tractable approximation to Equation A4. This is Lundberg and Lee's (2017) first key innovation; it makes Kernel SHAP a computationally feasible approximation to the exact Shapley values in Equation A1.

The second component required to calculate Shapley values is the vector $\mathbf{f}$, which consists of the function evaluation under all $2^{\mathcal{K}}$ possible combinations of covariates—or $2^{\mathcal{K}'}$ in case of the sampling scheme mentioned above. To assess $\hat{f}(\mathcal{S}) = \mathrm{E}[\hat{f}(x_i | x_i = x_{i,\mathcal{S}})]$, one approach is to integrate out all covariates that are not included in $\mathcal{S}$ from the expectation:

$$\mathrm{E}[\hat{f}(x_i) | x_i = x_{i,\mathcal{S}}] = \int \hat{f}(x_{i,\bar{\mathcal{S}}}, x_{i,\mathcal{S}}) \mathrm{p}(x_{i,\bar{\mathcal{S}}} | x = x_{i,\mathcal{S}}) d\mathbf{x}_{i,\bar{\mathcal{S}}}, \tag{A6}$$

where $\mathrm{p}(x_{i,\bar{\mathcal{S}}} | x = x_{i,\mathcal{S}})$ is the conditional distribution of the covariates not included in $\mathcal{S}$, indicated with $\bar{\mathcal{S}}$. In principle, this density is conditional on the observed values of the covariates that *are* included in $\mathcal{S}$: $x_{i,\mathcal{S}}$. The standard Kernel SHAP by Lundberg and Lee (2017) assumes independence among covariates and thus removes the conditionality in the distribution and simply implements $p(x_{i,\bar{\mathcal{S}}})$, the empirical density of the variables. When such independence assumptions are unlikely, various alternative solutions have been developed to incorporate the conditionality between $x_{i,\mathcal{S}}$ and $x_{i,\bar{\mathcal{S}}}$ (Aas et al. 2021).

Tree SHAP

An alternative estimation approach to $\hat{f}(\mathcal{S}) = \mathrm{E}[\hat{f}(x_i|x_i = x_{i,\mathcal{S}})]$ is available for tree-based methods, Tree SHAP, from Lundberg et al. (2020). By exploiting tree structures, exact Shapley values can be calculated instead of relying on the sampling approach applied in Kernel SHAP. Whereas exact Shapley value calculation would require a summation over all possible sets $\mathcal{S}$, Tree SHAP only evaluates the decision paths for a set $x_{i,\mathcal{S}}$ throughout the tree and is thus effectively capped through the number of partitions in the tree. The basic approach to estimate $\hat{f}(\mathcal{S}) = \mathrm{E}[f(x_i|x_i = x_{i,\mathcal{S}})]$ using a tree is the following recursive algorithm:

Algorithm 1. Tree SHAP algorithm.

```
G(j) : = ;
Receive node j;
if node j is a leaf then
    return fitted value of leaf
else
    if node's split variable x_j ∈ x_S then
        if x_j <= s_j then
            return G(j_L)
        else
            return G(j_R)
        end
    else
        return (G(j_L)·r_L + G(j_R)·r_R) / (r_L + r_R)
    end
end
```

This algorithm $G(\cdot)$ is initialized at the root node of the tree and first evaluates whether the node is a leaf—that is, a terminal node. If so, the leaf estimate is simply returned as $\hat{f}(\mathcal{S})$. If the node is not a leaf, the algorithm assesses whether the split variable x_j used to split the node further into $[j_L, j_R]$ is in the set $\mathcal{S}$. If so, we evaluate whether the observed value of x_j for the observation for which a prediction is being explained lies in j_L or j_R, and the algorithm is repeated on the child node within which the observation falls. If $x_j \notin \mathcal{S}$ the algorithm is applied to both nodes, and a weighted average of both is used based on r_L or r_R, which are the proportions of data points falling in either of the two nodes. In other words, recursively applying $G(\cdot)$ the algorithm follows the effective tree structure until it reaches a node j for which the split variable is not part of $\mathcal{S}$ and from there onward uses a weighted average of $G(\cdot)$ applied to both nodes.

A key development in Lundberg et al. (2020) is a computationally efficient version of Algorithm 1 that elegantly stores information of paths that have already been followed within previous passes of the algorithm (see the discussion in Lundberg et al. 2020:65, and the source code documentation, located at https://github.com/slundberg/shap). The advantage of the Tree SHAP approach is that the assumption of covariate independence is relaxed in the process of determining $\hat{f}(\mathcal{S}) = \mathrm{E}[\hat{f}(x_i|x_i = x_{i,\mathcal{S}})]$, as the approach is conditional on the tree Π to generate estimates of $\hat{f}(\mathcal{S})$ while using the information available in $\mathcal{S}$ (cf. Aas et al. 2021).

Pairwise Shapley values

The dramatic decrease in computation time required to estimate Tree SHAP as compared to Kernel SHAP allows for the calculation of exact Shapley values, but also for calculation of local interaction effects through pairwise Shapley values. These pairwise Shapley values go beyond the more simple linear decomposition of a prediction along the K covariates as reflected in Equation A3. SHAP interaction values calculate the difference between a Shapley value ϕ_{ik} for some observation i when covariate j is present versus when it is absent. Specifically, it is estimated through:

$$\Phi_{i,j,k}(f,x) = \sum_{\mathcal{S}\subseteq K/k,\,j} = \frac{|\mathcal{S}|!(|\mathcal{K}|-|\mathcal{S}|-2)!)}{2(|\mathcal{K}|-1)!} \nabla_{i,j,k}(f,x,\mathcal{S}), \tag{A7}$$

for $j \neq k$ where

$$\nabla_{i,j,k}(f,x_i,\mathcal{S}) = f_{x_i}(\mathcal{S}\cup j,k) - f_{x_i}(\mathcal{S}\cup j) - f_{x_i}(\mathcal{S}\cup k) + f_{x_i}(\mathcal{S}). \tag{A8}$$

Here, $f_{x_i}(\cdot)$ is the standard formula to estimate a Shapley value as defined in Equation A1. Effectively, the direct Shapley value for variable j is $\Phi_{j,j}(f,x)$ and is defined as the standard Shapley value $\phi_j(f,x)$ minus the interaction values $\sum_{j\neq k}\Phi_{i,j,k}(f,x)$. The intuition behind A8 is that the pairwise Shapley value between variable k and variable j is a weighted average of all sets of covariates excluding k and j: $\mathcal{S}/\{j,k\}$. For each set, the Shapley value is first estimated when including *both* k and j—the first term in Equation A8. From this, the Shapley values when including *either* k or j are subtracted—the second and third terms in A8. Finally, the Shapley value from including *neither* k nor j is added—the last term in A8. Again, using the same efficient implementation as discussed above allows for computationally tractable estimation of these pairwise values (Lundberg et al. 2020:66).

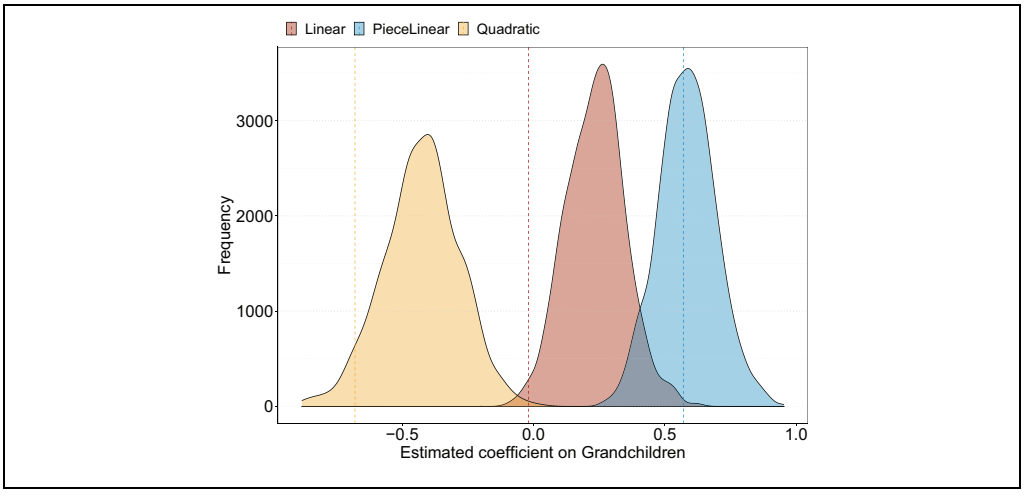


Figure A1. Bootstrapped distributions of the coefficient of interest from Step II of the framework.

