Incorporating Machine Learning into Sociological Model-building

Mark D. Verhagen^{*a,b}

^aLeverhulme Centre for Demographic Science, 42 Park End St, Oxford OX1 1JD, UK

^bNuffield College, University of Oxford, 1 New Rd, Oxford OX1 1NF, UK

November 1, 2022

Abstract

Throughout quantitative sociology, researchers define functional forms dictating how variables relate to one another. However, there is little guidance whether a functional form is correct. This can lead to misspecification bias and fosters interference of researcher degrees of freedom in sociological work. Machine Learning (ML) models have shown an impressive ability to find patterns in data, but have been mostly restricted to predictive questions within sociology. In this paper, I propose a framework that uses methods from the ML domain to provide an indication of the fit potential in a dataset containing the same covariates as a researcher’s hypothesized model. This potential can then be used to identify a possible lack of model fit in the researcher’s functional form. Advances in the field of Explainable AI (X-AI), notably Shapley values, can then be used to generate understanding into the ML model and improve the researcher’s original functional form whenever a large difference in fit is found. The framework thus aims to use ML to build better interpretable models, rather than replace existing methods or require rephrasing sociological question in a predictive context. Additionally, the framework improves transparency in empirical model-building. I illustrate the approach using both simulation and real-world examples.

^{*}All code and publicly available data underlying the analyses in this paper can be found at <https://github.com/MarkDVerhagen/Functional.Form.Complexity>. Correspondence may be addressed to: mark.verhagen@nuffield.ox.ac.uk.

1 Introduction

Sociologists are frequently required to specify a model of the world in their empirical work. Most commonly encountered is the linear additive functional form plugged into an exponential family probability distribution. Although the choice of functional form dictates the way in which variables can affect one another, it is often unclear whether our functional forms correctly represent the data generating process (DGP) in which we are interested. If incorrect, misspecification can lead to severe bias in empirical findings or simplify the true patterns in the data. For example, whenever there are important interactions or non-linearities present. Currently, such modeling nuances have to be encoded by the researcher themselves and can be easily overlooked. Furthermore, a lack of a principled approach to evaluating functional form specification opens the door for researcher degrees of freedom to interfere in sociological model-building (Muñoz and Young, 2018). By allowing the sociologist to specify their functional form in relative isolation, repeated evaluation of specifications invalidates statistical uncertainty and can lead to so-called ‘p-hacking’ (Gelman and Loken, 2013). In this paper, I propose to use methods from the machine learning (ML) domain to assist sociologists in the process of model-building and help address these issues.

ML methods have shown an impressive ability to find intricate patterns in data. As a result, they are increasingly being adopted in industry and research (Rahal et al., 2022; Athey, 2018) to assist in finding structure relating explanatory variables, X , with an outcome of interest, y . The key difference with standard empirical work in sociology is that ML methods exploit the exponential increase in computational resources available in this day and age to let the data dictate the relationship amongst variables, rather than relying on the researcher to hypothesize the functional relationship between the two.¹ Often, this flexibility leads to better fitting models that improve on researcher-hypothesized models in fitting the data (Brand et al., 2021; Grimmer et al., 2021). Presently, ML methods are almost exclusively applied within a predictive context (Mullainathan and Spiess, 2017; Shmueli, 2010) and its inclusion into sociological inquiry — which has an overwhelmingly explanatory focus — remains limited.² This is unfortunate, I argue, and the fact that ML methods can

¹The difference has been described by Breiman (2001) 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, it is assumed the functional form is fundamentally unknown and has to be learned from the data. As Breiman notes, we typically reside in the latter of the two worlds, although the majority of statistical work uses empirical approaches that assumes we reside in the former (p. 199).

²Indicative 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

find associations amongst variables holds value for much explanatory work as well.

Concretely, I propose introducing ML methods into the typical empirical workflow in the following way. Assume a researcher develops some hypothesized functional form, $\tilde{f}(\cdot)$, for which it is unclear whether it correctly represents the underlying relationship amongst variables in the data. First, the hypothesized functional form is estimated and, in parallel, a non-parametric ML model is estimated to the same dataset. The flexible model thus uses the same set of variables as identified by the researcher in their own hypothesized model, and follows the same inferential logic.³ However, the functional form in which variables relate to one another is left to be decided by the data. Second, the fit of both the hypothesized model and the flexible ML model is evaluated out-of-sample, and a possible lack of fit in the former is identified by comparing the two models. Third, in case a considerable lack of appropriate specification 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 crucially relies on a number of concepts from ML or closely adjacent fields. First, that out-of-sample model fit can be used to provide a comparable indication of a model’s fit across model domains (Stone, 1977; Rose, 2013; Verhagen, 2022). In this case to compare the researcher’s model with the ML alternative. Second, that ML methods can be used as efficient function approximators that can uncover relevant patterns in data (Van der Laan et al., 2007; Baćak and Kennedy, 2019). Third, that methods from the rapidly evolving field of Explainable A.I. (X-AI) — like the Shapley values approach explored in this paper — allow researchers to distill understanding from ML methods, which can be used to improve their originally 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 overarching goal of the framework is thus still to generate interpretable models, whilst using ML methods as a guide rather than a replacement for classical methods (Rudin, 2019; Agrawal et al., 2020).⁴ In other words, ML methods are used to complement the

questions into predictive ones. Similar assessments can be found in the reviews by Molina and Garip (2019) and Shmueli (2010).

³By following a set inferential logic (namely that of the original model), complexity in the proposed framework is introduced solely through the functional relationship amongst variables and distinguishes itself from ‘model complexity’ as is often understood as the problem of selecting amongst a (large) set of potential covariates (Bucca and Urbina, 2019).

⁴The iterative and complementary role envisioned for ML is in line with an increasing body of literature that recommends using ML in a supportive role throughout the analytical pipeline whilst warning against completely replacing interpretable methods for ML equivalents (Agrawal et al., 2020; Rudin et al., 2010; Rudin, 2019, p. 3). A similar iterative perspective on ML is discussed by Grimmer et al. (2021) who envision a ‘human in the loop’ approach, where ML models provide insights to the researcher but do not supplant them (p. 406-408).

60 standard empirical workflow of sociologists. The proposed framework has two main
61 goals, which were introduced above and are discussed in more detail below. They
62 are first and foremost to better align our functional forms with the empirical data,
63 and secondly to provide increased transparency in an often opaque empirical research
64 process.

65 As is common knowledge, valid inference crucially depends on a correctly specified
66 relationship between the outcome of interest, y , and the explanatory variables, X
67 (Buja et al., 2019a; Cameron and Trivedi, 2005; Long and Trivedi, 1992). In prac-
68 tice, though, much more attention is typically paid to identifying relevant variables
69 to include into a model than making sure the functional relationship amongst these
70 variables is correctly specified. This is evidenced by the fact that variables are of-
71 ten simply specified to affect the outcome in a linear and additive way (Hindman,
72 2015). There are various noteworthy examples throughout the social sciences — and
73 likely many more that have gone unnoticed — where more complicated functional
74 relationships, which might for instance include non-linearities or interactions, have
75 led to reversed findings (Muñoz and Young, 2018; Heckman et al., 2018; McClintock,
76 2017; Freedman, 2009; Christensen and Christensen, 2014; Dougherty et al., 2015)
77 (see also the simulation example in the next section). More generally, criticism of
78 the standard linear additive model, which remains the workhorse throughout most
79 empirical sociology, has been around for many decades (Duncan, 1984; Abbott, 1988;
80 Berk, 2004; Lundberg et al., 2021). Such criticism is effectively one of widespread
81 misspecification in empirical work.

82 Fortunately, model simplifications that were a necessity in the past are no longer
83 required in the computationally rich present.⁵ ML’s ability to find patterns in data
84 through exploiting computational power allows for a data-driven approach to model-
85 building that can include considerable levels of complexity in terms of non-linearities
86 and interactions amongst variables. Effectively, a much wider range of possible
87 model specifications is considered than a researcher would typically do during model-
88 building. This potential of ML-based methods to find improved patterns in data
89 has already been illustrated through a number of examples in sociology (Baćak and
90 Kennedy, 2019; Brand et al., 2021), psychology (Agrawal et al., 2020) and behav-
91 ioral work (Kleinberg et al., 2017; Peterson et al., 2021; Peysakhovich and Naecker,
92 2017). At the same time, their increased structure makes ML methods more appli-
93 cable from a practical perspective than fully non-parametric regression which suffers

⁵Linear 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, Chapter 4). This historic limitation still dictates the domination of these approaches in empirical work today.

94 from exponential data requirements (Yatchew, 1997; Bishop, 2006, Chapter 1).

95 Another consequence of a relative lack of emphasis on functional form specification is
96 that it allows researchers to effectively evaluate a large number of possible functional
97 forms and choose which ones to report (Sala-i Martin, 1997; Muñoz and Young, 2018;
98 Simonsohn et al., 2020). For example, whether to include a possible confounder in
99 a linear, logged or squared form. Such ‘researcher degrees of freedom’ are to blame
100 for a large number of important empirical findings turning out to be unreproducible
101 and a more general ‘crisis in science’ where results are often dependent on over-
102 engineered models (Gelman and Loken, 2013; Young, 2018). Allowing researchers
103 to cherry-pick specifications — knowingly or unknowingly (Epley and Gilovich, 2016;
104 Gerber and Malhotra, 2008) — invalidates reported certainty thresholds and leads to
105 false positives (Ioannidis, 2005).

106 Introducing ML methods into the empirical workflow does not solve the problem of
107 ‘p-hacking’ entirely, but it does introduce additional transparency into the research
108 process in two ways. First, comparing a researcher’s carefully tailored model to the
109 fit of a ML method estimated to the same data can reveal a severe lack of model
110 fit and provides a more general insight into the potential fit in the data that is not
111 primarily a result of a researcher’s decision-making process but rather a feature of
112 the data. Second, ML methods put considerable emphasis on re-estimating models
113 on many sampled subsets — for example through cross validation — which further
114 protects against ‘p-hacking’ (Gelman and Loken, 2013).

115 Beyond these two main goals, the proposed framework also provides a welcome re-
116 alignment of empirical methods with the computational reality of this day and age
117 and provides a way for ML methods to be used as complementary tools in the sociolo-
118 gist’s standard empirical workflow (Efron and Hastie, 2016; Muñoz and Young, 2018;
119 Rudin, 2019). Additionally, the introduction of ML methods into sociological work is
120 in line with increasing calls for a re-appreciation of the predictive power of sociologi-
121 cal theories, which can be observed more broadly throughout sociology (Watts, 2014;
122 Hofman et al., 2017; Watts, 2017; Verhagen, 2022).

123 The rest of this paper is structured as follows. In Section 2, I introduce a toy example
124 to illustrate the risk of model misspecification and researcher degrees of freedom and
125 provide a more general discussion on the topic. Then, in Section 3, I introduce each
126 of the three steps of the proposed framework and illustrate its potential by applying
127 the framework to the same toy example from the previous section. I compare the
128 proposed approach to other computational frameworks in Section 4. In Section 5,
129 the empirical part of this paper, I apply the framework to three cases; one simulated

and two real-world datasets. In doing so, I illustrate how the framework identifies i) missing non-linearities and interactions in a simulated dataset on the association between schooling and earnings, ii) non-linearities and spatial heterogeneity in a large dataset on London house sales, and iii) various intricate patterns in voting preferences amongst US voters in the General Social Survey (GSS). In Section 6, I provide a discussion and concluding remarks.

