A Pragmatist's Guide to Using Prediction in the Social Sciences

Mark Verhagen*

January 17, 2022

Abstract

Prediction is an underutilized tool in the social sciences, often for the wrong reasons. Many social scientists confuse prediction with unnecessarily complicated methods or with narrowly predicting the future. This is unfortunate. When we view prediction as the simple process of evaluating a model's ability to approximate an outcome of interest, it becomes a more generally applicable and disarmingly simple technique. For all its simplicity, the value of prediction should not be underestimated. Prediction can address enduring sources of criticism plaguing the social sciences, like a lack of assessing a model's ability to reflect the real world, or the use of overly simplistic models to capture social life. I illustrate these benefits with empirical examples that merely skim the surface of the many and varied ways in which prediction can be applied, staking the claim that prediction is a truly illustrious 'free lunch' that can greatly benefit empirical social scientists.

*All code and publicly available data underlying the analyses in this paper can be found at `https://github.com/MarkDVerhagen/Pragmatist_Guide_to_Prediction`.

Introduction

Social scientists should start using prediction more often. Prediction is the process of generating predicted values of a dependent variable by applying an estimated model to a set of explanatory variables. It brings a unique analytical perspective to empirical work. Prediction can also help address enduring sources of criticism facing the social sciences. Examples are a general lack of assessing research findings in terms of their real-world relevance, and the use of overly simplistic models to study the complexities of social life. In this article, I address common misconceptions about prediction and provide a simple definition that addresses existing barriers to adoption. I then discuss and illustrate some of the many benefits that prediction can bring when used as a complement to traditional empirical methods. I argue that prediction can and should become a fundamental part of the social scientist's empirical toolkit, but that this first requires us to look beyond the current dichotomy between prediction and explanation and instead view the two as complementary to one another.

The current lack of prediction in the social sciences stems from a seeming incompatibility between wanting to explain and wanting to predict, effectively forcing the researcher to choose between the two approaches. A case in point is the much-cited paper by Galit Shmueli – aptly titled 'To Predict or to Explain' – which outlines how a social scientist's empirical workflow differs in terms of data processing, modeling, and post-estimation diagnostics when choosing to either predict or explain [1]. Naturally, the paper assumes that a researcher would not normally attempt to do both. This is an accurate reflection of social science research. The apparent need to dogmatically choose between either approach means that, in practice, social scientists tend to stick to explanation almost exclusively. Illustratively, the terms "predict" and "prediction" are mentioned in less than 5% of abstracts over the last ten years in various flagship journals in economics, political science and sociology, and of the papers mentioning either term, only 13% proceed to generate actual predictions of the outcome variable (Table 1).[1,2]

So why does this dichotomy exist? In many cases, an unnecessarily narrow interpretation of prediction is to blame. For example, the type of prediction discussed by Shmueli refers to the practice of maximizing predictive power, which is a subset of the

[1] In most of the articles that mention the term 'predict' or 'prediction', the authors use the commonplace, conceptual meaning term – e.g. 'we predict that' or 'our theory makes several predictions'. The actual process of making predictions of the outcome variable is virtually non-existent in the literature cited. Note that the term "explain" or "explanation" only features in about 13% of abstracts, although this proportion likely does not reflect the proportion of work that is explanatory. Explanation is the default approach to empirical work, making it less relevant to explicitly mention the term in the abstract.

[2] The single sociological paper making predictions in fact generated mortality forecasts [2].

Table 1. Number of articles mentioning `predict` or `prediction` in the abstract, and actual usage of prediction from six flagship journal in Economics, Political Science, and Sociology, 2010 to 2021.

Journal	Total articles[†]	Mentions prediction	Actually makes predictions[*]
American Economic Review	2414	85	12
Quarterly Journal of Economics	458	47	3
American Journal of Political Science	800	61	14
American Political Science Review	743	33	4
American Journal of Sociology	394	8	0
American Sociological Review	523	24	1

[†]Data was collected using the Scopus API using the `predict` and `prediction` search queries.
[*]Understood as generating predictions of the outcome variable (including forecasting). Papers generating predicted probabilities by setting explanatory variables to their mean or median were excluded as such predictions don't reflect actual observation in the data.

more general practice of making predictions. As a result, prediction is often conflated with the use of complex non-linear models like those from the domain of machine learning, which have their own unique set of challenges [3]. There is no reason to transfer these challenges to the general process of making predictions, which can be done with any type of model. Prediction tends to be narrowly understood as forecasting, which is but one of numerous examples of making predictions [4].[3] The biggest culprit, however, is the enduring discussion whether prediction and explanation are conceptually the same, and whether the latter should imply the former. This philosophical debate, although interesting, is ultimately irrelevant to applying prediction in explanatory research. When viewing prediction as a simple tool to evaluate a model's ability to approximate the outcome of interest, it can be applied without exception to most social science questions, rendering a dogmatic choice between prediction or explanation unnecessary.