Note: A natural result of estimating the fit of the hypothesized model over a large amount of resampled sets during the second step of the framework is that it allows for insights into the variability of the coefficients of the researcher-hypothesized model across samples. In this case, 80 percent of the data are randomly sampled 1,000 times to generate these distributions.

Table A1. Super Learner performance using the Toy Example data.

Model	Mean	Std. Dev.	Min.	Max.
SL.GLM_PieceLinear	1.218	0.060	1.046	1.320
SL.GB_200_1_0.1	1.229	0.059	1.107	1.327
SL.GB_500_1_0.1	1.231	0.059	1.104	1.323
SL.GB_100_1_0.1	1.241	0.059	1.129	1.351
SL.GB_100_2_0.1	1.242	0.060	1.113	1.342
SL.GB_500_3_0.01	1.247	0.060	1.122	1.357
SL.GB_100_3_0.1	1.252	0.061	1.098	1.349
SL.GB_200_2_0.1	1.253	0.061	1.107	1.354
SL.GB_500_2_0.01	1.254	0.059	1.139	1.364
SL.GB_500_4_0.01	1.259	0.060	1.120	1.364
SL.GB_500_5_0.01	1.268	0.061	1.125	1.364
SL.GB_200_3_0.1	1.268	0.062	1.101	1.351
SL.GB_100_4_0.1	1.271	0.062	1.103	1.362
SL.GB_500_6_0.01	1.275	0.061	1.116	1.376
SL.GB_100_5_0.1	1.279	0.061	1.094	1.366
SL.GB_500_2_0.1	1.284	0.063	1.120	1.367
SL.GB_100_6_0.1	1.286	0.062	1.110	1.363
SL.GB_200_4_0.1	1.292	0.063	1.107	1.381
SL.GB_200_5_0.1	1.297	0.062	1.106	1.385
SL.GB_500_3_0.1	1.302	0.063	1.116	1.393
SL.GB_200_6_0.1	1.313	0.063	1.132	1.394
SL.GB_500_4_0.1	1.330	0.064	1.139	1.419
SL.GB_500_5_0.1	1.341	0.064	1.157	1.440
SL.GB_500_6_0.1	1.357	0.065	1.198	1.443
SL.GB_500_1_0.01	1.366	0.062	1.255	1.503
SL.RF_200_2	1.383	0.064	1.229	1.544
SL.GLM_Quadratic	1.384	0.063	1.174	1.536
SL.RF_200_4	1.393	0.066	1.224	1.538
SL.RF_500_4	1.394	0.066	1.220	1.545
SL.RF_500_2	1.398	0.065	1.233	1.578
SL.RF_100_4	1.398	0.066	1.229	1.552
SL.RF_100_2	1.408	0.067	1.250	1.675
SL.GLM_Linear	1.413	0.065	1.321	1.569
SL.RF_200_1	2.342	0.098	2.029	2.671
SL.RF_100_1	2.394	0.100	2.106	2.815
SL.GB_200_3_0.01	2.696	0.089	2.477	2.919
SL.GB_200_4_0.01	2.697	0.088	2.483	2.925
SL.GB_200_5_0.01	2.698	0.088	2.484	2.926
SL.GB_200_6_0.01	2.698	0.088	2.486	2.927
SL.GB_200_2_0.01	2.725	0.093	2.488	2.952
SL.GB_200_1_0.01	2.911	0.114	2.567	3.188

Note: The models included in the Super Learner are a GB model with parameter grid: `ntrees` = [100, 200, 500], `max_depth` = [1, 2, 3, 4, 5, 6], shrinkage = [0.01, 0.1], and a Random Forest model with parameter grid: mtry = $[\sqrt{n},\ 2\sqrt{n}]$, ntree = [100, 200, 500]. Super Learner performance based on RMSE & 10-fold CV. Selected model in boldface. Hypothesized models in italics. GB parameters: `n_rounds_max_depth_eta`. RF parameters: `n_trees_mtry`.

Table A2. Descriptive statistics of the simulated wage data based on a synthetic sample of the 2018 General Social Survey and using four DGPs to generate four different outcome variables.

Variable	*N*	Mean	Std. Dev.	Min.	Pctl. 25	Pctl. 75	Max.
Age	50,000	41.09	13.70	18	29	53	65
Schooling (years)	50,000	13.79	2.97	0	12	16	20
Work experience (years)	50,000	23.01	13.7	0	11	35	51
Sex: Female	50,000	50.1%	-	-	-	-	-
Sex: Male	50,000	49.9%	-	-	-	-	-
Log wages (Linear-I)	50,000	8.29	1.25	5.24	7.238	9.38	10.96
Log wages (Linear-II)	50,000	7.58	0.69	5.20	7.066	8.09	9.17
Log wages (Linear-III)	50,000	6.74	0.64	4.62	6.254	7.30	7.99
Log wages (Linear-IV)	50,000	6.70	0.80	3.73	6.325	7.16	8.56

Table A3. Super Learner performance, Mincerian wage simulation Linear-I and Linear-II.

Model	Mean	Std. Dev.	Min.	Max.
Mincerian wage simulation Linear-I.				
SL.GLM_Linear	0.12	0.00	0.12	0.12
SL.GB_100_3_0.1	0.12	0.00	0.12	0.13
SL.GB_200_3_0.1	0.12	0.00	0.12	0.12
SL.GB_500_3_0.1	0.12	0.00	0.12	0.12
SL.GB_100_4_0.1	0.12	0.00	0.12	0.12
SL.GB_200_4_0.1	0.12	0.00	0.12	0.12
SL.GB_500_4_0.1	0.12	0.00	0.12	0.12
SL.GB_100_5_0.1	0.12	0.00	0.12	0.12
SL.GB_200_5_0.1	0.12	0.00	0.12	0.12
SL.GB_500_5_0.1	0.12	0.00	0.12	0.12
SL.GB_500_5_0.01	0.13	0.00	0.13	0.14
SL.GB_500_4_0.01	0.13	0.00	0.13	0.14
SL.GB_500_3_0.01	0.14	0.00	0.14	0.14
SL.GB_200_4_0.01	1.07	0.00	1.07	1.08
SL.GB_200_5_0.01	1.07	0.00	1.07	1.08
SL.GB_200_3_0.01	1.08	0.00	1.08	1.09
SL.RF_3_200	0.47	0.00	0.46	0.48
SL.RF_3_500	0.47	0.00	0.46	0.49
SL.RF_1_100	0.48	0.00	0.47	0.50
SL.RF_3_100	0.48	0.00	0.46	0.51
SL.RF_1_200	0.48	0.00	0.45	0.50
SL.RF_1_500	0.48	0.00	0.45	0.49
SL.GB_100_3_0.01	2.90	0.01	2.90	2.92
SL.GB_100_4_0.01	2.90	0.01	2.89	2.92
SL.GB_100_5_0.01	2.90	0.01	2.89	2.91
Mincerian wage simulation Linear-II.				
SL.GLM_Linear	0.07	0.00	0.07	0.07
SL.GB_100_3_0.1	0.07	0.00	0.07	0.07
SL.GB_200_3_0.1	0.07	0.00	0.07	0.07
SL.GB_500_3_0.1	0.07	0.00	0.07	0.07
SL.GB_100_4_0.1	0.07	0.00	0.07	0.07
SL.GB_200_4_0.1	0.07	0.00	0.07	0.07
SL.GB_500_4_0.1	0.07	0.00	0.07	0.07
SL.GB_100_5_0.1	0.07	0.00	0.07	0.07
SL.GB_200_5_0.1	0.07	0.00	0.07	0.07
SL.GB_500_5_0.1	0.07	0.00	0.07	0.07
SL.GB_500_3_0.01	0.09	0.00	0.09	0.09

(continued)

Table A3. (Continued)

Mincerian wage simulation Linear-II.				
SL.GB_500_4_0.01	0.09	0.00	0.08	0.09
SL.GB_500_5_0.01	0.09	0.00	0.08	0.09
SL.RF_3_100	0.26	0.00	0.25	0.28
SL.RF_1_200	0.26	0.00	0.26	0.28
SL.RF_3_200	0.26	0.00	0.26	0.28
SL.RF_1_100	0.27	0.00	0.25	0.28
SL.RF_1_500	0.27	0.00	0.25	0.28
SL.RF_3_500	0.27	0.00	0.26	0.27
SL.GB_200_3_0.01	0.96	0.00	0.96	0.97
SL.GB_200_4_0.01	0.96	0.00	0.96	0.96
SL.GB_200_5_0.01	0.96	0.00	0.95	0.96
SL.GB_100_3_0.01	2.61	0.00	2.60	2.62
SL.GB_100_4_0.01	2.61	0.00	2.60	2.61
SL.GB_100_5_0.01	2.61	0.00	2.60	2.61

Note: The models included in the Super Learner are a GB model with parameter grid: `ntrees` = [100, 200, 500], `max_depth` = [3, 4, 5], `shrinkage` = [0.01, 0.1], and a Random Forest model with parameter grid: `mtry` = $[\sqrt{n}, 2\sqrt{n}]$, `ntree` = [100, 200, 500]. Super Learner performance based on RMSE & 10-fold CV. Selected model in boldface. True model in italics. GB parameters: `n_rounds_max_depth_eta`. RF parameters: `n_trees_mtry`.