2 Incorrect Models of the World and Risks to Inference

Consider the following toy example. We are interested in whether having grandchildren affects elderly individuals' worries about climate change. To this effect, we analyze a survey ($N = 900$) where respondents were asked their level of concern with climate change on a 7-point scale. In addition, we have information on the respondent's age, sex, whether they completed high school and, our interest variable, whether they have grandchildren or not. We might propose the following model ('Model 1'):

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

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

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

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 to conclude our empirical analysis at this stage.

Imagine, however, that the true effect of age is indeed non-linear, but piece-wise linear rather than a second-degree polynomial. Specifically: there is a strong increase in associated concerns amongst those aged 65-75 with more modest increases at other

	Model 1	Model 2	Model 3
(Intercept)	-19.02*** (0.51)	-2.16 (3.42)	9.97*** (0.23)
Age : linear	0.52*** (0.01)	0.03 (0.10)	
Grandchildren	-0.02 (0.21)	-0.68** (0.25)	0.57* (0.23)
Sex: female	-0.97*** (0.10)	-0.99*** (0.10)	-0.95*** (0.08)
High school	0.41*** (0.10)	0.38*** (0.10)	0.36*** (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

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

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.

ages. If we had estimated a functional form in line with this non-linearity, we would find results as reported in the third column of the left panel of Table 1, which shows 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 accompanying figure to the right of Table 1. The problem is that both the linear, as well as the quadratic function of age overshoots the effect as the respondent's age increases. Coupled with the fact that grandchildren are more prevalent under the older part of the population, this effect leads to incorrect inference.

To generalize the above, the central issue with model misspecification is that the assumption of exogeneity, i.e., $E(\epsilon|X) = 0$, is violated. Consider four bivariate relationships plotted in Figure 1.⁶ Three of the four models include a non-linear relationship, and although a linear coefficient estimated to the data ($N = 80$) is statistically significant at the conventional 5% level, exogeneity is violated in each of the three cases. This is most clearly visible in the quadratic specification where squared residuals are much larger at the ends and the middle of the X distribution and are clearly correlated with the covariate. This violation of exogeneity leads to a form of omitted variable bias that renders estimated coefficients inconsistent (Long and Trivedi, 1992;

⁶Choice of functional forms for this example have been inspired by Polley et al. (2011).

Figure 1. Implied association from a linear model (dashed line) relative to the true effect (solid line). The linear models are estimated based on 80 uniformly distributed observations for the X variable, plugged into one of four functional forms (see Section 6) with an additive normally distributed error. Data points are shown in green.

176 [Cameron and Trivedi, 2005](#), Chapter 8).⁷

177 To guard against misspecification, various statistical tests have been proposed in the
 178 literature. Most popular are White’s test for functional misspecification ([White, 1980](#),
 179 [1981](#)) and Ramsey’s RESET test ([Ramsey, 1969](#)). White’s test statistic is based on a
 180 comparison of the estimated coefficients from a hypothesized model $\hat{\beta}$ with those from
 181 a weighted regression $\hat{\beta}_{WLS}$, whereas the Ramsey test statistic compares the hypothe-
 182 sized model with a model which includes higher order versions of the explanatory
 183 variables and evaluates their statistical significance. Both provide computationally
 184 efficient statistics to evaluate basic misspecification of the functional form, although
 185 both also assume that the true complexity of the model lies within a specific class
 186 of functions $\mathcal{M}$ with possibly restrictive assumptions ([Golden et al., 2016](#)).⁸ Since
 187 the introduction of the White and Ramsey tests, more computationally intensive
 188 (semi-)parametric methods have been developed to estimate parameters of interest
 189 in case of non-linear effects in control variables which have been used as the basis for
 190 a different set of misspecification tests (discussed in more detail below) ([Robinson](#),
 191 [1988a,b](#); [Yatchew, 1997](#)). At the same time, there has been substantial follow-up re-

⁷Assume some covariate X affects the outcome in a non-linear way — e.g., through a polynomial relationship — then excluding the higher order terms effectively leads to omitted variable bias. Combined with the fact that the omitted variable is in general associated with the linear covariate guarantees endogeneity and bias in parameter estimates.

⁸For example, neither test will pick up an interaction of some continuous variable x_k and a binary variable D if the (linear) slopes are mirrored for both groups.

192 search building on the principles of the White and Ramsey tests (Golden et al., 2016).
193 However, there are various well-known limitations to these types of specification tests
194 (Buja et al., 2019a).

195 First, they are generally under-powered and struggle to identify misspecification in
196 multivariate settings (Buja et al., 2019b, p. 616).⁹ Second, they provide limited
197 insight into what to do next, whenever a test is rejected. Third, and perhaps most
198 importantly, the application of misspecification tests differs substantially throughout
199 published empirical work (Long and Trivedi, 1992; Open Science Collaboration, 2015),
200 and continues to this day.¹⁰ It is safe to say that the statistical frontier has left the
201 practical frontier far behind and the wealth of follow-up research on specification tests
202 has barely penetrated into typical empirical analysis. As a consequence, researchers
203 enjoy relative freedom in specifying their functional form, as well as the robustness
204 and specification tests they use to verify its appropriateness. In other words, the
205 crisis in reproducibility and the impact of researcher degrees of freedom on empirical
206 results extend to the (lack of) specification checks chosen by researchers to validate
207 their model assumptions (Simonsohn et al., 2020; Young, 2018).

208 Even without malicious intent, many functional forms are from an age of computa-
209 tional limitations that are no longer present (Muñoz and Young, 2018; Efron and
210 Hastie, 2016; Buja et al., 2019b, p. 615). The linear additive functional forms
211 plugged into exponential family probability distributions were historically preferred
212 due to their ease of estimation, not their de-facto appropriateness to study social
213 life. This pragmatic preference for parsimony has led to a prevalence of fairly sim-
214 plistic functional forms which remain in use to this day, even though the wealth of
215 qualitative research in the social sciences consistently implies that social life is in fact
216 highly complex and probably not appropriately modelled using such simple functional
217 forms (Abbott, 1988). As the toy example above illustrates, it is likely that under-
218 specification of functional forms is a prevalent problem throughout sociology that can
219 easily go unnoticed. Failure to consistently apply the tools to diagnose such a lack
220 of model specification will maintain this status quo and risks further increasing the
221 number of false and irreproducible findings throughout empirical work.

⁹To illustrate, misspecification tests are generally able to identify the bivariate non-linearities 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.

¹⁰As a case in point, the most recent five issues of the *American Sociological Review* at the time of writing (Volume 87, Issues 1-5, 2022) contain 30 journal articles. 19 of these include quantitative analyses of which 14 implement a conditioning-on-observables approach. None of these articles implements any of the standard misspecification tests like the White or RESET tests.

3 A Computational Framework to Improve Model-building

Instead of (selective) application of specification tests to hypothesized functional forms, 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 the hypothesized model, and thus used to diagnose a lack of specification in the latter. The proposed framework has a number of benefits. First, it is simple to apply and provides the researcher with an intuitive assessment of how well the data *could* be modelled versus how well the data *is* modelled by their hypothesized model. Second, as long as a broad suite of flexible form models is applied to generate the fit potential, there is no room for researcher degrees of freedom to affect findings. Third, using ML models to benchmark model fit allows for an explicit feedback loop to understand how to improve an underspecified hypothesized model. In case a lack of fit is found, the ML model outperforming the hypothesized model can be unpacked to understand what patterns are picked up and possibly missed by 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:

1. Estimate hypothesized model $\tilde{f}(\cdot)$ to the data, as well as a 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.

The proposed benchmarking framework including each of the three steps is illustrated schematically in Figure 2. Note that benchmarking uses the same variables as originally hypothesized by the researcher. The framework is not designed for data-mining where a large number of possible explanatory variables are included into the model, which risks including variables that might improve model fit but could harm our understanding of the underlying processes — typical examples would be to (accidentally) include post-treatment variables or colliders. Rather, the approach here is to evaluate the functional form assumptions given the exact same inferential curation of variables and logic as present in the researcher’s originally hypothesized model.¹¹

¹¹As discussed in [Grimmer et al. \(2021\)](#), the framework still uses variables that are relevant from an inferential perspective rather than those that give the best possible fit (p. 412). Given these

Figure 2. Schematic illustration of the proposed framework.

The framework relies crucially on three components. First, a flexible form *benchmarking model* to estimate the potential fit in the data. Second, a *fit metric* to compare the flexible model’s performance to that of the hypothesized model and identify a possible lack of model fit. Third — in case a considerable lack of fit is found — an *unpacking method* to generate understanding into why the hypothesized functional form underperformed the benchmark. In what follows, I will describe each of the three components of the proposed framework in-depth and illustrate them using the toy example introduced earlier.

Step I: The benchmarking model

The first component of the proposed framework is a flexible form model $\bar{f}(\cdot)$ to serve as a benchmark of the possible model fit in the data. This model should include the same variables and thus follow the same inferential logic as the hypothesized model $\tilde{f}(\cdot)$, but evaluate patterns not necessarily considered by the researcher’s functional form. As ML methods do not rely on pre-specified functional forms, but distill the functional form from the observed data they are highly appropriate for this task (Grimmer et al., 2021; Baćak and Kennedy, 2019). They also retain more structure than non-parametric approaches like local regression and suffer less from the ‘curse of dimensionality’ where the number of data points required scales exponentially with

variables, though, we *are* interested in a model that fits the data well. Note that ML methods are also used to identify relevant variables amongst a (large) set of possible covariates (Bucca and Urbina, 2019; Grimmer et al., 2021). This is another fruitful applications of ML into sociological model-building, but outside of the scope of this paper.

the covariate space (Bishop, 2006, Chapter 1). Many ML methods are feasible to estimate to datasets commonly encountered in the social sciences (Hastie et al., 2009; Athey, 2018).

In particular tree-based models — a subset of ML methods — lend themselves exceptionally well to the envisioned benchmarking approach. First, tree-based methods can deal with various types of data often encountered in the social sciences — e.g., categorical as well as numerical variables — and handle missing data well. Second, tree-based methods do not require exponential increases in sample size as the feature space increases, like other popular methods from the ML domain like neural networks. Third, the most popular tree-based methods like the Random Forest (RF) or Gradient Boosting (GB) model incorporate regularization and variable selection by design, requiring less tuning on the part of the researcher and allowing the method to be applied out-of-the-box or with limited effort to most datasets.¹² Finally, tree-based methods have been shown to perform well amongst ML methods across various applications whenever data is tabular, as is almost always the case in the social sciences (Lundberg et al., 2020; Chen and Guestrin, 2016).

Take the two most common tree-based methods, the RF and the GB model. Both approaches share the flexibility of a decision tree, but address risks of overfitting to the sample at hand by averaging many decision trees estimated to slightly altered (for RF) or weighted (for GB) versions of the data (Breiman, 1996; Freund and Schapire, 1996). This allows for the estimation of highly non-linear patterns without requiring any type of functional form specification on the part of the researcher. As a case in point, consider the non-linear associations in Figure 3, which had been introduced earlier. As before, the black lines illustrate true associations and the green dots a hypothetical dataset generated by the true association and some white noise. The blue and red dots are out-of-sample predicted values based on a GB and a RF model trained to 80 observed data points and fed with 100 uniformly distributed X values to generate $\hat{y}$ -predictions for. As can be seen, both models learn the non-linearity in the data well and do so without requiring any sort of structure or pre-hypothesized functional relationship.