The more relevant question is what prediction might bring to the table. To illustrate just one dimension, take the Fragile Families challenge (FFC), a case which I will

[3]Prediction is more commonly encountered in those social science domains that put emphasis on forecasting and projecting. Typical examples are the analysis of (financial) time-series, but can also include the prediction of conflicts or rare events [5], network science [6], and demography [7].

return to throughout this paper [8]. During the FFC, 160 research teams around the globe were asked to predict a number of important early-stage life outcomes of general interest to social scientists (e.g., eviction and material hardship). The idea was to evaluate the general predictability of these outcomes through a common task setup [9]. This setup mirrored the popular competition website Kaggle, where datasets with some outcome of interest and a number of possible explanatory variables are published online. Participants are challenged to estimate models that can accurately predict the outcome. These models are then tested on a partition of the data which is kept secret. Similar to Kaggle, the FFC made available a rich dataset to generate predictive models, while storing an evaluation set against which each team's predictive model was scored. The organizers encouraged the use of prediction-focused algorithms, rather than the explanatory methods already applied in hundreds of peer-reviewed articles using the same data.

The conclusions were telling for a number of reasons. First, many teams applied methods using flexible functional forms and variable selection techniques not often seen in the social sciences. Second, most models were nonetheless poorly able to predict life outcomes, although some did improve on benchmark models including a curated number of explanatory variables in a standard linear model. Poor overall predictability was thus a feature of both predictive and explanatory techniques. Third, and most important, the FFC was a rare occasion where the onus was truly on prediction rather than explanation. As a consequence, it put into sharp focus the fact that decades of explanatory research into the outcomes of interest had not led to much predictive ability. This somewhat awkward finding led the organizers to conclude that researchers had to 'find a way to reconcile a widespread belief that understanding has been generated by these data ... with the fact that the very same data could not yield accurate predictions of these important outcomes' [10] (p. 8402).

If prediction had been a more natural tool for social scientists, the main take-away of the FFC would likely not have taken so long to materialize. An earlier realization of the predictive limits of our knowledge might have stimulated a rigorous evaluation of the mechanisms hypothesized, the methods employed, and/or data collected at an earlier stage in the dataset's rich academic career and throughout life course research more generally. Importantly, it is unlikely that poor predictability is only a feature of life course research. Assessing the ability of our research findings to meaningfully predict outcomes we are interested in will most likely spur important debate in many other domains as well. The point is that our traditional preference for in-sample diagnostics means that we often don't assess our models ability to approximate the outcomes we care about. Prediction can, amongst other things, solve this problem.

The FFC is but a single example how prediction can shine a different light on empirical work and represents one of many approaches to making and evaluating predictions. More generally, this paper argues that prediction can bring the following three key virtues to the table of the social scientist:

1. Prediction provides improved insight into model fit.
2. Prediction provides a benchmarking tool across modeling domains.
3. Prediction can help generate insight into the behavior of complicated models.

These key virtues come in addition to other benefits. Some examples are an improved alignment of research findings and policy [3], providing a metric to align scientific efforts [11, 12], and improving transparency and the ability to scrutinize estimated models [13]. Viewing prediction as a complement to classical methods would also ease the incorporation of prediction-focused methods from machine learning into the social sciences [14].

To summarize, with this paper I aim to increase the use of prediction in explanatory research by challenging the unnecessary dichotomy between prediction and explanation, and illustrating the many benefits prediction brings when applied as a complement to explanatory analysis. Hopefully, this paper can serve as a pragmatic guide to the varied ways in which prediction can be successfully applied in the social sciences. The remainder of this paper is structured as follows. First, I will discuss several reasons why prediction is currently being underutilized, and provide a definition of prediction which should address these obstacles to adoption. I then provide a number of ways in which prediction can be operationalized, dependent on the case at hand. To showcase the benefits of prediction, I will then present three sets of empirical examples – in line with the three virtues outlined above – to illustrate the application of prediction. I conclude the paper with a summary and discussion of the main claims and findings.