Table A4. Super Learner performance, Mincerian wage simulation Linear-III and Linear-IV.

Model	Mean	Std. Dev.	Min.	Max.
Mincerian wage simulation Linear-III.				
SL.GLM_Linear	0.07	0.00	0.07	0.07
SL.GB_100_3_0.1	0.07	0.00	0.07	0.07
SL.GB_200_3_0.1	0.07	0.00	0.07	0.07
SL.GB_500_3_0.1	0.07	0.00	0.07	0.07
SL.GB_100_4_0.1	0.07	0.00	0.07	0.07
SL.GB_200_4_0.1	0.07	0.00	0.07	0.07
SL.GB_500_4_0.1	0.07	0.00	0.07	0.07
SL.GB_100_5_0.1	0.07	0.00	0.07	0.07
SL.GB_200_5_0.1	0.07	0.00	0.07	0.07
SL.GB_500_5_0.1	0.07	0.00	0.07	0.07
SL.GB_500_5_0.01	0.08	0.00	0.08	0.08
SL.GB_500_4_0.01	0.08	0.00	0.08	0.08
SL.GB_500_3_0.01	0.09	0.00	0.08	0.09
SL.RF_1_100	0.24	0.00	0.23	0.26
SL.RF_3_100	0.24	0.00	0.23	0.24
SL.RF_1_200	0.24	0.00	0.22	0.26
SL.RF_3_200	0.24	0.00	0.23	0.25
SL.RF_1_500	0.24	0.00	0.23	0.25
SL.RF_3_500	0.24	0.00	0.23	0.25
SL.GB_200_3_0.01	0.85	0.00	0.85	0.85
SL.GB_200_4_0.01	0.85	0.00	0.84	0.85
SL.GB_200_5_0.01	0.85	0.00	0.84	0.85
SL.GB_100_4_0.01	2.30	0.00	2.30	2.30
SL.GB_100_3_0.01	2.30	0.00	2.30	2.31
SL.GB_100_5_0.01	2.30	0.00	2.30	2.30
Mincerian wage simulation Linear-IV.				
SL.GB_200_3_0.1	0.09	0.00	0.09	0.09
SL.GB_500_3_0.1	0.09	0.00	0.09	0.09
SL.GB_100_4_0.1	0.09	0.00	0.09	0.09
SL.GB_200_4_0.1	0.09	0.00	0.09	0.09

(continued)

Table A4. (Continued)

Mincerian wage simulation Linear-IV.				
SL.GB_500_4_0.1	0.09	0.00	0.09	0.09
SL.GB_100_5_0.1	0.09	0.00	0.09	0.09
SL.GB_200_5_0.1	0.09	0.00	0.09	0.09
SL.GB_500_5_0.1	0.09	0.00	0.09	0.09
SL.GB_100_3_0.1	0.10	0.00	0.10	0.10
SL.GB_500_5_0.01	0.10	0.00	0.10	0.10
SL.GB_500_4_0.01	0.11	0.00	0.11	0.11
SL.GB_500_3_0.01	0.13	0.00	0.12	0.13
SL.RF_1_100	0.30	0.00	0.29	0.33
SL.RF_3_100	0.30	0.00	0.28	0.32
SL.RF_1_200	0.31	0.00	0.29	0.32
SL.RF_3_200	0.30	0.00	0.29	0.32
SL.RF_1_500	0.30	0.00	0.28	0.33
SL.RF_3_500	0.31	0.00	0.29	0.32
SL.GLM_Linear	0.44	0.00	0.44	0.45
SL.GB_200_4_0.01	0.85	0.00	0.85	0.86
SL.GB_200_5_0.01	0.85	0.00	0.84	0.85
SL.GB_200_3_0.01	0.87	0.00	0.86	0.87
SL.GB_100_3_0.01	2.30	0.00	2.30	2.31
SL.GB_100_4_0.01	2.30	0.00	2.29	2.30
SL.GB_100_5_0.01	2.29	0.00	2.29	2.30

Note: The models included in the Super Learner are a GB model with parameter grid: `ntrees` = [100, 200, 500], `max_depth` = [3, 4, 5], `shrinkage` = [0.01, 0.1], and a Random Forest model with parameter grid: `mtry` = $[\sqrt{n}, 2\sqrt{n}]$, `ntree` = [100, 200, 500]. Super Learner performance based on RMSE & 10-fold CV. Selected model in boldface. Hypothesized model in italics. GB parameters: `n rounds max depth eta`. RF parameters: n trees mtry.

Table A5. Descriptive statistics of the London house price data.

Variable	*N*	Mean	Std. Dev.	Min.	Pctl. 25	Pctl. 75	Max.
Log House Price	629,669	12.92	0.63	9.21	12.49	13.25	17.44
House size (square meter)	629,669	88.89	46.45	10	60	104	959.74
Number of rooms	629,669	4.17	1.71	1	3	5	20
Distance to local center (minutes)	629,669	11.52	4.79	5	8.87	13.28	60.74
Crime and disorder index	629,669	0.35	0.57	–2.01	–0.05	0.75	2.17
Deprivation index	629,669	21.72	11.85	1.7	12.17	29.75	66.21
Property type: Detached	629,669	0.06	-	-	-	-	-
Property type: Flats/ Maisonettes	629,669	0.44	-	-	-	-	-
Property type: Semi-Detached	629,669	0.18	-	-	-	-	-
Property type: Terraced	629,669	0.33	-	-	-	-	-
Transaction type: Marketed sale	629,669	0.76	-	-	-	-	-
Transaction type: Private rental	629,669	0.17	-	-	-	-	-
Transaction type: Non-marketed sale	629,669	0.01	-	-	-	-	-
Transaction type: Marketed Sale	629,669	0.76	-	-	-	-	-
Transaction type: Other	629,669	0.05	-	-	-	-	-
New property	629,669	0.01	-	-	-	-	-
Year	629,669	2015	2.40	2011	2013	2017	2019
Month	629,669	6.61	3.36	1	4	9	12

Source: Reshare Project, UK Data Services, Department for Transport (2011) and Ministry of Housing, Communities & Local Government (2010).

Table A6. Super Learner performance, London house price data.

Model	Mean	Std. Dev.	Min.	Max.
SL.GB_1500_7_0.3	0.227	0.001	0.226	0.228
SL.GB_1750_7_0.3	0.227	0.001	0.224	0.229
SL.GB_1250_6_0.3	0.227	0.001	0.226	0.228
SL.GB_1500_7_0.3	0.227	0.001	0.224	0.230
SL.GB_1250_7_0.3	0.227	0.001	0.227	0.228
SL.GB_1250_6_0.4	0.228	0.001	0.226	0.229
SL.GB_1250_7_0.3	0.228	0.001	0.227	0.229
SL.GB_1000_6_0.4	0.229	0.001	0.227	0.231
SL.GB_1000_7_0.3	0.229	0.001	0.228	0.231
SL.GB_1000_6_0.3	0.230	0.001	0.229	0.231
SL.GB_1250_6_0.5	0.231	0.001	0.229	0.232
SL.GB_1500_8_0.3	0.231	0.001	0.229	0.232
SL.GB_1250_8_0.3	0.231	0.001	0.229	0.232
SL.GB_1500_8_0.3	0.231	0.001	0.228	0.233
SL.GB_1250_7_0.4	0.231	0.001	0.230	0.232
SL.GB_1000_6_0.5	0.231	0.001	0.229	0.233
SL.GB_1000_7_0.4	0.231	0.001	0.230	0.233
SL.GB_1750_8_0.3	0.231	0.001	0.229	0.233
SL.GB_1250_8_0.3	0.231	0.001	0.230	0.233
SL.GB_1000_8_0.3	0.231	0.001	0.230	0.233
SL.GB_1500_7_0.4	0.232	0.001	0.229	0.233
SL.GB_1750_7_0.4	0.232	0.001	0.230	0.234
SL.GB_750_7_0.4	0.232	0.001	0.231	0.234
SL.GB_750_6_0.5	0.233	0.001	0.231	0.234
SL.GB_750_8_0.3	0.233	0.001	0.232	0.235
SL.GB_750_6_0.4	0.233	0.001	0.231	0.235
SL.GB_750_7_0.3	0.233	0.001	0.231	0.235
SL.GB_1250_9_0.3	0.235	0.001	0.233	0.236
SL.GB_1500_9_0.3	0.235	0.001	0.234	0.236
SL.GB_750_7_0.5	0.235	0.001	0.234	0.236
SL.GB_1000_7_0.5	0.236	0.001	0.234	0.237
SL.GB_1250_7_0.5	0.236	0.001	0.235	0.237
SL.GB_1000_8_0.4	0.237	0.001	0.235	0.237
SL.GB_750_8_0.4	0.237	0.001	0.235	0.238
SL.GB_1250_8_0.4	0.237	0.001	0.236	0.238
SL.GB_1500_8_0.4	0.237	0.001	0.236	0.239
SL.GB_750_6_0.3	0.237	0.001	0.236	0.238
SL.GB_1750_8_0.4	0.238	0.001	0.236	0.239
SL.GB_500_6_0.4	0.242	0.001	0.239	0.244
SL.GB_750_8_0.5	0.242	0.001	0.240	0.244
SL.GB_1000_8_0.5	0.243	0.001	0.241	0.244
SL.GB_1250_8_0.5	0.244	0.001	0.242	0.245
SL.GLM	0.421	0.002	0.419	0.423