Besides a particular tree-based model, other ML or non-parametric methods could be

¹²For some ML methods the optimal solution is dependent on the scale of the explanatory variables. For example, least absolute shrinkage and selection operator regression 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 which provide limited improvement to fit towards zero. Clearly, dividing or multiplying an explanatory variable x_k 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.

Figure 3. Predicted outcome using a GB (blue) and RF (red) model fit to noisy data (green). 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 Section 6). 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.

used as well. More generally, limiting the benchmarking step to a single researcher-curated ML model to serve as $\tilde{f}(\cdot)$ would clearly invite some of the very risks this framework is designed to address in terms of researcher degrees of freedom in model-building. Tuning parameters can easily be calibrated in a way to diminish model fit and thus imply appropriate fit of $\tilde{f}(\cdot)$ when this is not the case.¹³ Relatedly, some methods fit certain patterns in data better than others with often limited a-priori guidance (Rose, 2013; Baćak and Kennedy, 2019; Berk and Bleich, 2013). A principled approach would be to estimate not a single but an ensemble of flexible form models during the benchmarking phase.

An ensemble or ‘Super Learner’ can consist of multiple models, as well as the same model with different parameter settings. From 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 amongst the evaluated models, and the price of including large numbers of models into the Super Learner is minor in terms of performance (Van der Vaart et al., 2006; Van der Laan et al., 2007; Baćak and Kennedy, 2019). Therefore, many models can and should be evaluated and ensembles are of-

¹³For example by forcing trees to be unreasonably shallow or deep, possibly leading to limited fit or severe overfitting with poor performance out-of-sample.

ten implemented as universal function approximators (Hornik et al., 1989; Agrawal et al., 2020). The Super Learner can also be specified prior to model estimation or even pre-registered, improving transparency into 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 (as is illustrated in the examples below).¹⁴

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. In addition, I include the two researcher-hypothesized models to the Super Learner as well, specifically ‘SL.GLM.Linear’ and ‘SL.GLM.Quadratic’ reflecting the model including linear age and quadratic age, respectively. For illustrative purposes, I also add the true model, ‘SL.GLM.PieceLinear’, even though I did not in fact hypothesize this model originally. A truncated output from the Super Learner can be found in Table 2 and shows the root mean squared error (RMSE) of various models included in the Super Learner, estimated based on 10 folds using cross-validation (for the full output see Table A.1). As can be seen, the best performing model is — unsurprisingly — the true model, but its performance is closely tracked by a GB model which strongly outperforms both the Linear as well as the Quadratic specifications. What is more, 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 underperformance in terms of fit. This further illustrates the need to include various models during benchmarking. The results from the Super Learner suggests, correctly, that a linear or quadratic specification misses important patterns in the data which are picked up by the GB model.

Step II: Estimating and comparing model fit

The second component of the proposed framework is a metric to compare the fit of the selected benchmarking model $\bar{f}^*(\cdot)$ from the previous step with that of the hypothesized model $\tilde{f}(\cdot)$. Comparative statistics like the AIC or BIC, or LR tests (in case of nested models) are typically 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 known to be challenging for ML models (Janson et al., 2015). As a result, ML methods are typically evaluated on their out-of-sample performance (Shmueli, 2010; Hastie et al., 2009, Chapter 7). The comparability of

¹⁴Throughout the empirical examples in this paper, I will use the Super Learner approach as implemented in the R package **SuperLearner** although various alternatives exist in different programming languages (Polley et al., 2021).

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

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 Table A.1 in the Appendix for the full output.

Table 2. SuperLearner performance of the Toy Example

out-of-sample prediction across models makes them an attractive fit metric for the framework (Stone, 1977; Van Der Laan and Dudoit, 2003; Verhagen, 2022).

Evaluating a model out-of-sample typically means splitting the total dataset $\mathcal{D}$ into two partitions, an estimation-set, $\mathcal{D}_{\text{estimation}}$, used to estimate the model, and an evaluation set, $\mathcal{D}_{\text{evaluation}}$, used to evaluate model fit. Estimating the model to the former and generating predictions using the latter allows for common summary metrics of fit to be assessed, like the RMSE for continuous outcomes or the F1-score or AUC for binary outcomes (Hastie et al., 2009). Given any of these out-of-sample fit metrics, a principled approach to evaluating an improvement in model fit is to calculate the difference in out-of-sample model fit between the flexible form model and the hypothesized model for a large number of different partitions of the data into estimation- and evaluation-sets. This ensures robustness of the comparison against a specific partition.¹⁵ A complementary benefit of this routine is that bootstrapped estimates of the coefficients for the hypothesized model $\tilde{f}(\cdot)$ are obtained which should help identify a lack of robustness in an estimated coefficient and p-hacking.¹⁶ In what follows, I report both the RMSE and out-of-sample R^2 over a large number of splits of the data into estimation- and evaluation-sets.

It is useful to briefly reflect on the practice of optimizing model fit. As has been noted

¹⁵A common approach in the literature is K -fold cross validation, which splits the data into K equal sized folds. The model is estimated K times, omitting one of the folds and using the held out fold to generate predictions (Kohavi et al., 1995). Some limitations have been noted in the literature. For example, the standard errors from such routines can be biased downwards by including the same datapoints in subsequent folds for training and testing (Bates et al., 2021; Vehtari et al., 2017).

¹⁶In effect, simple bootstrapped standard errors of the coefficient estimates are calculated (Cameron and Trivedi, 2005, Chapter 11). In case a model has been over-engineered to a sample at hand this should be reflected by the empirical standard errors from re-estimating the model to resampled datasets compared to that of the researcher’s preferred specification.

by others, fit needn't automatically equate with the best model from an inferential perspective (Muñoz and Young, 2018; Grimmer et al., 2021). However, this discussion is often concerned with including colliders or post-treatment variables into a model or simply trying as many variables as a researcher can possibly think off. Such practices typically improve fit, but can invalidate inference. Within the proposed framework, the flexible model uses the exact same inferential reasoning in terms of which explanatory variables are included as the originally hypothesized model. Much of the typical criticism regarding blind 'fit-hunting', where all causal or 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, p. 412).¹⁷ Improvements are solely driven by more appropriate specification of the variables in the model.

Putting the second step into practice, I compare the out-of-sample performance of the GB model identified in the previous step with the two hypothesized alternatives for the toy example. As mentioned above, I include two fit metrics: the out-of-sample R^2 and the RMSE, and evaluate both using 1,000 splits of the dataset into estimation and evaluation sets. The results are summarized in Table 3. The top half of the table reflects the RMSE, whereas the lower half reflects the R^2 . For both metrics, the flexible model identified by the Super Learner, and the two hypothesized models are evaluated. Again for illustration's sake, the true piece-wise model is included as well. The bottom three rows for both metrics show the difference in the fit metric between the flexible model and the hypothesized models. As can be seen, the flexible model strongly outperforms the linear and quadratic models, but does not improve on the true model as should be expected. This leads us to conclude there is substantial fit improvement possible for the two hypothesized models.¹⁸

Step III: Unpacking the ML model

The third and final component of the proposed framework is a method to generate understanding into the ML model when it outperforms the hypothesized model in a

¹⁷A 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 invalidates any inference as it is a result (i.e., post-treatment) of the very process of interest. Although this example showcases accurately that a singular focus on fit is problematic, it mostly emphasizes that any analysis of an inferential question still requires causal or inferential logic.

¹⁸As mentioned, the 1,000 repeated estimations of the hypothesized models has the added benefit of providing bootstrapped standard errors of the coefficient estimates which are presented in Figure A.1 in the Appendix, together with the estimated coefficient on the full set. As can be seen, the reported coefficient lies at the extremes of the bootstrapped distribution of β 's for the two misspecified models.

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	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	0.026	0.007	0.004	0.015	0.039	0.049
GB vs. Quadratic	1000	0.022	0.007	-0.001	0.011	0.034	0.041
GB vs. PieceLinear	1000	-0.003	0.003	-0.015	-0.007	0.002	0.006

Table 3. Three linear additive models are estimated on a train-set consisting of 80% of the total dataset. The three linear additive models assume the effect of age to be respectively linear, quadratic or piece-wise linear. 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 (Table 2). RMSE and OOS R^2 is calculated for each model based on the remaining 20% of the data. The last three rows of the RMSE and R^2 parts of the Table 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-set are evaluated.

way that warrants re-evaluation of the originally hypothesized functional form $\tilde{f}(\cdot)$.¹⁹ Increasingly, researchers and practitioners have called for better understandable ML models given their broad application in everyday life and the risks a lack of understanding might bring when deploying ML in practice.²⁰ These developments have led to the emergence of X-AI, which focuses on understanding *why* ML models generate the predictions they do (Ribeiro et al., 2016; Samek et al., 2019). Two approaches can be identified: global and local explainability. The former attempts to describe the mechanics of a model in general terms, i.e., which variables tend to be important

¹⁹As mentioned above, a lack of model fit need not be problematic in and of itself given a specific research question. In other words, the decision to unpack the flexible form model depends on the researcher’s assessment of whether the lack of fit is potentially problematic and whether or not they feel comfortable defending an apparent lack of fit in the proposed model. However, as the flexible model includes the exact same variables as the originally hypothesized model, clear improvements in model fit will reflect improved estimation of the relationship amongst variables which can be expected to impact inference.

²⁰Common examples include unnoticed ethnic profiling, or models fit to aspects that do not generalize outside of the sample used (Subrahmanian and Kumar, 2017; Lundberg and Lee, 2017). As a case in point, recent legislation in the European Union now increasingly requires companies that use algorithms to support decision-making to be explainable, the so-called ‘Right to Explanation’ (Goodman and Flaxman, 2017).

for the model’s overall performance. Variable importance measures are an example of this approach (Baćak and Kennedy, 2019; Brand et al., 2021). Local explainability takes a more detailed approach and attempts to explain the drivers of individual predictions made by a model. As global explainability only gives a high-level understanding of which variables are important to model fit rather than an understanding into the functional relationship or direction of associations amongst variables, local explanation methods are more appropriate for the framework.

There are many local explanation methods available, all of which attempt to provide understanding into why a ML model makes the predictions it does. Which method is more appropriate fundamentally depends on the substantive question being asked of the model (Doshi-Velez and Kim, 2017) and interpretability methods tends to differ across substantive and expert domains (Zhou et al., 2021; Lipton, 2018).²¹ For example, much explainability research is focused on providing practitioners making (high-stakes) decisions with additional context into the model’s predictions. In such cases, the ability to quickly calculate and extract information from an explanation could be more relevant than emphasizing fidelity. Conversely, an ethical reviewer of a system-in-action might require more fine-grained and detailed explanations. In low-stakes cases with well-defined performance metrics, this may again be different.

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.²² Recent advances have made their estimation computationally more feasible and rekindled interest into their use more broadly (Aas et al., 2021; Samek et al., 2019; Heskes et al., 2020). More importantly, they are shown to align closely with human intuitions of model interpretation making them appropriate for the iterative process envisioned in the proposed framework (Doshi-Velez and Kim, 2017; Lundberg and Lee, 2017; Rudin, 2019). Shapley values have been applied in various fields, notably medicine

²¹The 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. Temporal efficiency and alignment with operational processes, which sometimes constrain explainability when deployed as a decision support system, is not relevant in this setup.

²²Another 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 paper.

Figure 4. Prediction $\hat{y}$ and mean value y_0 based on a model with four explanatory variables. The difference between the prediction and overall mean is decomposed along the Shapley values of each of the four variables.