A new perspective on prediction for the Social Sciences

The social sciences are currently dominated by a focus on explanation. This often boils down to estimating models reflecting some explanatory mechanism and assessing the in-sample coefficient estimates of these models. Prediction – which broadly reflects an interest in how well the models we estimate are able to approximate the dependent variable – plays, at best, an auxiliary role.[4] Below, I identify three reasons why

[4]Whenever prediction is applied, it is usually in the form of an auxiliary regression, e.g. Heckman selection methods, 2SLS or Matching methods. These predictions should not be considered as pure predictions given that they are meant to support standard in-sample evaluation methods and the

a predictive focus in the social sciences is lacking. Then, I will provide a simple definition of prediction which should not suffer from such barriers to adoption.

Before doing so, it is appropriate to briefly reflect on the intriguing philosophical debate whether explanation and prediction are conceptually the same. Some authors have forcefully claimed that causal explanation should always have predictive implications [4, 15, 16] whereas others (equally strongly) qualify this viewpoint [17]. This paper does not seek to wade into this debate for two reasons. First, because the debate has been documented extensively elsewhere [4, 18, 19]. Second, and more importantly, because the formal (in)equality between prediction and explanation is not strictly required to apply prediction for explanatory purposes. Therefore, I do not aim to support or assume either view going forward and encourage others to take a similarly pragmatic approach when considering to use prediction in their work. In that respect, none of the examples I use in this article requires a strong position on the above.

Prediction is often misperceived as deterministic forecasting Prediction is underutilized in the social sciences in part due to misperceptions of what prediction actually is. Prediction is often understood to deal with predicting outcomes or events in the future – i.e., outside of the time frame on which we have current data – and to be intrinsically deterministic – i.e., as making statements with certainty. This type of prediction is at best a small subset of the general process of making predictions.[5]

Predictions need not be made on future events, nor does prediction have to exclusively concern time-varying data. A mechanistic theory describing the effect of some variable X on an outcome y via some model $y = f(X)$ can lead to predictions in future, current, and past cases as long as the data used for prediction is similar to that used in estimating the model. For example, predictions can be made for a small partition of the dataset collected to study some mechanism, which is set aside and not used for model estimation but purely for predictive evaluation. This is the typical approach to prediction observed in the field of machine learning [9]. Tellingly, most machine learning applications do not concern time-varying events at all [20]. The only thing conceptually required to predict is a set of data similar to that used in estimation.

Accordingly, predictions are made using estimated models and should thus be viewed

predictions typically are not assessed substantively.

[5]Similar points have been made within the prediction versus explanation debate in Sociology [4]. In this particular work, the author implies another possible reason why prediction might be underutilized by explanatory researchers, noting that 'explanations will also become less satisfying' when forced to be predictive (p. 313). In other words, prediction might be actively avoided by researchers as it restricts the types of explanations one can plausibly argue for.

from a probabilistic perspective, just like classic techniques like (logistic) regression are inherently probabilistic in nature, too. What makes prediction different is an explicit focus on the outcome variable. There is no reason to assume determinism any more when making predictions using some model, than determinism is involved when evaluating the estimated coefficients of that very model.

Historical limitations limit the use of prediction in the present Historically, there were considerable limitations on both data and computational resources available to researchers. This still affects the use of prediction in the present. A parallel can be drawn to the enduring imbalance between Bayesian and Frequentist approaches to inference in the social sciences. Bayesian statistics require a relative abundance of computational resources, compared to a Frequentist approach. This made the use of fairly simplistic linear models – plugged into exponential family probability distributions with computationally convenient properties – the preferred methodological approach for social scientists during the latter half of the 20th century [21]. This dominance persists up to this very day. Choices which were reasonable and necessary at the time have led to an analytical mono-culture today [22].