Note: The models included in the Super Learner are a GB model with parameter grid: `ntrees` = [100, 200, 500, 750, 1000, 1250, 1500, 1750], `max_depth` = [4, 5, 6, 7, 8, 9], `shrinkage` = [0.3, 0.4, 0.5], and a Random Forest model with parameter grid: `mtry` = $[\sqrt{n},\ 2\sqrt{n}]$, `ntree` = [200, 500, 1000]. Results from the Random Forest models are omitted for brevity, as are results from the Gradient Boosting models for `ntrees` = [100, 200] and `max_depth` = [4, 5]. None of the omitted models featured among the better-performing models. Super Learner performance based on RMSE & 10-fold CV. Selected model in boldface. Hypothesized model in italics. GB parameters: `n_rounds_max_depth_eta`. RF parameters: `n_trees_mtry`.

Table A7. Selected model from the Super Learner routine when applied to subsets of the full covariate set for the London house price data.

Covariate subset	Selected model	Mean RMSE (linear model)	Mean RMSE (flexible model)
Area	SL.GB_100_4_0.1	0.463	0.447
Temp	SL.GB_500_2_0.3	0.610	0.609
Area + Temp	SL.GB_100_4_0.1	0.439	0.421
Area + Temp + Housing	SL.GB_100_4_0.1	0.437	0.419
Area + Temp + Housing + Transaction	SL.GB_100_4_0.1	0.437	0.419
Area + Temp + Housing + Transaction + IMD	SL.GB_1000_6_0.3	0.431	0.341
Area + Temp + Housing + Transaction + Crime	SL.GB_750_4_0.1	0.432	0.404
Area + Temp + Housing + Transaction + Travel	SL.GB_1000_6_0.3	0.437	0.305
Area + Temp + Housing + Transaction + Travel + Crime	SL.GB_1000_6_0.3	0.431	0.272
Area + Temp + Housing + Transaction + Travel + IMD	SL.GB_1000_6_0.3	0.430	0.265
Area + Temp + Housing + Transaction + IMD + Crime	SL.GB_1000_6_0.3	0.431	0.274

Note: The models included in the Super Learner are a GB model with parameter grid: `ntrees` = [100, 200, 500, 750, 1000, 1250, 1500, 1750], `max_depth` = [4, 5, 6, 7, 8], `shrinkage` = [0.1, 0.2, 0.3, 0.4, 0.5], and a Random Forest model with parameter grid: `mtry` = $[\sqrt{n},\ 2\sqrt{n}]$, `ntree` = [200, 500, 1000]. Super Learner performance based on RMSE & 10-fold CV. GB parameters: `n_rounds_max_depth_eta`. RF parameters: `n_trees_mtry`.

Table A8. Descriptive statistics of the GSS.

Variable	*N*	Mean	Std. Dev.	Min.	Pctl. 25	Pctl. 75	Max.
Party Identification	26,011	3.67	1.99	1	2	6	7
Age	26,011	45.31	17.14	18	31	58	89
Schooling (Years)	26,011	13.13	3.00	0	12	15	20
Income	26,011	10.19	2.74	1	9	12	12
Sex: Female	26,011	0.56	-	-	-	-	-
Sex: Male	26,011	0.44	-	-	-	-	-
Race: Black	26,011	0.14	-	-	-	-	-
Race: Other	26,011	0.05	-	-	-	-	-
Race: White	26,011	0.81	-	-	-	-	-
Survey Year	26,011	1996	13.46	1974	1987	2010	2018

Source: The General Social Survey (2018).

Table A9. Super Learner performance, GSS data.

Model	Mean	Std. Dev.	Min.	Max.
SL.GB_50_4_0.15	1.86	0.01	1.83	1.89
SL.GB_75_4_0.15	1.86	0.01	1.83	1.89
SL.GB_100_3_0.1	1.86	0.01	1.83	1.89
SL.GB_75_3_0.15	1.86	0.01	1.83	1.89
SL.GB_100_4_0.1	1.86	0.01	1.83	1.89
SL.GB_75_4_0.1	1.86	0.01	1.83	1.89
SL.GB_100_3_0.15	1.86	0.01	1.83	1.89
SL.GB_50_5_0.15	1.86	0.01	1.83	1.89
SL.GB_50_5_0.1	1.86	0.01	1.83	1.89
SL.GB_75_3_0.1	1.86	0.01	1.84	1.89
SL.GB_50_3_0.15	1.86	0.01	1.84	1.89
SL.GB_75_5_0.1	1.86	0.01	1.83	1.89
SL.GB_50_4_0.1	1.86	0.01	1.84	1.89
SL.GB_100_3_0.1	1.86	0.01	1.82	1.91
SL.GB_100_4_0.15	1.86	0.01	1.83	1.89
SL.GB_100_5_0.1	1.86	0.01	1.83	1.89
SL.GB_200_3_0.1	1.86	0.01	1.82	1.90
SL.GB_100_4_0.1	1.86	0.01	1.81	1.91
SL.GB_75_5_0.15	1.86	0.01	1.83	1.89
SL.GB_50_3_0.1	1.86	0.01	1.84	1.89
SL.GB_100_3_0.2	1.86	0.01	1.82	1.91
SL.GB_100_5_0.15	1.86	0.01	1.84	1.90
SL.GB_200_3_0.2	1.87	0.01	1.82	1.91
SL.GB_100_5_0.1	1.87	0.01	1.82	1.92
SL.GB_200_4_0.1	1.87	0.01	1.82	1.91
SL.GB_100_4_0.2	1.87	0.01	1.82	1.91
SL.RF_4_200	1.87	0.01	1.83	1.93
SL.RF_4_100	1.87	0.01	1.83	1.93
SL.RF_2_100	1.87	0.01	1.83	1.93
SL.RF_4_500	1.87	0.01	1.83	1.93
SL.RF_2_500	1.87	0.01	1.83	1.93
SL.RF_2_200	1.87	0.01	1.83	1.93
SL.GB_200_5_0.1	1.87	0.01	1.83	1.93
SL.GB_100_5_0.2	1.87	0.01	1.83	1.92
SL.GB_200_4_0.2	1.87	0.01	1.83	1.92
SL.GB_200_5_0.01	1.89	0.01	1.84	1.94
SL.GB_200_4_0.01	1.89	0.01	1.84	1.95
SL.GB_200_5_0.2	1.89	0.01	1.85	1.95
SL.GB_200_3_0.01	1.90	0.01	1.84	1.95
SL.GLM	1.92	0.01	1.90	1.95
SL.GB_100_5_0.01	2.04	0.02	1.98	2.10
SL.GB_100_4_0.01	2.05	0.02	1.98	2.11
SL.GB_100_3_0.01	2.05	0.02	1.98	2.11

Note: The models included in the Super Learner are a GB model with parameter grid: `ntrees` = [50, 75, 100, 200, 500], `max_depth` = [1, 2, 3, 4, 5, 6], `shrinkage` = [0.01, 0.1, 0.15, 0.2], and a Random Forest model with parameter grid: `mtry` = $[\sqrt{n},\ 2\sqrt{n}]$, `ntree` = [100, 200, 500]. Results from the Gradient Boosting models were omitted for brevity for `ntrees` = [500] and `max_depth` = [1, 2, 6]. None of the omitted models featured among the better-performing models. Super Learner performance based on RMSE & 10-fold CV. Selected model in boldface. Hypothesized model in italics. GB parameters: `n_rounds_max_depth_eta`. RF parameters: `n_trees_mtry`.