(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) as well as to resolve path dependencies in decomposition metrics like segregation indices (Elbers, 2021).

Shapley values for model explanation

Shapley values — originally developed within cooperative game theory (Shapley, 1953) — decompose each prediction made by a model into a sum of K Shapley values, one for each variable in the model:

$$\hat{y}_i = \phi_0 + \sum_{k=1}^M \phi_k x_{k,i}, \quad (3)$$

where $\hat{y}_i$ is a prediction made by a model, and ϕ_0 resembles the overall mean across predictions. The Shapley values thus decompose the deviance of $\hat{y}_i$ with respect to the mean prediction in an additive way, leading to the typical waterfall plots associated with Shapley values illustrated in Figure 4. As every prediction is decomposed into K Shapley values, the approach effectively leads to a matrix of $N \times K$ values, one for every variable x_{ki} in the model. Thus, every observed value of x_{ki} has an associated Shapley value ϕ_{ki} reflecting that variable’s impact on a specific outcome $\hat{y}_i$. In a linear additive model without an intercept, the Shapley value for variable k would simply be the coefficient β_k times the observed value x_{ki} — i.e., $\phi_{ki} := \beta_k x_{ki}$ (Aas et al., 2021). Plotting pairs of Shapley values and the values of their associated explanatory variable $[\phi_{ki}, x_{ki}]$ thus provides a visual representation of the global functional relationship between x_k and the outcome which can be used to generate understanding of the

456 patterns encoded in the ML model (Lundberg et al., 2020, p. 59).

457 Calculating the Shapley value ϕ_k for variable k given some prediction $\hat{y}_i$ relies on a
 458 perturbation-based approach where the impact of omitting information on variable k
 459 in the model’s prediction is assessed.²³ This is done by defining all possible informa-
 460 tion sets — which is simply a set of $s \leq K$ variables for which information is available
 461 to the model — including variable k , $\mathcal{M}^k$, and its complement where each information
 462 set excludes information on variable k , $\mathcal{M}^k/k$. For each of the information sets in
 463 $\mathcal{M}^k$, a prediction is made using that set as well as its equivalent in $\mathcal{M}^k/k$ leading to
 464 two predictions — one including information on k and one without — which are then
 465 differenced.²⁴ The Shapley value ϕ_k is defined as the weighted mean over all these
 466 differences in predictions for each information set in $\mathcal{M}^k$:

$$\phi_k = \underbrace{\sum_{\mathcal{I} \in \mathcal{M}^k}}_{\text{Sum over all information sets}} \underbrace{\frac{|\mathcal{I}|!(|\mathcal{M}| - |\mathcal{I}| - 1)!}{\mathcal{M}!}}_{\text{Weighting function}} \overbrace{(\bar{f}(\mathcal{I} \cup k) - \bar{f}(\mathcal{I}))}^{\text{Difference in prediction}}. \quad (4)$$

467 This process is then repeated for every observation and variable in the dataset leading
 468 to the $N \times k$ matrix of Shapley values.

469 Shapley values have a number of unique theoretical properties that make them at-
 470 tractive as a method to generate understanding into models. First, the sum of all
 471 Shapley values ϕ_k and the mean prediction, ϕ_0 , match the actually observed predic-
 472 tion $\hat{y}_i$. Second, whenever the inclusion of variable k into the information set has no
 473 impact on the model’s prediction — $\bar{f}(\mathcal{I}) = \bar{f}(\mathcal{I} \cup k)$ for all $\mathcal{I}$ — its Shapley value
 474 is zero. Third, whenever two features j and k contribute equally to every prediction
 475 — $\bar{f}(\mathcal{I} \cup j) - \bar{f}(\mathcal{I}) = \bar{f}(\mathcal{I} \cup k) - \bar{f}(\mathcal{I}) \quad \forall \mathcal{I} \in \mathcal{M}^k$ — their Shapley values are the
 476 same: $\phi_j = \phi_k$. Fourth, they are consistent with respect to addition and multiplica-
 477 tion across models. Shapley values have been shown to be the only local explanation
 478 method to possess these properties (see Shapley (1953); Young (1985); Lundberg and
 479 Lee (2017); Lundberg et al. (2020) for detailed discussions and proofs).

480 Calculating Shapley values poses a number of computational challenges that have

²³The approach has many resemblances to the *do*-operator popularized by Pearl (2009), as it effectively concerns intervention on the covariate space (Lundberg and Lee, 2017; Heskes et al., 2020).

²⁴Consider calculating the Shapley value for variable k in a model $\hat{f}(\cdot)$ which is estimated to four explanatory variables. First, all 2^{K-1} possible information sets containing combinations of the remaining $K - 1$ variables have to be considered. When $K = 4$ we consider 8 unique information sets of the three other variables, ranging from including none of the variables — $\mathcal{I} = [0, 0, 0]$ — to including all three — $\mathcal{I} = [1, 1, 1]$ — into the prediction. For each information set $\mathcal{I}$, the difference in prediction is calculated when adding information on the variable k — i.e., the observed x_k — to the prediction: $\bar{f}(\mathcal{I} \cup k) - \bar{f}(\mathcal{I})$ where $\bar{f}(\cdot)$ is the flexible model.

limited their applicability in the past. This computational burden stems from three
 elements. First, the number of possible information sets is 2^K and scales exponen-
 tially with the number of features, quickly becoming intractable. 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 $E[x|x^* = x_s]$ 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 explained to assure
 sufficient tuples $[\phi_k, x_k]$ to infer the underlying functional form, thus further scaling
 the computational requirements linearly. In recent years, computationally efficient
 methods have been developed to calculate Shapley values, notably the Kernel SHAP
 approach due to the seminal paper by Lundberg and Lee (2017). Instead of consider-
 ing the entire space of information sets, Kern SHAP implements a sampling approach
 to approximate the exact Shapley values (see the Appendix for a more detailed dis-
 cussion).

In this paper, I use the Shapley value estimation method optimized for tree-based
 models, Tree SHAP, due to Lundberg et al. (2020). By exploiting the structure of
 the estimated tree, a much smaller set of relevant information sets can be deter-
 mined. Specifically, those information sets that only differ in terms of variables that
 do not feature in branches of the tree can be ignored as they will not lead to differ-
 ent predictions per definition. This has made estimation efficient to the point that
 exact Shapley values can be calculated, rather than having to rely on numerical ap-
 proximations (Lundberg et al., 2020, p. 64-65) and even renders pairwise Shapley
 values feasible to calculate in reasonable time (Lundberg et al., 2020; Lapuschkin
 et al., 2019). Pairwise Shapley values assess the joint omission of *two* variables from
 the model and the subsequent impact on the outcome. In the standard Shapley de-
 composition in (3), interactions would be divided amongst the relevant interacting
 variables.²⁵ 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 detailed discussion of the Tree SHAP approach and pairwise
 Shapley values is provided in the Appendix.

To illustrate how Shapley values can be used to generate understanding into a ML
 model, I calculate the Shapley values for the age variable in the GB model identified in

²⁵Interactions can still be identified from visualizing Shapley values for a variable x_k separately,
 conditional on values of another variable x_j (as is illustrated in the first empirical example). However,
 they do not explicate interactions.

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

the toy example above. This leads to N Shapley values associated with the N values of x_{age} . A joint scatterplot is presented in Figure 5. These show the tuples $[\phi_{ki}, x_{ki}]$ for all N observations in the data, where ϕ_{ki} reflects the contribution of variable x_{ki} to the prediction $\hat{y}_i$ relative to ϕ_0 . As the underlying model is linear additive, these Shapley values should represent the underlying coefficient β_k times the observed value of the covariate x_{ki} for each observation. Note that I normalize the Shapley values as well as the true underlying effect such that they can be visualized jointly.²⁶ As can be seen, the Shapley values accurately recover the piecewise linear association in the underlying data, and would provide us with concrete insights into the way in which our hypothesized functional forms could be improved from the linear and quadratic specifications of age. Again, no functional form specification was required to generate these insights.

4 Resemblances to Other Computational Frameworks

Before applying the framework to three empirical use 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 in-

²⁶The Shapley value for variable k reflects the impact of x_k into the information set relative to ϕ_0 . In linear regression including an intercept, interpretation is relative to the overall mean.

debted 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.²⁷ Such approaches can identify a broader range of non-linear functional forms than those for example considered by the RESET test, 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 local estimates scales exponentially with the number of independent variables (Bishop, 2006; Yatchew, 2003). The ML methods in the framework retain considerably more structure and are more efficient from a data perspective than non-parametric tests, improving efficiency and allowing for better post-estimation analysis.²⁸

The proposed framework also shares similar goals to the field of model robustness. The emphasis in model robustness lies on 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 amongst variables (Young and Holsteen, 2017).²⁹ Model robustness primarily emphasizes inference into a single parameter, e.g. β_k , and its consistency and substantive size across the exclusion of (sets of) variables. The functional complexity as well as the fit of the models is of limited importance (Slez, 2019; Young, 2019). The proposed framework instead focuses on the functional relationship amongst a set number of variables rather than variable inclusion: the emphasis is on specification. The focus of model robustness on excluding variables to generate ‘plausible’ specifications also raises questions whether the various alternative specifications are equally

²⁷Typically 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, Chapter 1).

²⁸Note that 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 are even applicable for very small N cases (Liao et al., 2017; Athey, 2018).

²⁹Muñoz and Young (2018) 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)’ (p. 8). As the proposed framework shows, there is no reason to exclusively implement ML in high dimensional settings.

plausible in practice (Slez, 2019; O’Brien, 2018).³⁰ The proposed framework could of course easily be combined with model robustness by applying the approach to different sets of explanatory variables, as is illustrated in the two empirical case studies in this paper. Despite their differences, both frameworks are similarly indebted to the growth in computational power to enable researchers to become more critical of modeling assumptions and to improve transparency into the research process (Muñoz and Young, 2018, p. 4).³¹

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 (Kleinberg et al., 2017; Agrawal et al., 2020; 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. In case the ML method leads to better predictions than a behavioral theory, those 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 this paper proposes. 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.³² However, the central premise of using ML to find patterns in data not necessarily hypothesized by a researcher is a key resemblance amongst the two approaches. More generally, the framework is aligned with an increasing interest in finding the predictive potential of social theories — and using ML to do so — that can be observed in sociology more generally (Watts, 2014; Salganik et al., 2020).

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 hypoth-

³⁰For example, if an important confounder is omitted in 50% of the alternative ‘reasonable’ specifications, we shouldn’t be surprised to see high variation in the estimates of the coefficient of interest.

³¹Model robustness has recently been expanded to ‘Multiverse’ analysis by Young and Stewart (2021) who, in addition to varying the number of variables in the model, further include 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.

³²Effectively, one would abort the framework after the second step and instead investigate which observations tend to be fitted better by the ML model. A similar approach to studying predictions has been suggested in Salganik et al. (2020) and Verhagen (2022).

esizing might yield. The autometrics approach is to include 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 in this paper. 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).³³ Related, some advocate a similarly stacked model with interactions but apply variable selection through least absolute shrinkage and selection operator (LASSO) regression instead (Blackwell and Olson, 2021; Beiser-McGrath and Beiser-McGrath, 2020). The proposed framework shares the same broad intuition of the above approaches, but uses a broader set of methods such that a wider range of patterns can be identified. The proposed framework further emphasizes out-of-sample evaluation over in-sample evaluation, reducing the risk of overfitting and path dependence in eliminating regressors from the model.