Making predictions is similarly expensive: in some cases a part of the dataset has to be put aside for evaluation or models have to be estimated many times for robust inferences into the predictive performance of a model. Limits on data and computational resources have thus strengthened a (historical) preference for in-sample inferential methods [21]. In a day and age of ever larger datasets and computational power, however, these issues are a problem of the past. Just as the increases in both data and computational resources have led to a burgeoning growth of methods using Bayesian approaches, the use of prediction should no longer be held back by practical concerns. Even in small N settings, techniques have been developed that still allow the prediction to be applied.[6]

Prediction is conflated with the use of convoluted models More recently, prediction is approached with hesitance due to the astronomic rise of techniques from the domain of machine learning which place a strong emphasis on prediction [1, 11]. This has led to the risk that the limitations of machine learning methods are blindly transferred to prediction in general. To illustrate, reviews discussing the potential of machine learning for the social sciences have appeared in various important journals

[6]Increases in data size are a key feature of the past decade, although some of the larger datasets available to social scientists needn't be on par in terms of data quality [23]. Small N settings are not restrictive, as Leave-One-Out prediction – discussed later – still allows a predictive perspective to be pursued in such cases.

[11, 24–26]. All these reviews discuss the benefits of machine learning – e.g., increased model complexity and the lack of reliance on pre-specified functional forms – as well as the key difference: machine learning's focus on prediction.

Machine learning methods have various limitations and risks associated with them, most notably highly convoluted models with a profound lack of interpretability [27].[7] These risks have little to do with the general process of making predictions. Decoupling prediction from black-box methods is crucial to break the misperception that predictive accuracy is something which is naturally maximized at the cost of interpretability. Predictions can be made as easily using an additive linear model as with the complicated non-linear algorithms commonly applied in machine learning. That researchers within machine learning almost exclusively predict doesn't mean that prediction is exclusive to machine learning.

A Simple Definition of Prediction

Prediction understood as the process of *evaluating a model in terms of its ability to accurately approximate the outcome* should not suffer from the definitional confusion outlined above. Prediction simply calls for a renewed emphasis on our model's ability to model the dependent variable in our data. Based on this definition, making predictions consists of the following simple steps:

1. Define an `estimation set` to fit the model, and an `evaluation` set to generate predictions for;
2. Estimate the `model` using the `estimation set`;
3. Make `predictions` of the outcome using the `model` and the data in the `evaluation set`;
4. Evaluate the performance of the `predictions` against the `observed outcome`.

Clearly, the above subsumes the more narrow definitions of prediction like forecasting, or the use of machine learning, which fall within the confines of this broader definition.

To make the above more concrete, assume we estimate some functional form $f_\mu(\cdot)$ in order to find evidence for the association of years of education, A, on wages, y – an example I will return to later. We include work experience, B, as a control variable leading to the model $y = f_\mu(y, A, B)$. Typically, $f_\mu(\cdot)$ is a linear additive model plugged into an exponential family probability distribution with parameter-vector μ,

[7] Note that considerable developments in the field of 'Explainable A.I.' are advancing the interpretability of complex model spaces, and can be used to inform functional form development in typical exponential family models as well [28–30].

although more complicated algorithms can be applied without loss of generality. Prediction is as simple as estimating $\hat{f}_\mu(\cdot)$ using information on $y^{\text{estimation}}$, $A^{\text{estimation}}$, and $B^{\text{estimation}}$ from some dataset $\mathcal{D}^{\text{estimation}}$ and generating predictions using information on $A^{\text{evaluation}}$, and $B^{\text{evaluation}}$ from some dataset $\mathcal{D}^{\text{evaluation}}$:

$$\hat{y}^{\text{evaluation}} = \hat{f}_\mu(A^{\text{evaluation}}, B^{\text{evaluation}}). \tag{1}$$

The predictions $\hat{y}^{\text{evaluation}}$ can then be evaluated, for example by comparing them against the actually observed $y^{\text{evaluation}}$. There are many summary metrics of fit available for this purpose – e.g., the Root-Mean-Squared-Error or F1-score – but one can also compare (sets of) individual predictions against observed outcomes. Of interest is the broad ability of $\hat{f}_\mu(\cdot)$ to accurately model the outcome.

Based on this definition, the only decision a researcher has to make is how to define the set used for estimating the model, and how to define the set used to evaluate its predictive performance. I identify three general approaches which I discuss below.