Table A10. Selected model from the Super Learner routine when applied to subsets of the full covariate set, GSS data.

Covariate subset	Selected model	Mean RMSE (linear model)	Mean RMSE (flexible model)
Year	SL.GB_600_2_0.01	1.985	1.978
Year + Age	SL.GB_500_3_0.01	1.985	1.974
Year + Age + Sex	SL.GB_500_3_0.01	1.980	1.969
Year + Age + Race	SL.GB_50_3_0.1	1.933	1.888
Year + Age + Sex + Race	SL.GB_50_2_0.2	1.927	1.881
Year + Age + Sex + Race + Income	SL.GB_50_3_0.15	1.919	1.873
Year + Age + Sex + Race + Education	SL.GB_200_3_0.05	1.921	1.865

Note: The models included in the Super Learner are a GB model with parameter grid: `ntrees` = [50, 75, 100, 200, 300, 400, 500, 600], `max_depth` = [1, 2, 3, 4, 5, 6], `shrinkage` = [0.01, 0.1, 0.15, 0.2], and a Random Forest model with parameter grid: `mtry` = $[\sqrt{n},\ 2\sqrt{n}]$, `ntree` = [100, 200, 500].

SIMULATIONS

Nonlinear DGPs

The four DGPs used to generate the true data in Figures 1 and 3 are as follows:

$$y_{\text{linear}} = -1 + 0.9X + \epsilon_1 \tag{A9}$$

$$y_{\text{polynomial}} = -6 + 0.4X - 0.36X^2 + 0.005X^3 + \epsilon_2 \tag{A10}$$

$$y_{\text{sine}} = 2.83\sin(\tfrac{\pi}{2X}) + \epsilon_3 \tag{A11}$$

$$y_{\text{step}} = -2\text{I}(X>0) + 4\text{I}(X>2) - \text{I}(X>3) + \epsilon_4 \tag{A12}$$

Where $X \sim \text{Unif}(-4, 4)$ and each error term is normally distributed and calibrated in such a way that the total variance of each outcome is 5.

Mincerian wage equation simulation

The four DGPs used to generate the four Mincerian wage outcomes in Figures 6 and 7 are the following:

Linear-I

$$\begin{aligned} \ln(\text{wages}) &= 4.5 + 0.125x_{\text{educ}} + 0.09x_{\text{exp}} + \epsilon \\ \epsilon &\sim \mathcal{N}(0, 0.07) \end{aligned} \tag{A13}$$

Linear-II

$$\begin{aligned} \ln(\text{wages}) &= 4.5 + 0.125x_{\text{educ}} + 0.09x_{\text{exp}} - 0.001x_{\text{educ}}^2 + \epsilon \\ \epsilon &\sim \mathcal{N}(0, 0.07) \end{aligned} \tag{A14}$$

Linear-III

$$\begin{aligned}\ln(\text{wages}) = {} & 4.5 + 0.02x_{\text{educ\_0\_8}} + 0.03x_{\text{educ\_9\_10}} + 0.3x_{\text{educ\_11\_12}} + \\ & 0.06x_{\text{educ\_13\_14}} + 0.06x_{\text{educ\_15+, i}} + 0.09x_{\text{exp}} + \\ & 0.001x_{\text{exp}}^2 + \epsilon \\ & \epsilon \sim \mathcal{N}(0, 0.07)\end{aligned} \tag{A15}$$

Linear-IV

$$\begin{aligned}\ln(\text{wages}) = {} & I(x_{\text{sex}} = \text{Female})[3.5 + 0.025x_{\text{educ\_0\_8}} + 0.06x_{\text{educ\_9\_10}} + \\ & 0.35x_{\text{educ\_11\_12}} + 0.06x_{\text{educ\_13\_14}} + 0.06x_{\text{educ\_15+}} + \\ & 0.1x_{\text{exp}} - 0.0005x_{\text{exp}}^2 + \epsilon] + \\ & I(x_{\text{sex}} = \text{Male})[5.5 + 0.015x_{\text{educ\_0\_8}} + 0.02x_{\text{educ\_9\_10}} + \\ & 0.25x_{\text{educ\_11\_12}} + 0.04x_{\text{educ\_13\_14}} + 0.04x_{\text{educ\_15+}} + \\ & 0.06x_{\text{exp}} - 0.001x_{\text{exp}}^2 + \epsilon^*]\end{aligned} \tag{A16}$$

$$\epsilon \sim \mathcal{N}(0, 0.09) \tag{A17}$$

$$\epsilon^* \sim \mathcal{N}(0, 0.08)$$

ORCID iD

Mark D. Verhagen https://orcid.org/0000-0003-2746-0309

Data Availability Statement

All code and publicly available data underlying the analyses in this article can be found at https://github.com/MarkDVerhagen/Functional_Form_Complexity.

Notes

1. Linear additive functional forms plugged into exponential family distributions have convenient properties that make estimation considerably more straightforward from a computational perspective than non-parametric or Bayesian approaches to inference (Efron and Hastie 2016). This historic limitation still dictates the dominance of the former approach in empirical work today.
2. Breiman (2001) describes the difference as "the two cultures in the use of statistical modeling." In the first, we assume the researcher knows the underlying functional form and simply has to estimate its parameters from data. In the second, we assume the functional form is fundamentally unknown and has to be learned from the data. As Breiman (2001:199) notes, we typically reside in the latter of the two worlds, although the majority of statistical work uses empirical approaches that assume we reside in the former.

3. The potential of ML-based methods to find intricate patterns in data has been illustrated in sociology (Baćak and Kennedy 2019; Brand et al. 2021), psychology (Agrawal et al. 2020), and behavioral work (Kleinberg et al. 2017; Peterson et al. 2021; Peysakhovich and Naecker 2017). Furthermore, their increased structure makes ML methods more applicable from a practical perspective than fully non-parametric regression, which suffers from exponential data requirements (Bishop 2006; Yatchew 1997).
4. Indicative of the predictive focus of ML is the discussion by Mullainathan and Spiess (2017), who explicitly call for the application of ML in predictive contexts or even reformulating explanatory questions into predictive ones. Similar assessments can be found in reviews by Molina and Garip (2019) and Shmueli (2010).
5. By following a set inferential logic (namely that of the original model), complexity in the proposed framework is introduced solely through the functional relationship among variables and distinguishes itself from "model complexity" as is often understood as the problem of selecting among a (large) set of potential covariates (Bucca and Urbina 2019).
6. Grimmer et al. (2021:406–408) discuss a similar iterative perspective on ML, envisioning a "human in the loop" approach, where ML models provide insights to the researcher but do not completely supplant the researcher in model-building.
7. In cases where there is no need to understand the underlying functional form (e.g., in the imputation of missing data), there is little need to combine both parametric and ML methods, and one can simply resort to the optimal model.
8. Choice of functional forms for this example were inspired by Polley, Rose, and Van der Laan (2011).
9. Assume some covariate X affects the outcome in a nonlinear way (e.g., through a polynomial relationship), then the exclusion of the higher-order terms effectively boils down to omitted variable bias.
10. An example of the restrictive nature of the White and RESET tests is that neither will pick up an interaction of some continuous variable x and a binary variable D if the (linear) slopes are mirrored for the interacting groups.
11. To illustrate, misspecification tests are generally able to identify the bivariate nonlinearities in Figure 1, although a standard RESET test assessing a possible second-degree polynomial fails to reject the step function and requires a third-degree polynomial. For the toy example, White's test fails to reject misspecification of the quadratic model.
12. One of the 32 papers used Huber-White standard errors to address possible heteroskedasticity of the error term. The coded articles are available from the author on request.
13. In the taxonomy laid out by Grimmer et al. (2021:412), the framework would still use variables that are relevant from an inferential perspective rather than those that give the best possible fit. Given these variables, though, we *are* interested in a model that fits the data well. Note that a different approach to ML methods is to identify relevant variables among a (large) set of possible covariates (Bucca and Urbina 2019; Grimmer et al. 2021). This is another fruitful application of ML into sociological model-building, but outside the scope of this article.
14. Neural networks with sigmoidal activation functions have been formally shown to approximate functions of arbitrary complexity given sufficient hidden units (Hornik, Stinchcombe, and White 1989). Ensembles including tree-based methods like Random Forests also approximate highly complex functions, although mathematical proof of universal approximation has proven complicated due to the piece-wise linear nature inherent in tree-based methods, as opposed to the continuous approximation of neural networks (Breiman 2001). However, Biau (2012), for example, provides theoretical properties of the Random Forest model suggestive of its flexibility to estimate highly complex functions.
15. Throughout the empirical examples in this article, I use the Super Learner approach as implemented in the R package `SuperLearner`; various alternatives exist in different programming languages (Polley et al. 2021).