5 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 studies 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 which has been actively innovated over the past decades and provides an illustration of how the proposed framework can speed up model-building. The hedonic regression and voting examples are chosen because both enjoy considerable academic interest and models typically include a number of standard control variables with a (novel) interest variable. Since the latter is often correlated with the control variables, a correct functional relationship is crucial for inference. For these latter two case studies I will evaluate whether the typical complexity in which control variables feature in the functional form is sufficient, and improve them where necessary.

³³The 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).

Each of these use cases is meant to illustrate the potential of the proposed framework to identify more intricate patterns in the data. In the first, both a non-linearity and interaction effect are identified using the framework. In the second, the framework is used to identify considerable non-linearities in a number of explanatory variables, as well as that spatial heterogeneity is poorly controlled for using neighborhood variables. In the third, the potential of pairwise Shapley values is used to identify a wide range of interactions amongst variables. Note that none of the examples are designed with a specific substantive research question in mind.³⁴

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 so-called ‘return to education’ (Lemieux, 2006). As mentioned above, this example is appropriate to showcase the framework since the functional form used to estimate the Mincerian wage equation has been innovated over the past decades. In other words: model specification has improved. The original functional form simply related log yearly wages with years of education and work experience:

$$\ln(\text{wages}_i) = \beta_0 + \beta_1 x_{\text{educ},i} + \beta_2 x_{\text{exp},i} + \epsilon_i. \quad (5)$$

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

$$\ln(\text{wages}_i) = \beta_0 + \beta_1 x_{\text{educ},i} + \beta_2 x_{\text{exp},i} + \beta_3 x_{\text{exp},i}^2 + \epsilon_i. \quad (6)$$

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

$$\begin{aligned} \ln(\text{wages}_i) = & \beta_0 + \beta_1 x_{\text{educ}_0-8,i} + \beta_2 x_{\text{educ}_9-10,i} + \beta_3 x_{\text{educ}_{11-12},i} + \\ & \beta_4 x_{\text{educ}_{13-14},i} + \beta_5 x_{\text{educ}_{15+},i} + \beta_6 x_{\text{exp},i} + \beta_7 x_{\text{exp},i}^2 + \epsilon_i. \end{aligned} \quad (7)$$

³⁴Code and instructions to obtain the data in order to reproduce each of the empirical examples in this paper, including the toy example introduced in Section 2, is available in this paper's online GitHub repository.

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 4. Descriptive statistics of the simulated wage data

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

$$\begin{aligned}
\ln(\text{wages}_i) = & I(x_{\text{sex},i} = \text{Female})[\beta_0^* + \beta_1^*x_{\text{educ}.0.8,i} + \beta_2^*x_{\text{educ}.9.10,i} + \beta_3^*x_{\text{educ}.11.12,i} + \\
& \beta_4^*x_{\text{educ}.13.14,i} + \beta_5^*x_{\text{educ}.15+,i} + \beta_6^*x_{\text{exp},i} + \beta_7^*x_{\text{exp},i}^2 + \epsilon_i] + \\
& I(x_{\text{sex},i} = \text{Male})[\beta_0 + \beta_1x_{\text{educ}.0.8,i} + \beta_2x_{\text{educ}.9.10,i} + \beta_3x_{\text{educ}.11.12,i} + \\
& \beta_4x_{\text{educ}.13.14,i} + \beta_5x_{\text{educ}.15+,i} + \beta_6x_{\text{exp},i} + \beta_7x_{\text{exp},i}^2 + \epsilon_i].
\end{aligned} \tag{8}$$

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. I vary the coefficients of sex to fall within a standard error of the coefficients found in the literature to generate the fourth DGP (see this paper’s Appendix) (Lemieux, 2006; Heckman et al., 2018).

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 used is based on the GSS 2018, such that the distribution of the explanatory variables is representative for 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 Table 4. Linear-I refers to the first functional form above provided by (5), Linear-II refers to (6) and so on.

As a result, there are four datasets consisting of a vector of outcomes y and the same matrix of explanatory variables X . Each outcome vector comes from a different underlying functional relationship of X and y . The first is based on a world where

both education and work experience affect the outcome linearly, the second where work experience follows a second degree polynomial 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 (5)-(7), as well as a fourth functional form which is equal to (7) but includes a dummy for sex. This means that none of the four hypothesized models aligns with the true DGP in the fourth dataset, which follows the DGP provided in (8). I include this fourth form for illustrative purposes, as it is common practice to ‘control’ for group differences by simply including a dummy variable whereas this might be simplifying the true pattern in the data.

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 form 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. In addition to the four hypothesized functional forms, I estimate a Super Learner including various tree-based ML methods (Tables A.2 and A.3 in the Appendix) — the first step of the framework. As can be seen, the GB model performs best across all four datasets. For each dataset, the best performing model is selected as the flexible model (indicated in boldface). Next, when comparing the out-of-sample model fit of the flexible model with the four hypothesized models — Step II in the framework — 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 the fourth dataset (Figure 6). The framework thus identifies the underlying model without requiring the researcher to hypothesize a functional form in each of the 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 under-specified functional forms, but even if we didn’t; the Shapley values easily identify the correct association of the independent variables with the outcome (Figure 7). As can be seen, the Shapley values capture the linearity of both explanatory variables in the first dataset, the non-linearity in years of work experience in the second dataset, as well as the step-wise function in the third dataset. What is more, the Shapley values pick up the interaction with sex in the fourth dataset, as is illustrated by varying the colors of the Shapley values by sex. Again, without requiring the researcher to formulate any functional relationship between variables a-priori.

Figure 6. Out-of-sample R^2 for the four datasets with varying DGPs, using four functional forms and the GB model. As can be seen, the GB model converges on the correct functional form's model fit in each of the datasets, whereas the pre-specified models only reach the maximum explanatory power when the functional form's complexity equals or outperforms that of the DGP. 100 splits into an estimation set of 80% and evaluation set of 20% are used.

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

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 to be 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 lies in the price elasticity of certain traits, e.g. 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 that is implicit in this functional form might be unreasonable (Malpezzi, 2003; Fan et al., 2006).

In this application I use typical house characteristics, like the 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 centres 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 & Local Government.³⁵ I also include the year and month of the transaction. Descriptive statistics of the dataset can be found in Table 5.

As the hypothesized model, I estimate a typical linear additive functional form relating log house prices to the variables depicted in Table 5. As is customary, I assume linear trends for all continuous variables, as well as for the temporal variables. The exact functional form is as follows:

$$\begin{aligned} \ln(\text{House price}_i) = & \beta_0 + \beta_1 x_{\text{area},i} + \beta_2 x_{\text{rooms},i} + \beta_{3,\dots,5} x_{\text{property type},i} + \beta_6 x_{\text{new},i} + \\ & \beta_7 x_{\text{travel.time},i} + \beta_8 x_{\text{crime},i} + \beta_9 x_{\text{deprivation},i} + \beta_{10} x_{\text{year},i} + \\ & \beta_{11} x_{\text{month},i} + \beta_{12,\dots,16} x_{\text{transaction.type},i} + \epsilon_i. \end{aligned} \quad (9)$$

³⁵I include these indices at the Lower Super Output Area level, for which the average population in London is around 1700.

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 centre (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

Table 5. Descriptive statistics of the London house price data

Following the framework, I start by estimating both the hypothesized model and a Super Learner to the data. The best fitting model is again a GB approach, which will be used as the flexible model (see Table A.4 in the Appendix, selected model in boldface). In the next step, the flexible model is compared in terms of its out-of-sample fit with the hypothesized model. The results are summarized in Figure 8 and are striking. Already amongst the simplest housing characteristics like size and property type the ML model improves on the linear additive functional form by about 5.5 pp in terms of the out-of-sample R^2 . However, the largest improvement by far is observed when adding neighborhood variables like ‘travel time to hub’ to the model, with the difference in R^2 more than 30 pp. This strongly implies that spatial heterogeneity is poorly addressed by the functional form in (9), a point to which I will return below.

To evaluate why the flexible model outperforms the linear additive framework, I calculate Shapley values for the flexible model next. These are visualized for six explanatory variables: the size of the house, the travel time to a local hub, the crime index, number of rooms in the house, and the two temporal variables (year and month of sale). The results can be found in Figure 9. As can be seen, non-linearities are observed 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 modelled in a non-linear way. When including these non-linearities into the linear additive model, the fit improves and is strictly preferred according to a LR

Figure 8. Out-of-sample R^2 for the hypothesized model and the flexible model. ‘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 to a local hub of the neighborhood, respectively. 100 splits into an estimation set of 80% and evaluation set of 20% are used.

test³⁶ although a remaining gap in model fit still points at further interactions and non-linearities amongst 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, and including non-linearities 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 ML 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 greatly improves to 84%, which is much closer to that

³⁶Adding squared versions of the neighborhood indicators, area, number of rooms and temporal variables improves the adjusted R^2 by 2.5 pp. A LR test accordingly prefers the updated model ($P < 0.001$).

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). A LOESS fit is included into the Shapley plots to indicate the association between each variable and the outcome. Implied effects are scaled to allow for visual comparisons.

of the ML model. In other words, the heterogeneity amongst 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, it seems like the ML model’s ability to estimate flexible patterns has led to the identification of individual neighborhoods thus mimicking a random intercept approach. The framework illustrates that spatial heterogeneity is highly predictive, but poorly accounted for by the available neighborhood characteristics (Elhorst, 2010).

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 has been of substantive interest in the social sciences (Freedman et al., 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. Usual suspects include the respondent’s age, sex, race, educational attainment and income. Typically, another variable of substantive interest is included into the analysis, e.g.

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	1996.00	13.46	1974	1987	2010	2018

Table 6. Descriptive statistics of the GSS

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. The GSS also encompasses many waves and spans multiple decades. Temporal changes might thus be relevant as well.

Table 6 shows descriptive statistics of the GSS containing information on voting preferences and demographic characteristics between 1974 and 2018. The outcome of interest is a 7-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. In addition, the year of the survey wave is available. As hypothesized model, I use a standard linear additive model as often encountered in the literature (Freedman 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:

$$y = \beta_0 + \beta_1 x_{\text{age}} + \beta_2 x_{\text{female}} + \beta_3 x_{\text{black}} + \beta_4 x_{\text{other}} + \beta_7 x_{\text{education}} + \beta_8 x_{\text{income}} + \beta_9 x_{\text{survey_year}} + \epsilon_i \quad (10)$$

and resembles that found in e.g. Meisenberg (2015).

I start by estimating (10) as well as a Super Learner containing various RF and GB models. The best performing model is again the GB model, although the RF also performs well (see Table A.5 in the Appendix, selected model in boldface). I use the GB approach as the flexible model. The out-of-sample R^2 when including different sets of explanatory variables to the hypothesized and flexible model is visualized in Figure 10 and show that the explanatory power of the flexible model is almost double that of the hypothesized model indicating a considerable lack of appropriate

Figure 10. Out-of-sample R^2 for the hypothesized model as compared to the flexible model. 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 model from the Super Learner routine has been estimated to 80% of the data, and the out-of-sample R^2 is calculated on the remaining 20%. The data has been split 1,000 times into estimation- and evaluation-sets.

specification. The flexible model improves most considerably when race is added to the model, although the hypothesized model already underperforms when simply including temporal, age, and sex variables. Note also that overall fit is comparatively low.