In-sample evaluation A first option is to simply use the same data used for estimation to make predictions (Figure 1, Panel A). This choice would effectively lead to an in-sample assessment of model fit, and the well-known R^2 is an example of aggregate fit under this choice of evaluation set. In-sample prediction is sometimes also applied by researchers interpreting coefficients in non-linear models where the coefficient estimates lack straightforward interpretation, like categorical outcome models ([31, 32]).[8]

In effect, in-sample evaluation boils down to assessing the fitted values of the model estimated in step 2, above. For parametric models with sufficient sample sizes, this is an efficient approach as it uses all the available data for both estimation and evaluation. The downside is the risk that in-sample predictions can be overfit – leading some to argue that predictions should be made exclusively out-of-sample [12]. Overfitting is the reason why aggregated in-sample fit metrics like the Adjusted R^2 or information criteria are scaled downwards based on the degrees of freedom in a model.[9] When evaluating predictions at a lower level of aggregation – e.g., for a subset

[8] When researchers use prediction in the context of categorical outcome variables it is more common to perform *simulated* prediction, where many covariates are set to their mean or median values. This approach does not actually reflect the model's ability to approximate the outcome, as the data used need not be representative of the true population.

[9] In the case of parametrized models without shrinkage terms, the correction term of the unexplained residuals is $\frac{N-1-p}{N-1}$ where p is the number of degrees of freedom in the model. This correction term converges to one quickly for moderately-sized datasets. Therefore practically, in-sample predictions might suffice for moderate N and small p. Note, however, that this correction is meant to be applied to the aggregated fit metric and not individual predictions.

of the data – or when models become more parametrized – e.g., multilevel models or when using regularization techniques – using separate estimation and evaluation sets is strongly advised.

Cross Validated evaluation A second approach is to partition the existing dataset into disjunct estimation and evaluation sets. This is the typical approach often observed in machine learning and has the added benefit that any risk of overfitting the data is explicitly addressed. Predictions are only ever made for data which was not used to estimate the model. The most common such approach is K-fold cross validation which consists of dividing the dataset into K equal-sized 'folds' – typically, K is set to 5 or 10. The model is then estimated K times, each time omitting one of the folds from the estimation process and using the omitted fold as the evaluation set (Figure 1, Panel B). This ensures that predictions are generated for every observation in the dataset, thus maximizing the number of predictions given the available data. It does make the routine computationally more expensive as the model has to be fit K times. K-fold cross validation also means that estimation is only ever done on $n - n_k$ data points per run which can lead to a loss in efficiency and precision of the estimates.

There are various alternatives to implementing cross validation. For example, in low N situations or when efficiency in estimation is paramount, one could use Leave-One-Out (LOO) cross validation which is the special case of K-fold cross validation where K is set to N [34]. As the number of folds K approaches N, the loss in efficiency decreases although the computational cost of performing the cross validation increases. For relatively straightforward estimation like OLS, the overall increase in computational time is negligible, but it can become prohibitive if the time required to perform a single estimation of the model is already considerable. Beyond varying the number of folds K, additional robustness to random variation in splitting the data into folds can be incorporated by repeating the entire routine multiple times.[10] Overall, K-fold cross validation remains the most commonly applied approach.

External evaluation A third and final choice for the evaluation set can be a set of data which is completely 'unseen' by the researcher (Figure 1, Panel C). For example,

[10]Common examples include repeating the cross validation routine M times – so-called Repeated cross validation. Another variant is Monte Carlo cross validation, where again M runs of cross validation are done, but each run only uses a single split of the data into estimation and evaluation sets. Many of these approaches tend to converge to the same results in the limit, see [33] for a review. For most social science applications, the number of Monte Carlo simulations M can be relatively low as the evaluation set is typically about 20%-30% in order to accurately reflect the original data. Therefore, with M around 100 the impact of assigning data to folds should be approximated well.