16. In contrast to tree-based methods, for some ML approaches the optimal solution is dependent on the scale of the explanatory variables. Examples include least absolute shrinkage and selection operator regression that introduces an additional term to the squared error loss of a standard OLS regression, which reflects the absolute size of all coefficients $\hat{\beta}$. This effectively shrinks large coefficients that provide limited improvement to fit toward zero. Clearly, dividing or multiplying an explanatory variable x_{ik} by 100 would focus or divert attention of the algorithm on $\hat{\beta}_k$. Decision trees are scale invariant, as single binary splits are placed on individual explanatory variables.
17. Given a vector of predictions $\hat{y}$ and a vector of actually observed outcomes y the RMSE is defined as $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$.
18. Even within a conventional conditioning-on-observables approach, the use of a test set is generally recommended whenever a model is intermittently assessed during model-building, as is often the case in quantitative sociology (Cameron and Trivedi 2005).
19. Prior work has noted some limitations to cross-validation, specifically that the standard errors of the fit estimates from such routines can be biased downward (Vehtari, Gelman, and Gabry 2017). In case correct standard errors are of interest, a truly heldout set of data is preferred or post-estimation scaling of the standard errors can be applied (Bates, Hastie, and Tibshirani 2021).
20. By exposing the researcher-hypothesized model to cross-validation, bootstrapped standard errors of the coefficient estimates are calculated (Cameron and Trivedi 2005). In case a model has been overengineered to a sample at hand, this should be reflected by the empirical standard errors from reestimating the model to resampled datasets compared to those from estimating the researcher's preferred specification on the full data.
21. A case in point is the example brought up by Young (2019) of a study into the gender wage gap done at Google, which included job promotions as an explanatory variable. Clearly, including promotions would improve fit, but it would invalidate any inference, as it is a result (i.e., post-treatment) of the very process of interest. This example shows that a singular focus on fit is problematic, but it mostly emphasizes that any analysis of an inferential question still requires causal or inferential logic.
22. However, including relevant predictors and correct model specification will generally improve precision of estimation and is generally advised.
23. As mentioned, a complementary benefit of using a cross-validation approach is that the 1,000 repeated estimations of the hypothesized models provide bootstrapped standard errors of the coefficient estimates. These are visualized in Appendix Figure A1, together with the estimated coefficient on the full set. The reported coefficients when estimating the models to the full set lie at the extremes of the bootstrapped distribution of $\hat{\beta}$'s for the two misspecified models.
24. As a case in point, recent legislation in the European Union requires companies that use algorithms to support decision-making to have explainable algorithms, the so-called "Right to Explanation" (Goodman and Flaxman 2017).
25. The variety of perspectives in explainability have led some to call for a more rigorous reporting of the envisioned goals of generating explanations (Doshi-Velez and Kim 2017). Following the taxonomy of Doshi-Velez and Kim (2017), explanation in this framework is meant to generate understanding into the unknown patterns in the data, leading to the choice for local explainability at the model level. Conversely, temporal efficiency and alignment with operational processes, which sometimes constrain explainability when deployed as a decision support system, are not relevant in this setup.
26. Another consequence of the wide variety in perspectives on what matters in explanation is that there are few generally accepted quantitative metrics to assess explanation quality, and different methodological approaches can provide differing explanations. As explanation becomes more critical to ML research, such metrics and comparative approaches are increasingly being introduced into the field (Krishna et al. 2022). The uniqueness of Shapley values in providing a number of theoretical guarantees makes them especially attractive for generating model understanding as understood in this article.
27. The approach has many resemblances to the *do*-operator popularized by Pearl (2009), as it effectively concerns intervention on the covariate space (Heskes et al. 2020; Lundberg and Lee 2017).

28. Assume a decision tree is fit to a dataset with three covariates and the resultant model only splits the data on two of the three covariates. For this model, it is only necessary to consider information sets that differ in terms of the two covariates used in the splitting process; variation in the third omitted covariate will have no bearing on the model's behavior.
29. Interactions can still be identified without the use of pairwise Shapley values. This can be done by separately visualizing Shapley values for a variable x_{ik}, conditional on values of another variable x_{ij}. This approach is illustrated in the first empirical example.
30. The Shapley value for variable k reflects the effect of x_{ik} relative to ϕ_0. In linear regression including an intercept, interpretation is relative to the overall mean.
31. Typically in non-parametric regression, a general function $y_i = m_h(x_i)$ is estimated as a simple weighted average of observations close to the observed x_i in Euclidean space. Inclusion of data points is generally governed by a bandwidth parameter h defining a region $[x_i - h, x_i + h]$. Different weighting functions or kernels can be used to determine the importance of the data points within the interval, for example, by weighing points closer to x_i more than those further away (Yatchew 2003). As the dimensionality of X increases, the number of data points required increases exponentially, limiting the practical applicability of the approach— the so-called "curse of dimensionality" (Bishop 2006).
32. Some ML methods are similarly affected by the "curse of dimensionality." Notably, methods like neural networks require considerable amounts of data to estimate (Bishop 2006). This is the main reason why tree-based or shrinkage methods are usually applied within the social sciences, as such methods remain applicable for very small N cases (Athey 2018; Liao 2017).
33. Muñoz and Young (2018:8) explicitly contrast their approach to ML methods by stating that ML methods are "generally about model selection in high-dimensional space (where conventional approaches to variable selection are impractical—e.g., when there are thousands of potential predictor variables)." As the proposed framework shows, there is no reason to exclusively implement ML in high dimensional settings.
34. The focus of model robustness on variable inclusion raises questions of whether the various alternative specifications are equally plausible in practice (see O'Brien 2018; Slez 2019). For example, if an important confounder is omitted in 50 percent of the alternative "reasonable" specifications, we should not be surprised to see high variation in estimates of the coefficient of interest.
35. Young and Stewart (2021) have expanded model robustness to "Multiverse" analysis by, in addition to varying the number of variables in the model, including a number of "reasonable" functional forms to assess the degree to which different findings can be supported by data. Like model robustness, fit is of secondary importance, and the framework relies on an a priori specified set of reasonable functional forms. However, in spirit it is more closely aligned to the proposed framework, even though it remains fundamentally reliant on the researcher to determine what constitutes "plausible" specifications, rather than letting computational power and model fit dictate the patterns in the data.
36. Effectively, one would abort the framework after the second step and instead investigate which observations tend to be fitted better by the ML model. Salganik et al. (2020) and Verhagen (2022) suggest a similar approach to studying predictions, in which the predictive output of ML models is assessed.
37. The MARS algorithm takes a similar approach to finding better-specified models from a stacked specification by iteratively trimming variables from a full specification (Friedman 1991).
38. Code and instructions to obtain the data to reproduce each empirical example in this article, including the toy example, are available in this article's online GitHub repository (https://github.com/MarkDVerhagen/Functional_Form_Complexity).
39. I include these indices at the Lower Super Output Area level, for which the average population in London is around 1,700.
40. Adding squared versions of the neighborhood indicators, area, number of rooms, and temporal variables improves the adjusted R^2 by 2.5 pp. An LR test accordingly prefers the updated model ($p < 0.001$).

41. Adding square terms to all three variables and separate intercepts for each survey year improves the R^2 from 6.1 to 7.0 percent. Unsurprisingly, an LR test comparing both models prefers the updated model ($p < 0.001$).
42. Adding age by race and sex by race interactions further improves the adjusted R^2 to 7.2 percent. An LR test again prefers the updated model ($p < 0.001$). Finally, adding income by race improves the adjusted R^2 by another 0.1 pp, with an LR test preferring the added flexibility to the functional form ($p < 0.001$).
43. Classic White and RESET tests failed to identify all the misspecification across the toy example and simulations discussed in the Misspecification and Risks to Inference section (see note 11). The RESET test did identify misspecification of the standard linear model for the Housing and GSS examples, as should be expected given the considerable nonlinearities in the data.
44. The Shapley kernel weights are not defined for $k(\mathcal{K}, \mathcal{K})$ as the denominator is zero. This can be addressed by setting $k(\mathcal{K}, \mathcal{K})$ to some high constant or setting $\phi_0 = \hat{f}(\varnothing)$ and $\sum_{k=0}^{\mathcal{K}} = \hat{f}(\mathcal{K})$ as additional constraints instead (Lundberg and Lee 2017:6).

References

Aas, Kjersti, Martin Jullum, and Anders Løland. 2021. "Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values." *Artificial Intelligence* 298: 103502. https://doi.org/10.1016/j.artint.2021.103502.