To understand why the GB model outperforms the linear additive model, I illustrate both Shapley values and Shapley interaction values for the age, schooling, and income variables in Figure 11. These Shapley values show a number of interesting dynamics. First of all, there are clear non-linearities 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 11, panel A).³⁷

These direct effects can be further decomposed into a number of interactions by estimating the pairwise Shapley values discussed earlier. These interactions allow Shapley values for a single variable to be decomposed into a direct and indirect effect. The indirect effects indicate how Shapley values for certain observed characteristics

³⁷Adding square terms to all three variables and separate intercepts for each survey year improves the R^2 from 6.1% to 7.0%. Unsurprisingly, a LR test comparing both models prefers the updated model ($P < 0.001$).

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

deviate from the overall effect of the variable conditional on the height of the second variable. The estimates in panels B-E all show indirect effects and illustrate that younger respondents associated less strongly with the Democrats in earlier waves than in later waves (Figure 11, Panel B), but also that Whites have a stronger positive age effect than Non-whites (Figure 11, Panel C). Similarly interesting dynamics can be observed when interacting sex and race, which show that the sex effect is more pronounced for Blacks but is less pronounced for Others (Figure 11, Panel D). The effect of income shows similarly complicated dynamics where higher income is associated more strongly with a Republican identification for Whites than for Non-whites (Figure 11, Panel E). 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.³⁸

6 Discussion & Concluding Remarks

This paper set out to incorporate ML methods into the empirical workflow of sociologists to address a key problem: that we often don't have a clear indication what the appropriate functional form might be to model a dataset. Coupled with a historic preference for parsimonious and easy-to-estimate models, this has led to a prevalence of relatively simplistic functional forms that run the risk of lacking the appropriate flexibility to study the social world. In addition, a lack of emphasis on developing appropriate specifications prolongs the risk posed by researcher degrees of freedom in muddying empirical findings. That many functional forms are often simplistic is well-known, and criticism of the linear additive framework has been consistent, but remains without a clear pathway to improve. Classic misspecification tests are known to be underpowered in multivariate settings, and often limited to specific types of misspecification. They lack guidance on how to improve misspecification going forward and, perhaps most importantly, their application in empirical work varies strongly and can be prone to cherrypicking.

I propose 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 modelled when the functional relationship amongst variables is dictated by the data rather than a researcher's hypothesizing. Such an estimate provides an indication whether a researcher's own functional form might miss important nuances like in-

³⁸Adding age by race and sex by race interactions further improves the adjusted R^2 to 7.2%. A LR test prefers the updated model again ($P < 0.001$). Finally, adding income by race improves the adjusted R^2 by another 0.1 pp with a LR test also preferring the added flexibility to the functional form ($P < 0.001$).

842 teractions or non-linearities. Crucially, the fit potential is a feature of the data and
843 not a result of choices on the part of the researcher, improving transparency into
844 the empirical process. The framework also uses out-of-sample model evaluation and
845 repeated estimation to guard against over-fitting. Most importantly, in case a ML
846 method finds more intricate patterns in the data than our own models do, we can
847 unpack the former to provide guidance on how to improve the latter. This is contrary
848 to popular belief that ML models are fundamentally black boxes which cannot convey
849 intuition into the patterns they identify.

850 The approach presented in this paper is not a panacea, and has a number of limi-
851 tations. First, uncovering intricate patterns from the data will still depend on the
852 available data and might provide limited guidance in low N settings, although many
853 of the methods proposed here can be applied to datasets typically encountered by
854 sociologists. Second, although considerable progress has been made in recent years,
855 unpacking flexible ML methods remains challenging. This is in part because patterns
856 can become complex to the degree that even local explanation methods like Shapley
857 values will not yield easily digestible insights into the underlying functional form,
858 especially when multiple interactions may be at play. More generally, active debate
859 regarding the practical as well as philosophical aspects of explanation methods' ac-
860 curacy to reflect how ML methods operate remains ongoing. However, the constant
861 developments in the X-AI field are both reassuring and exciting. Finally, when causal
862 questions rather than optimal fit are of interest, there is limited guidance whether
863 a difference in model fit between a hypothesized model and ML alternative is actu-
864 ally problematic. That having been said, the framework is not a blind search for fit
865 across a host of variables but follows the same inferential logic as a researcher's own
866 model. Therefore, any improvement in fit should reflect missed functional relation-
867 ships amongst variables that have been purposefully curated by a researcher and their
868 functional relationship should thus be expected to affect inference.

869 Already in its current form, I have shown how the framework can be fruitfully ap-
870 plied in practice. Using the canonical example of the Mincerian wage equation, I
871 show how the historic process of model-development could have been sped up con-
872 siderably. The framework effortlessly identified underspecification, and subsequent
873 analysis of the ML models identified the necessary improvements to the functional
874 form. The other two empirical examples, a hedonic regression of house prices and a
875 model explaining party identification, both illustrate how often encountered model-
876 ing strategies lead to considerably lower fit than a flexible ML model. In the case
877 of house prices in London, a number of simple non-linearities were identified. Most
878 importantly, the fundamental inability to address spatial heterogeneity through the

inclusion of standard neighborhood characteristics became clear through the immense 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 non-linearities between the explanatory variables and outcome, which are not commonly implemented in the functional forms used to study party identification.

At the root of almost every empirical sociological finding lies a functional form which is assumed to be correctly specified. There is limited evaluation whether this functional form is appropriate for the data, leading to a number of serious risks. The model might not accurately reflect the patterns in the data, affecting the validity of statistical inference. Simple reporting of a single or curated number of functional forms further exposes sociology to p-hacking. These practices stem from a state of the world where 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. This paper proposes 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. A symbiosis of quantitative sociological work and the computational riches of this day and age that is long overdue, and will only bear more fruit as the goals of the ML community increasingly align with the explanatory focus of sociologists.

Appendices

Figures

Figure A.1. Bootstrapped distributions of the coefficient of interest from Step II of the framework. A natural result of estimating the fit of the hypothesized model over a large amount of resampled sets is that it allows for insights into the variability of the found coefficients across samples. In this case, 80% of the data is randomly sampled 1,000 times to generate these distributions.

Tables

Model	Ave	SD	Min	Max
<i>SL.glm.Step</i>	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
<i>SL.glm.Quadratic</i>	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
<i>SL.glm.Linear</i>	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

Super Learner performance based on RMSE & 10-fold CV.

Selected model in boldface. Hypothesized models in italic.

GB parameters: `n.rounds_max.depth.eta`.

RF parameters: `n.trees_mtry`.

Table A.1. Super Learner performance, simulated data.

Model	Ave	SD	Min	Max
<i>SL.Linear</i>	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

Super Learner performance based on RMSE & 10-fold CV.
Selected model in boldface. True model in italic.
GB parameters: `n_rounds_max_depth_eta`.
RF parameters: `n_trees_mtry`.

Model	Ave	SD	Min	Max
<i>SL.Linear</i>	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
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

Super Learner performance based on RMSE & 10-fold CV.
Selected model in boldface. True model in italic.
GB parameters: `n_rounds_max_depth_eta`.
RF parameters: `n_trees_mtry`.

Table A.2. Super Learner performance, Mincerian wage simulation Linear-I (top) and Linear-II (bottom).

Model	Ave	SD	Min	Max
<i>SL.Linear</i>	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

Super Learner performance based on RMSE & 10-fold CV.
Selected model in boldface. True model in italic.
GB parameters: `n_rounds_max_depth_eta`.
RF parameters: `n_trees_mtry`.

Model	Ave	SD	Min	Max
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
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
<i>SL.Linear</i>	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

Super Learner performance based on RMSE & 10-fold CV.
Selected model in boldface. Hypothesized model in italic.
GB parameters: `n_rounds_max_depth_eta`.
RF parameters: `n_trees_mtry`.

Table A.3. Super Learner performance, Mincerian wage simulation Linear-III (top) and Linear-IV (bottom).

Model	Ave	SD	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.GB_500_5_0.4	0.254	0.001	0.252	0.255
SL.RF_500	0.369	0.001	0.366	0.370
<i>SL.GLM</i>	0.421	0.002	0.419	0.423

Super Learner performance based on RMSE & 10-fold CV.

Selected model in boldface. Hypothesized model in italic.

GB parameters: `n_rounds_max_depth_eta`.

RF parameters: `n_trees_mtry`.

Table A.4. Super Learner performance, London house price data.

Model	Ave	SD	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
<i>SL.GLM</i>	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

Super Learner performance based on RMSE & 10-fold CV.
Selected model in boldface. Hypothesized model in italic.
GB parameters: `n_rounds_max_depth_eta`.
RF parameters: `n_trees_mtry`.

Table A.5. Super Learner performance, GSS data.

Simulations

Non-linear DGP's

Below the four DGPs used to generate the true data in Figures 1 and 3 are provided:

$$y_{\text{linear}} = -1 + 0.9X + \epsilon_1 \quad (\text{A.1})$$

$$y_{\text{polynomial}} = -6 + 0.4X - 0.36X^2 + 0.005X^3 + \epsilon_2 \quad (\text{A.2})$$

$$y_{\text{sine}} = 2.83\sin\left(\frac{\pi}{2X}\right) + \epsilon_3 \quad (\text{A.3})$$

$$y_{\text{step}} = -2\mathbf{I}(X > 0) + 4\mathbf{I}(X > 2) - \mathbf{I}(X > 3) + \epsilon_4 \quad (\text{A.4})$$

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

Mincerian wage equation simulation

Below the four DGPs used to generate the four Mincerian wage outcomes in Figure 6 and 7 are provided:

Linear-I

$$\begin{aligned} \ln(\text{wages}) &= 4.5 + 0.125x_{\text{years_educ}} + 0.09x_{\text{years_exp}} + \epsilon \\ \epsilon &\sim \mathcal{N}(0, 0.07) \end{aligned} \quad (\text{A.5})$$

Linear-II

$$\begin{aligned} \ln(\text{wages}) &= 4.5 + 0.125x_{\text{years_educ}} + 0.09x_{\text{years_exp}} - 0.001x_{\text{years_exp_squared}} + \epsilon \\ \epsilon &\sim \mathcal{N}(0, 0.07) \end{aligned} \quad (\text{A.6})$$

Linear-III

$$\begin{aligned}\ln(\text{wages}) &= 4.5 + 0.02x_{\text{years_educ_0_8}} + 0.03x_{\text{years_educ_9_10}} + 0.3x_{\text{years_educ_11_12}} + \\ &0.06x_{\text{years_educ_13_14}} + 0.06x_{\text{years_educ_15+}} + 0.09x_{\text{years_exp}} + \\ &0.001x_{\text{years_exp_squared}} + \epsilon \\ \epsilon &\sim \mathcal{N}(0, 0.07)\end{aligned}\tag{A.7}$$

Linear-IV

$$\begin{aligned}\ln(\text{wages}) &= I(x_{\text{sex}} = \text{Female})[3.5 + 0.025x_{\text{years_educ_0_8}} + 0.06x_{\text{years_educ_9_10}} + \\ &0.35x_{\text{years_educ_11_12}} + 0.06x_{\text{years_educ_13_14}} + 0.06x_{\text{years_educ_15+}} + \\ &0.1x_{\text{years_exp}} - 0.0005x_{\text{years_exp_squared}} + \epsilon] + \\ &I(x_{\text{sex}} = \text{Male})[5.5 + 0.015x_{\text{years_educ_0_8}} + 0.02x_{\text{years_educ_9_10}} + \\ &0.25x_{\text{years_educ_11_12}} + 0.04x_{\text{years_educ_13_14}} + 0.04x_{\text{years_educ_15+}} + \\ &0.06x_{\text{years_exp}} - 0.001x_{\text{years_exp_squared}} + \epsilon^*] \\ \epsilon &\sim \mathcal{N}(0, 0.09) \\ \epsilon^* &\sim \mathcal{N}(0, 0.08)\end{aligned}\tag{A.9}$$

Shapley values

In game theory, Shapley values are defined as follows. Consider a game $\mathcal{G}$ consisting of $\mathcal{M}$ possible players. Further assume that 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 have been 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 $v(\mathcal{S})$ amongst $\mathcal{S}$ but Shapley values are the only division that satisfy a number of properties discussed below. Shapley values are estimated as follows:

$$\phi_j(v(\mathcal{S})) = \phi_j = \sum_{\mathcal{S} \subseteq \mathcal{M}/j} \frac{|\mathcal{S}|!(M - |\mathcal{S}| - 1)!}{M!} (v(\mathcal{S} \cup j) - v(\mathcal{S})), \quad j = 1, \dots, M. \quad (\text{A.10})$$

Where each player j gets their part of the payoff, equal to ϕ_j .

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 $\mathcal{S}$ when including player j to the set of players, versus the payoff without including player j .

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

$$\begin{aligned} \phi_1 &= \frac{1}{4}(v([1, 2, 3, 4]) - v([2, 3, 4])) + \frac{1}{12}[(v([1, 3, 4]) - v([3, 4])) + \\ &\quad (v([1, 2, 4]) - v([2, 4])) + (v([1, 2, 3]) - v([2, 3])) + (v([1, 4]) - v([4])) + \\ &\quad (v([1, 2]) - v([2])) + (v([1, 3]) - v([3]))] + \frac{1}{4}(v([1]) - v([\emptyset])) \\ \phi_2 &= \frac{1}{4}(v([1, 2, 3, 4]) - v([1, 3, 4])) + \frac{1}{12}[(v([2, 3, 4]) - v([3, 4])) + \\ &\quad (v([1, 2, 4]) - v([1, 4])) + (v([1, 2, 3]) - v([1, 3])) + (v([1, 2]) - v([1])) + \\ &\quad (v([2, 4]) - v([4])) + (v([2, 3]) - v([3]))] + \frac{1}{4}(v([2]) - v([\emptyset])) \\ \phi_3 &= \frac{1}{4}(v([1, 2, 3, 4]) - v([1, 2, 4])) + \frac{1}{12}[(v([2, 3, 4]) - v([2, 4])) + \\ &\quad (v([1, 3, 4]) - v([1, 4])) + (v([1, 2, 3]) - v([1, 2])) + (v([1, 3]) - v([1])) + \\ &\quad (v([2, 3]) - v([2]))] + \frac{1}{4}(v([3]) - v([\emptyset])) \\ \phi_4 &= \frac{1}{4}(v([1, 2, 3, 4]) - v([1, 2, 3])) + \frac{1}{12}[(v([2, 3, 4]) - v([2, 3])) + \\ &\quad (v([1, 3]) - v([1, 3, 4])) + (v([1, 2, 4]) - v([1, 2])) + (v([1, 4]) - v([1])) + \\ &\quad (v([2, 4]) - v([2]))] + \frac{1}{4}(v([4]) - v([\emptyset])). \end{aligned} \quad (\text{A.11})$$

Note that by adding all four Shapley values together and defining $\phi_0 = v(\emptyset)$ 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 up to the outcome of the game. This is the so-called 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 (A.10) 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:

$$\hat{f}(x_i) = \phi_0 + \sum_{k=1}^K \phi_k, \quad (\text{A.12})$$

where $\phi_0 = \mathbb{E}(\hat{f}(x))$. In other words, the Shapley values ϕ_k capture the difference between $\hat{f}(x_i)$ 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 sub-sets $\mathcal{S} \subseteq \mathcal{K}$ of covariates instead of players. In other words, each player's contribution to the game's outcome in the game theoretical setup now represents each covariate's contribution to the prediction of a model. This requires evaluating the function $\hat{f}(\cdot)$ whilst excluding some covariates from the model. That is, replacing $v(\mathcal{S})$ in (A.10) with $\mathbb{E}[\hat{f}(x|x = x_{\mathcal{S}})]$. Generally speaking, calculating exact Shapley values is computationally challenging as the number of sets $\mathcal{S}$ is $2^{|\mathcal{K}|}$.

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

$$\sum_{\mathcal{S} \subseteq \mathcal{K}} (\mathbb{E}[\hat{f}(x|x = x_{\mathcal{S}})] - (\phi_0 + \sum_{k \in \mathcal{S}} \phi_k))^2 \mathbf{k}(\mathcal{K}, \mathcal{S}). \quad (\text{A.13})$$

In this setup, the function $k(\cdot)$ reflects the so-called Shapley kernel weights: $k(\mathcal{K}, \mathcal{S}) = \frac{|\mathcal{K}|-1}{\binom{|\mathcal{K}|}{|\mathcal{S}|}|\mathcal{S}|(\mathcal{K}-|\mathcal{S}|)}$.³⁹ 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 $\hat{f}(\mathcal{K}) = \mathbb{E}[\hat{f}(x|x = x_\mathcal{S})]$ for all $\mathcal{S}$ and ordered in line with $\mathbf{Z}$, the setup in (A.13) can be rewritten as:

$$(\mathbf{f} - \mathbf{Z}\phi)^T \mathbf{W} (\mathbf{f} - \mathbf{Z}\phi), \quad (\text{A.14})$$

which leads to the standard WLS solution $\phi = (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{W} \mathbf{f}$. The benefit of this formulation is that the matrix multiplication $\mathbf{R} = (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{W}$ only has to be calculated once as it does not depend on $\mathcal{K}$. 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 (A.13). This is the first key innovation by [Lundberg and Lee \(2017\)](#) and makes Kernel SHAP a computationally feasible approximation to the exact Shapley values in (A.10).

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 $\hat{f}(\mathcal{S}) = \mathbb{E}[\hat{f}(x|x = x_\mathcal{S})]$, one approach is to integrate out all covariates that are not included in $\mathcal{S}$ from the expectation:

$$\mathbb{E}[\hat{f}(x)|x = x_\mathcal{S}] = \int \hat{f}(x_{\bar{\mathcal{S}}}, x_\mathcal{S}) p(x_{\bar{\mathcal{S}}}|x = x_\mathcal{S}) d\mathbf{x}_{\bar{\mathcal{S}}}, \quad (\text{A.15})$$

where $p(x_{\bar{\mathcal{S}}}|x = x_\mathcal{S})$ 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 those covariates which *are* included in $\mathcal{S}$: $x_\mathcal{S}$. The standard Kernel SHAP by [Lundberg and Lee \(2017\)](#) assumes independence amongst features and thus replaces the conditionality in the distribution and simply implements $p(x_{\bar{\mathcal{S}}})$, the empirical density of the variables. In case such independence assumptions are unlikely, various alternative solutions have been developed to incorporate the conditionality between $x_\mathcal{S}$ and $x_{\bar{\mathcal{S}}}$ ([Aas et al., 2021](#)).

³⁹Note that 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 $\phi_0 = \hat{f}(\emptyset)$ and $\sum_{k=0}^\mathcal{K} = \hat{f}(\mathcal{K})$ as additional constraints instead ([Lundberg and Lee, 2017](#), p. 6).