Abbott, Andrew. 1988. "Transcending General Linear Reality." *Sociological Theory* 6(2):169–86.

Agrawal, Mayank, Joshua C. Peterson, and Thomas L. Griffiths. 2020. "Scaling Up Psychology via Scientific Regret Minimization." *Proceedings of the National Academy of Sciences* 117(16):8825–35.

Athey, Susan. 2018. "The Impact of Machine Learning on Economics." Pp. 507–47 in *The Economics of Artificial Intelligence: An Agenda*, edited by Agrawal, A., J. Gans, and A. Goldfarb. Chicago, IL: University of Chicago Press.

Baćak, Valerio, and Edward H. Kennedy. 2019. "Principled Machine Learning Using the Super Learner: An Application to Predicting Prison Violence." *Sociological Methods & Research* 48(3):698–721.

Bates, Stephen, Trevor Hastie, and Robert Tibshirani. 2021. "Cross-Validation: What Does It Estimate and How Well Does It Do It?" arXiv. https://doi.org/10.48550/arXiv.2104.00673

Beiser-McGrath, Janina, and Liam F. Beiser-McGrath. 2020. "Problems with Products? Control Strategies for Models with Interaction and Quadratic Effects." *Political Science Research and Methods* 8(4): 707–30.

Berk, Richard A. 2004. *Regression Analysis: A Constructive Critique*, Vol. 11. Thousand Oaks, CA: Sage.

Berk, Richard A., and Justin Bleich. 2013. "Statistical Procedures for Forecasting Criminal Behavior: A Comparative Assessment." *Criminology and Public Policy* 12(3):513–44.

Biau, Gérard. 2012. "Analysis of a Random Forests Model." *Journal of Machine Learning Research* 13: 1063–95.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. New York, NY: Springer.

Blackwell, Matthew, and Michael P. Olson. 2022. "Reducing Model Misspecification and Bias in the Estimation of Interactions." *Political Analysis* 30(4):495–514.

Brand, Jennie E., Xu Jiahui, Koch Bernard, and Geraldo Pablo. 2021. "Uncovering Sociological Effect Heterogeneity Using Tree-Based Machine Learning." *Sociological Methodology* 52(2):189–223.

Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24(2):123–40.

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16(3):199–231.

Bucca, Mauricio, and Daniela R. Urbina. 2019. "Lasso Regularization for Selection of Log-Linear Models: An Application to Educational Assortative Mating." *Sociological Methods & Research* 50(4): 1763–800.

Buja, Andreas, Lawrence Brown, Richard Berk, Edward George, Emil Pitkin, Mikhail Traskin, Kai Zhang, and Linda Zhao. 2019. "Models as Approximations I: Consequences Illustrated with Linear Regression." *Statistical Science* 34(4):523–44.

Buja, Andreas, Arun Kumar Kuchibhotla, Richard Berk, Edward George, Eric Tchetgen Tchetgen, and Linda Zhao. 2019. "Models as Approximations–Rejoinder." *Statistical Science* 34(4):606–20.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge, UK: Cambridge University Press.

Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System." Pp. 785–94 in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: Association for Computing Machinery.

Chi, Bin, Adam Dennett, Thomas Oléron-Evans, and Robin Morphet. 2021. "A New Attribute-Linked Residential Property Price Dataset for England and Wales, 2011 to 2019." *UCL Open: Environment* 3: e019.

Christensen, Björn, and Sören Christensen. 2014. "Are Female Hurricanes Really Deadlier Than Male Hurricanes?" *Proceedings of the National Academy of Sciences* 111(34):E3497–98.

Davis, James A., and Tom W. Smith. 1991. *The NORC General Social Survey: A User's Guide*. Thousand Oaks, CA: Sage.

Doornik, Jurgen A., and David F. Hendry. 2015. "Statistical Model Selection with 'Big Data.'" *Cogent Economics & Finance* 3(1):1045216.

Doshi-Velez, Finale, and Been Kim. 2017. "Towards a Rigorous Science of Interpretable Machine Learning." arXiv. https://doi.org/10.48550/arXiv.1702.08608

Dougherty, Michael R., Rick P. Thomas, Ryan P. Brown, Jeffrey S. Chrabaszcz, and Joe W. Tidwell. 2015. "An Introduction to the General Monotone Model with Application to Two Problematic Data Sets." *Sociological Methodology* 45(1):223–71.

Duncan, Otis Dudley. 1984. *Notes on Social Measurement: Historical and Critical*. New York, NY: Russell Sage Foundation.

Efron, Bradley, and Trevor Hastie. 2016. *Computer Age Statistical Inference*. Cambridge, UK: Cambridge University Press.

Elbers, Benjamin. 2023. "A Method for Studying Differences in Segregation across Time and Space." *Sociological Methods & Research* 52(1):5–42.

Elhorst, J. Paul. 2010. "Applied Spatial Econometrics: Raising the Bar." *Spatial Economic Analysis* 5(1): 9–28.

Fan, Gang-Zhi, Seow Eng Ong, and Hian Chye Koh. 2006. "Determinants of House Price: A Decision Tree Approach." *Urban Studies* 43(12):2301–15.

Freeden, Michael, Lyman Tower Sargent, and Marc Stears. 2013. *The Oxford Handbook of Political Ideologies*. Oxford, UK: Oxford University Press.

Freedman, David A. 2009. *Statistical Models: Theory and Practice*. Cambridge, UK: Cambridge University Press.

Freund, Yoav, and Robert E. Schapire. 1996. "Experiments with a New Boosting Algorithm." *ICML* 96: 148–56.

Friedman, Jerome H. 1991. "Multivariate Adaptive Regression Splines." *Annals of Statistics* 19(1):1–67.

Gelman, Andrew, and Eric Loken. 2013. "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No 'Fishing Expedition' or '*P*-Hacking' and the Research Hypothesis Was Posited Ahead of Time."Department of Statistics, Columbia University, New York, NY.

Golden, Richard M., Steven S. Henley, Halbert White, and Michael T. Kashner. 2016. "Generalized Information Matrix Tests for Detecting Model Misspecification." *Econometrics* 4(4):46.

Goodman, Bryce, and Seth Flaxman. 2017. "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation'." *AI Magazine* 38(3):50–57.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2021. "Machine Learning for Social Science: An Agnostic Approach." *Annual Review of Political Science* 24:395–419.

Hastie, Trevor, and Robert Tibshirani. 1987. "Generalized Additive Models: Some Applications." *Journal of the American Statistical Association* 82(398):371–86.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

Heckman, James J., John Eric Humphries, and Veramendi Gregory. 2018. "Returns to Education: The Causal Effects of Education on Earnings, Health, and Smoking." *Journal of Political Economy* 126(S1):S197–246.

Heskes, Tom, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. 2020. "Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models." *Advances in Neural Information Processing Systems* 33:4778–89.

Hindman, Matthew. 2015. "Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences." *ANNALS of the American Academy of Political and Social Science* 659(1):48–62.

Hofman, Jake M., Amit Sharma, and Duncan J. Watts. 2017. "Prediction and Explanation in Social Systems." *Science* 355(6324):486–88.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. "Multilayer Feedforward Networks Are Universal Approximators." *Neural Networks* 2(5):359–66.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2(8): e124.

Janson, Lucas, William Fithian, and Trevor J. Hastie. 2015. "Effective Degrees of Freedom: A Flawed Metaphor." *Biometrika* 102(2):479–85.

Kleinberg, Jon, Annie Liang, and Sendhil Mullainathan. 2017. "The Theory Is Predictive, but Is It Complete? An Application to Human Perception of Randomness." Pp. 125–26 in *Proceedings of the 2017 ACM Conference on Economics and Computation*. New York, NY: Association for Computing Machinery.

Kohavi, Ron. 1995. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." *IJCAI* 14:1137–45.

Krishna, Satyapriya, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. 2022. "The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective." arXiv. https://doi.org/10.48550/arXiv.2202.01602

Lapuschkin, Sebastian, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. "Unmasking Clever Hans Predictors and Assessing What Machines Really Learn." *Nature Communications* 10(1):1–8.

Lemieux, Thomas. 2006. "The 'Mincer Equation' Thirty Years After Schooling, Experience, and Earnings." Pp. 127–45 in *Jacob Mincer a Pioneer of Modern Labor Economics*, edited by Grossbard, S.. New York, NY: Springer.

Liao, Yung-Sheng. 2017. "Machine Learning in Macro-Economic Series Forecasting." *International Journal of Economics and Finance* 9(12):71–76.

Lipton, Zachary C. 2018. "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery." *Queue* 16(3):31–57.