Tree SHAP An alternative estimation approach to $\hat{f}(\mathcal{S}) = \mathbb{E}[\hat{f}(x|x = x_{\mathcal{S}})]$ is available for tree-based methods, Tree SHAP, due to [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 $x_{\mathcal{S}}$ throughout the tree and is thus effectively capped through the number of partitions in the tree. The basic approach to estimate $\hat{f}(\mathcal{S}) = \mathbb{E}[f(x|x = x_{\mathcal{S}})]$ 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  $x_j \in x_{\mathcal{S}}$  then
    |   if  $x_j \leq s_j$  then
    |     | return  $G(j_L)$ 
    |   else
    |     | return  $G(j_R)$ 
    |   end
  | else
  |   return  $\frac{G(j_L) \cdot r_L + G(j_R) \cdot r_R}{r_L + r_R}$ 
  | end
end

```

This algorithm $G(\cdot)$ is initialized at the root node of the tree and first evaluates whether the node is a leaf — i.e., 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, it is evaluated 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 onwards uses a weighted average of $G(\cdot)$ applied to both nodes.

A key development in [Lundberg et al. \(2020\)](#) is a much more computationally efficient version of Algorithm 1 that elegantly stores information of paths that have already been followed within previous passes of the algorithm. The reader is referred to the

discussion in [Lundberg et al. \(2020\)](#) (p. 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 $\hat{f}(\mathcal{S}) = \mathbb{E}[\hat{f}(x|x = x_{\mathcal{S}})]$, as the approach is conditional on the tree Π to generate estimates of $\hat{f}(\mathcal{S})$ whilst 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 (A.12). SHAP interaction values calculate the difference between a Shapley value ϕ_k for some observation i when covariate j is present versus when it is absent. Specifically, it is estimated through:

$$\Phi_{j,k}(f, x) = \sum_{\mathcal{S} \subseteq \mathcal{K} \atop k, j} = \frac{|\mathcal{S}|!(|\mathcal{K}| - |\mathcal{S}| - 2)!}{2(|\mathcal{K}| - 1)!} \nabla_{j,k}(f, x, \mathcal{S}), \quad (\text{A.16})$$

for $j \neq k$ where

$$\nabla_{j,k}(f, x, \mathcal{S}) = f_x(\mathcal{S} \cup j, k) - f_x(\mathcal{S} \cup j) - f_x(\mathcal{S} \cup k) + f_x(\mathcal{S}). \quad (\text{A.17})$$

Here, $f_x(\cdot)$ is the standard formula to estimate a Shapley value as defined in (A.10). 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 $\sum_{j \neq k} \Phi_{j,k}(f, x)$. The intuition behind (A.17) 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 (A.17). From this, the Shapley values when *either* including k or j are subtracted — the second and third terms in (A.17). Finally, the Shapley value from including *neither* k and j is added — the last term in (A.17). Again, using the same efficient implementation as discussed above allows for computationally tractable estimation of these pairwise values ([Lundberg et al., 2020](#), p. 66).

References

- Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298, 2021.
- Andrew Abbott. Transcending general linear reality. *Sociological Theory*, pages 169–186, 1988.
- Mayank Agrawal, Joshua C Peterson, and Thomas L Griffiths. Scaling up psychology via scientific regret minimization. *Proceedings of the National Academy of Sciences*, 117(16):8825–8835, 2020.
- Susan Athey. The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda*, pages 507–547. University of Chicago Press, 2018.
- Valerio Baćak and Edward H Kennedy. Principled machine learning using the Super Learner: An application to predicting prison violence. *Sociological Methods & Research*, 48(3):698–721, 2019.
- Stephen Bates, Trevor Hastie, and Robert Tibshirani. Cross-validation: what does it estimate and how well does it do it? *arXiv*, 2021. doi: 2104.00673.
- Janina Beiser-McGrath and Liam F Beiser-McGrath. Problems with products? Control strategies for models with interaction and quadratic effects. *Political Science Research and Methods*, 8(4):707–730, 2020.
- Richard A Berk. *Regression analysis: A constructive critique*, volume 11. Sage, 2004.
- Richard A Berk and Justin Bleich. Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Pub. Pol’y*, 12(3):513–544, 2013.
- Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.
- Matthew Blackwell and Michael P Olson. Reducing model misspecification and bias in the estimation of interactions. *Political Analysis*, pages 1–20, 2021.
- Jennie E Brand, Jiahui Xu, Bernard Koch, and Pablo Geraldo. Uncovering sociological effect heterogeneity using tree-based machine learning. *Sociological Methodology*, 52(2):189–223, 2021.
- Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.
- Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.
- Mauricio Bucca and Daniela R Urbina. Lasso regularization for selection of log-linear models: An application to educational assortative mating. *Sociological Methods & Research*, 50(4):1763–1800, 2019.
- Andreas Buja, Lawrence Brown, Richard Berk, Edward George, Emil Pitkin, Mikhail

- Traskin, Kai Zhang, and Linda Zhao. Models as approximations i: Consequences illustrated with linear regression. *Statistical Science*, 34(4):523–544, 2019a.
- Andreas Buja, Arun Kumar Kuchibhotla, Richard Berk, Edward George, Eric Tchetgen Tchetgen, and Linda Zhao. Models as approximations—rejoinder. *Statistical Science*, 34(4):606–620, 2019b.
- A Colin Cameron and Pravin K Trivedi. *Microeconometrics: methods and applications*. Cambridge University Press, 2005.
- Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.
- Bin Chi, Adam Dennett, Thomas Oléron-Evans, and Robin Morphet. A new attribute-linked residential property price dataset for england and wales, 2011 to 2019. *UCL Open: Environment Preprint*, 2021.
- Björn Christensen and Sören Christensen. Are female hurricanes really deadlier than male hurricanes? *Proceedings of the National Academy of Sciences*, 111(34):E3497–E3498, 2014.
- JA Davis and TW Smith. *The NORC General Social Survey: A User Guide*. SAGE, 1991.
- Jurgen A Doornik and David F Hendry. Statistical model selection with “Big Data”. *Cogent Economics & Finance*, 3(1), 2015.
- Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv*, 2017. doi: 1702.08608.
- Michael R Dougherty, Rick P Thomas, Ryan P Brown, Jeffrey S Chrabaszcz, and Joe W Tidwell. An introduction to the general monotone model with application to two problematic data sets. *Sociological Methodology*, 45(1):223–271, 2015.
- Otis Dudley Duncan. *Notes on social measurement: Historical and critical*. Russell Sage Foundation, 1984.
- Bradley Efron and Trevor Hastie. *Computer age statistical inference*. Cambridge University Press, 2016.
- Benjamin Elbers. A method for studying differences in segregation across time and space. *Sociological Methods & Research*, page 0049124121986204, 2021.
- J Paul Elhorst. Applied spatial econometrics: Raising the bar. *Spatial Economic Analysis*, 5(1):9–28, 2010.
- Nicholas Epley and Thomas Gilovich. The mechanics of motivated reasoning. *Journal of Economic Perspectives*, 30(3):133–40, 2016.
- Gang-Zhi Fan, Seow Eng Ong, and Hian Chye Koh. Determinants of house price: A decision tree approach. *Urban Studies*, 43(12):2301–2315, 2006.

- Michael Freeden, Lyman Tower Sargent, and Marc Stears. *The Oxford handbook of political ideologies*. Oxford University Press, 2013.
- David A Freedman. *Statistical models: theory and practice*. Cambridge University Press, 2009.
- Yoav Freund and Robert E Schapire. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156, 1996.
- Jerome H Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 4(3):1–67, 1991.
- Andrew Gelman and Eric Loken. 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*, 348:1–17, 2013.
- Alan S Gerber and Neil Malhotra. Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods & Research*, 37(1):3–30, 2008.
- Richard M Golden, Steven S Henley, Halbert White, and T Michael Kashner. Generalized information matrix tests for detecting model misspecification. *Econometrics*, 4(4):46, 2016.
- Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a “right to explanation”. *AI magazine*, 38(3):50–57, 2017.
- Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24: 395–419, 2021.
- Trevor Hastie and Robert Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009.
- James J Heckman, John Eric Humphries, and Gregory Veramendi. Returns to education: The causal effects of education on earnings, health, and smoking. *Journal of Political Economy*, 126(S1):S197–S246, 2018.
- Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in Neural Information Processing Systems*, 33:4778–4789, 2020.
- Matthew Hindman. Building better models: Prediction, replication, and machine learning in the social sciences. *The ANNALS of the American Academy of Political and Social Science*, 659(1):48–62, 2015.
- Jake M Hofman, Amit Sharma, and Duncan J Watts. Prediction and explanation in social systems. *Science*, 355(6324):486–488, 2017.

- Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.
- John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.
- Lucas Janson, William Fithian, and Trevor J Hastie. Effective degrees of freedom: A flawed metaphor. *Biometrika*, 102(2):479–485, 2015.
- Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. The theory is predictive, but is it complete? An application to human perception of randomness. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 125–126, 2017.
- Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.
- Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner’s perspective. *arXiv*, 2022. doi: 2202.01602.
- Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1–8, 2019.
- Thomas Lemieux. The “Mincer equation” thirty years after schooling, experience, and earnings. In *Jacob Mincer a pioneer of modern labor economics*, pages 127–145. Springer, 2006.
- Yun Liao et al. Machine learning in macro-economic series forecasting. *International Journal of Economics and Finance*, 9(12):71–76, 2017.
- Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.
- J Scott Long and Pravin K Trivedi. Some specification tests for the linear regression model. *Sociological Methods & Research*, 21(2):161–204, 1992.
- Ian Lundberg, Rebecca Johnson, and Brandon M Stewart. What is your estimand? defining the target quantity connects statistical evidence to theory. *American Sociological Review*, 86(3):532–565, 2021.
- Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.
- Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.
- Stephen Malpezzi. Hedonic pricing models: a selective and applied review. *Housing Economics and Public Policy*, 1:67–89, 2003.

- Elizabeth Aura McClintock. Occupational sex composition and gendered housework performance: Compensation or conventionality? *Journal of Marriage and Family*, 79(2):475–510, 2017.
- Gerhard Meisenberg. Verbal ability as a predictor of political preferences in the United States, 1974–2012. *Intelligence*, 50:135–143, 2015.
- Mario Molina and Filiz Garip. Machine learning for sociology. *Annual Review of Sociology*, 2019.
- Stephen L Morgan and Jiwon Lee. Social class and party identification during the clinton, bush, and obama presidencies. *Sociological Science*, 4:394–423, 2017.
- Sendhil Mullainathan and Jann Spiess. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.
- John Muñoz and Cristobal Young. We ran 9 billion regressions: Eliminating false positives through computational model robustness. *Sociological Methodology*, 48(1):1–33, 2018.
- Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.
- Robert M O’Brien. Comment: Some challenges when estimating the impact of model uncertainty on coefficient instability. *Sociological Methodology*, 48(1):34–39, 2018.
- Judea Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, 2009.
- Joshua C Peterson, David D Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L Griffiths. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214, 2021.
- Alexander Peysakhovich and Jeffrey Naecker. Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. *Journal of Economic Behavior & Organization*, 133:373–384, 2017.
- Eric Polley, Erin LeDell, Chris Kennedy, Sam Lendle, and Mark van der Laan. Package ‘superlearner’, 2021.
- Eric C Polley, Sherri Rose, and Mark J Van der Laan. Super learning. In *Targeted learning*, pages 43–66. Springer, 2011.
- Charles Rahal, Mark D Verhagen, and David S Kirk. The rise of machine learning in the academic social sciences. *AI and Society*, 2022. doi: 10.1007/s00146-022-01540-w.
- James Bernard Ramsey. Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 31(2):350–371, 1969.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why should I trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd*

- ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.
- Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 56(4):931–954, 1988a.
- Peter M Robinson. Semiparametric econometrics: A survey. *Journal of Applied Econometrics*, 3(1):35–51, 1988b.
- Sherri Rose. Mortality risk score prediction in an elderly population using machine learning. *American Journal of Epidemiology*, 177(5):443–452, 2013.
- Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.
- Cynthia Rudin, Rebecca J Passonneau, Axinia Radeva, Haimonti Dutta, Steve Ierome, and Delfina Isaac. A process for predicting manhole events in manhattan. *Machine Learning*, 80(1):1–31, 2010.
- Xavier X Sala-i Martin. I just ran four million regressions. Technical report, National Bureau of Economic Research, 1997.
- Matthew J Salganik, Ian Lundberg, Alexander T Kindel, Caitlin E Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M Altschul, Jennie E Brand, Nicole Bohme Carnegie, Ryan James Compton, et al. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403, 2020.
- Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*. Springer Nature, 2019.
- Lloyd S Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10):1095–1100, 1953.
- Galit Shmueli. To explain or to predict? *Statistical science*, 25(3):289–310, 2010.
- Uri Simonsohn, Joseph P Simmons, and Leif D Nelson. Specification curve analysis. *Nature Human Behaviour*, 4(11):1208–1214, 2020.
- Adam Slez. The difference between instability and uncertainty: Comment on Young and Holsteen (2017). *Sociological Methods & Research*, 48(2):400–430, 2019.
- Mervyn Stone. 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, 1977.
- VS Subrahmanian and Srijan Kumar. Predicting human behavior: The next frontiers. *Science*, 355(6324):489–489, 2017.
- Siyi Tang, Amirata Ghorbani, Rikiya Yamashita, Sameer Rehman, Jared A Dunnmon, James Zou, and Daniel L Rubin. Data valuation for medical imaging using

- shapley value and application to a large-scale chest x-ray dataset. *Scientific reports*, 11(1):1–9, 2021.
- Jan van der Laan, Edwin de Jonge, Marjolijn Das, Saskia Te Riele, and Tom Emery. A whole population network and its application for the social sciences. *European Sociological Review*, 2022.
- Mark J Van Der Laan and Sandrine Dudoit. 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*, 2003.
- Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.
- Aad W Van der Vaart, Sandrine Dudoit, and Mark J van der Laan. Oracle inequalities for multi-fold cross validation. *Statistics & Decisions*, 24(3):351–371, 2006.
- Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27(5): 1413–1432, 2017.
- Mark D Verhagen. A pragmatist’s guide to using prediction in the social sciences. *Socius*, 8:23780231221081702, 2022.
- Duncan J Watts. Common sense and sociological explanations. *American Journal of Sociology*, 120(2):313–351, 2014.
- Duncan J Watts. Should social science be more solution-oriented? *Nature Human Behaviour*, 1(1):1–5, 2017.
- Halbert White. Using least squares to approximate unknown regression functions. *International Economic Review*, 21(1):149–170, 1980.
- Halbert White. Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association*, 76(374):419–433, 1981.
- Adonis Yatchew. An elementary estimator of the partial linear model. *Economics Letters*, 57(2):135–143, 1997.
- Adonis Yatchew. *Semiparametric regression for the applied econometrician*. Cambridge University Press, 2003.
- Cristobal Young. Model uncertainty and the crisis in science. *Socius*, 4:1–7, 2018.
- Cristobal Young. The difference between causal analysis and predictive models: Response to “comment on Young and Holsteen (2017)”. *Sociological Methods & Research*, 48(2):431–447, 2019.
- Cristobal Young and Katherine Holsteen. Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research*, 46(1):3–40, 2017.

- Cristobal Young and Sheridan A Stewart. Functional form robustness: Advancements in multiverse analysis. *Unpublished*, 2021.
- H Peyton Young. Monotonic solutions of cooperative games. *International Journal of Game Theory*, 14(2):65–72, 1985.
- Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.