Long, J. Scott, and Pravin K. Trivedi. 1992. "Some Specification Tests for the Linear Regression Model." *Sociological Methods & Research* 21(2):161–204.

Lundberg, Scott M., Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. "From Local Explanations to Global Understanding with Explainable AI for Trees." *Nature Machine Intelligence* 2(1):56–67.

Lundberg, Ian, Rebecca Johnson, and Brandon M. Stewart. 2021. "What is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory." *American Sociological Review* 86(3): 532–65.

Lundberg, Scott M., and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30: 4765–4774.

Malpezzi, Stephen. 2003. "Hedonic Pricing Models: A Selective and Applied Review." *Housing Economics and Public Policy* 1:67–89.

McClintock, Elizabeth Aura. 2017. "Occupational Sex Composition and Gendered Housework Performance: Compensation or Conventionality?" *Journal of Marriage and Family* 79(2):475–510.

Meisenberg, Gerhard. 2015. "Verbal Ability as a Predictor of Political Preferences in the United States, 1974–2012." *Intelligence* 50:135–43.

Molina, Mario, and Filiz Garip. 2019. "Machine Learning for Sociology." *Annual Review of Sociology* 45: 27–45.

Morgan, Stephen L., and Jiwon Lee. 2017. "Social Class and Party Identification during the Clinton, Bush, and Obama Presidencies." *Sociological Science* 4:394–423.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31(2):87–106.

Muñoz, John, and Cristobal Young. 2018. "We Ran 9 Billion Regressions: Eliminating False Positives Through Computational Model Robustness." *Sociological Methodology* 48(1):1–33.

O'Brien, Robert M. 2018. "Comment: Some Challenges When Estimating the Impact of Model Uncertainty on Coefficient Instability." *Sociological Methodology* 48(1):34–39.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349(6251):aac4716.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge University Press.

Peterson, Joshua C., David D. Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L. Griffiths. 2021. "Using Large-Scale Experiments and Machine Learning to Discover Theories of Human Decision-Making." *Science* 372(6547):1209–14.

Peysakhovich, Alexander, and Jeffrey Naecker. 2017. "Using Methods from Machine Learning to Evaluate Behavioral Models of Choice under Risk and Ambiguity." *Journal of Economic Behavior & Organization* 133:373–84.

Polley, Eric, Erin LeDell, Chris Kennedy, Sam Lendle, and Mark van der Laan. 2021. "Package Superlearner." https://cran.r-project.org/web/packages/SuperLearner/index.html. Accessed 1 November 2022.

Polley, Eric C., Sherri Rose, and Mark J. Van der Laan. 2011. "Super Learning." Pp. 43–66 in *Targeted Learning*, edited by M. J. van der Laan, and S. Rose. New York, NY: Springer.

Rahal, Charles, Mark D. Verhagen, and David S. Kirk. 2022. "The Rise of Machine Learning in the Academic Social Sciences." *AI and Society*. Advance online publication. https://doi.org/10.1007/s00146-022-01540-w

Ramsey, James Bernard. 1969. "Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis." *Journal of the Royal Statistical Society: Series B (Methodological)* 31(2): 350–71.

Robinson, Peter M. 1988a. "Root-N-Consistent Semiparametric Regression." *Econometrica: Journal of the Econometric Society* 56(4):931–54.

Robinson, Peter M. 1988b. "Semiparametric Econometrics: A Survey." *Journal of Applied Econometrics* 3(1):35–51.

Rose, Sherri. 2013. "Mortality Risk Score Prediction in an Elderly Population Using Machine Learning." *American Journal of Epidemiology* 177(5):443–52.

Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1(5):206–15.

Rudin, Cynthia, Rebecca J. Passonneau, Axinia Radeva, Haimonti Dutta, Steve Ierome, and Delfina Isaac. 2010. "A Process for Predicting Manhole Events in Manhattan." *Machine Learning* 80(1):1–31.

Sala-i Martin, Xavier X. 1997. "I Just Ran Four Million Regressions." Technical report. Cambridge, MA: National Bureau of Economic Research.

Salganik, Matthew J., Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul, et al. 2020. "Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration." *Proceedings of the National Academy of Sciences* 117(15): 8398–403.

Samek, Wojciech, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. 2019. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. New York, NY: Springer Nature.

Shapley, Lloyd S. 1953. "Stochastic Games." *Proceedings of the National Academy of Sciences* 39(10): 1095–100.

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25(3):289–310.
Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson. 2020. "Specification Curve Analysis." *Nature Human Behaviour* 4(11):1208–14.
Slez, Adam. 2019. "The Difference between Instability and Uncertainty: Comment on Young and Holsteen (2017)." *Sociological Methods & Research* 48(2):400–30.
Stone, Mervyn. 1977. "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion." *Journal of the Royal Statistical Society: Series B (Methodological)* 39(1):44–47.
Tang, Siyi, Amirata Ghorbani, Rikiya Yamashita, Sameer Rehman, Jared A. Dunnmon, James Zou, and Daniel L. Rubin. 2021. "Data Valuation for Medical Imaging Using Shapley Value and Application to a Large-Scale Chest X-Ray Dataset." *Scientific Reports* 11(1):1–9.
Van der Laan, Jan, Edwin de Jonge, Marjolijn Das, Saskia Te Riele, and Tom Emery. 2022. "A Whole Population Network and Its Application for the Social Sciences." *European Sociological Review* 39(1): 145–60.
Van Der Laan, Mark J., and Sandrine Dudoit. 2003. "Unified Cross-Validation Methodology for Selection among Estimators and a General Cross-Validated Adaptive Epsilon-Net Estimator: Finite Sample Oracle Inequalities and Examples." U.C. *Berkeley Division of Biostatistics* Working Paper Series. Berkeley, CA: The Berkeley Electronic Press.
Van der Laan, Mark J., Eric C. Polley, and Alan E. Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6(1):1–23.
Van der Vaart, Aad W., Sandrine Dudoit, and Mark J. van der Laan. 2006. "Oracle Inequalities for Multi-Fold Cross Validation." *Statistics & Decisions* 24(3):351–71.
Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. "Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC." *Statistics and Computing* 27(5):1413–32.
Verhagen, Mark D. 2022. "A Pragmatist's Guide to Using Prediction in the Social Sciences." *Socius* 8. https://doi.org/10.1177/23780231221081702
Watts, Duncan J. 2014. "Common Sense and Sociological Explanations." *American Journal of Sociology* 120(2):313–51.
Watts, Duncan J. 2017. "Should Social Science Be More Solution-Oriented?" *Nature Human Behaviour* 1(1):1–5.
White, Halbert. 1980. "Using Least Squares to Approximate Unknown Regression Functions." *International Economic Review* 21(1):149–70.
White, Halbert. 1981. "Consequences and Detection of Misspecified Nonlinear Regression Models." *Journal of the American Statistical Association* 76(374):419–33.
Yatchew, Adonis. 1997. "An Elementary Estimator of the Partial Linear Model." *Economics Letters* 57(2):135–43.
Yatchew, Adonis. 2003. *Semiparametric Regression for the Applied Econometrician*. Cambridge, UK: Cambridge University Press.
Young, H. Peyton. 1985. "Monotonic Solutions of Cooperative Games." *International Journal of Game Theory* 14(2):65–72.
Young, Cristobal. 2018. "Model Uncertainty and the Crisis in Science." *Socius* 4. https://doi.org/10.1177/2378023117737206
Young, Cristobal. 2019. "The Difference between Causal Analysis and Predictive Models: Response to 'Comment on Young and Holsteen (2017).'" *Sociological Methods & Research* 48(2):431–47.
Young, Cristobal, and Katherine Holsteen. 2017. "Model Uncertainty and Robustness: A Computational Framework for Multimodel Analysis." *Sociological Methods & Research* 46(1):3–40.
Young, Cristobal, and Sheridan A. Stewart. 2021. "Functional Form Robustness: Advancements in Multiverse Analysis." Unpublished manuscript. http://cristobalyoung.com/development/wp-content/uploads/2021/08/Multiverse-Aug-2021.pdf. Accessed 1 November 2022.
Yousef, Waleed A. 2020. "A Leisurely Look at Versions and Variants of the Cross Validation Estimator." *stat* 1050:9.

Zhou, Jianlong, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. 2021. "Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics." *Electronics* 10(5):593. https://doi.org/10.3390/electronics10050593

Author Biography

Mark Verhagen is a researcher at the Leverhulme Centre for Demographic Science and member of Nuffield College, Oxford. His work uses computational methods to improve the construction and understanding of explanatory models. He has published in a wide range of outlets, including *The Proceedings of the National Academy of Sciences, BMC Medicine, PloS One, The International Review of Law and Economics*, and *Socius.*