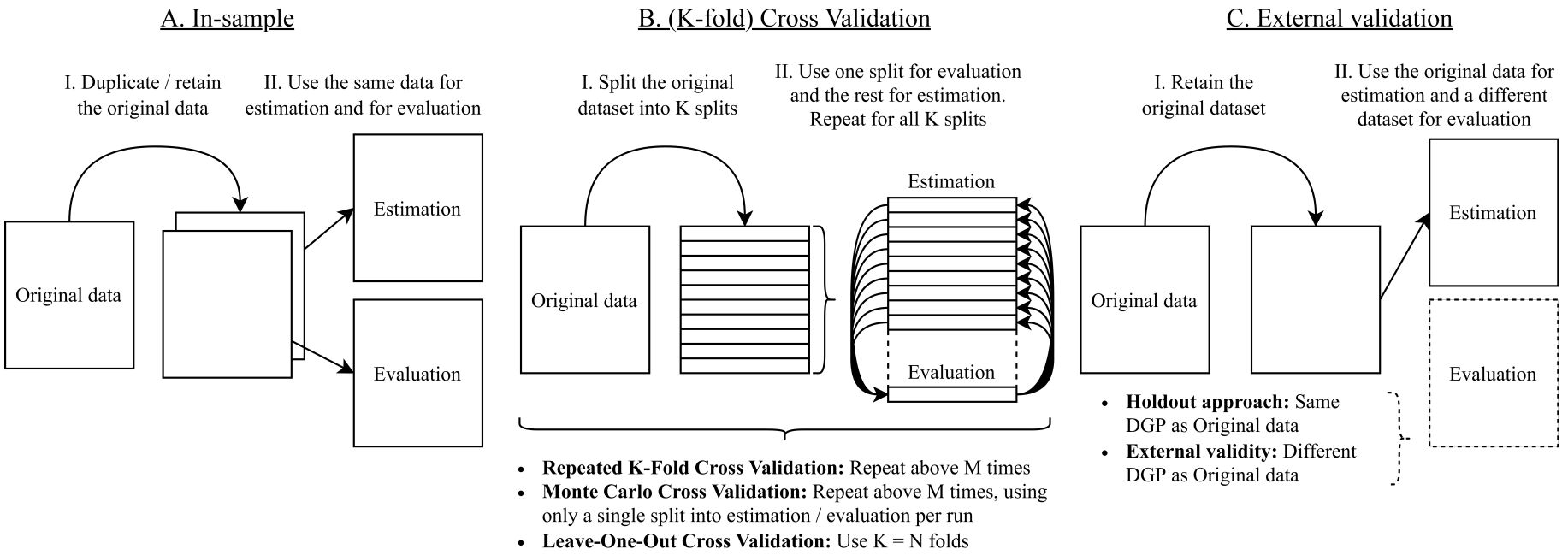

Figure 1. Three different strategies to define an estimation set and an evaluation set. The first strategy (A) uses the original data as both the estimation set as well as the evaluation set. The second strategy (B) splits the original data into K splits. Each split is used once as an evaluation set, to evaluate the model estimated to the remaining $K-1$ splits. The model is thus estimated K times. This step can be repeated M times, leading to repeated K-fold cross validation, or Monte Carlo cross validation in case only a single estimation / evaluation cycle is done per run instead of K [33]. The third strategy (C) uses the original data to estimate the model, and uses a different dataset to evaluate the model, which can be a holdout set from the same Data Generating Proces (DGP) but partitioned off prior to analysis, or collected separately.

by immediately partitioning off a part of the data into a holdout set which is kept separate from the entire estimation process or, ideally, never even shared with the researcher(s) – the typical approach in Kaggle-style competitions. This is called the 'Holdout' approach and provides the most truthful assessment of a model's predictive performance. Unfortunately, it is expensive as a sufficient number of observations are required to make a sufficient number of predictions and these observations cannot be used for estimation. Thus leading to both reduced efficiency in estimation, and a reduced number of predictions to evaluate.

External validation can also be done by assessing model predictions on a completely new set of collected data. For example, similar data that was collected at a different time – e.g., separate waves of a survey – or place – e.g., regional comparisons. The choice of an external evaluation set speaks directly to calls for increased attention to the external validity of research findings in empirical work [35, 36], and can be particularly useful to assess the transferability of research findings outside of the sample used for estimation. By choosing an external validation set, prediction provides a simple framework to assess model fit outside of the sample at hand.

In practice, the choice of splitting the data into estimation and evaluation sets will be made on a case-by-case basis. If a parametrized model is estimated with few coefficients and a considerable data size – e.g., $N > 500$ – the risk of overfitting will generally be low and in-sample prediction could be considered. When the number of parameters in a model increases, it is advisable to use some form of cross validation, either Leave-One-Out in case the number of observations is limited, or K-Fold with K typically about 10 [37]). If N is sufficiently large that setting aside a portion of the data does not meaningfully affect model estimation, the holdout approach can be applied where 20%-30% of the data is typically partitioned off as the holdout set.[11]

When using prediction to improve model understanding – as most of the examples in this paper do – Leave-One-Out cross validation is generally attractive if computationally feasible. Prediction is out-of-sample and the maximal number of data points $(n-1)$ are used to estimate the models used for each prediction. However, when the goal is to select an optimal predictive model to deploy amongst a set of candidate models, one might be interested in the expected prediction error on a completely new observation and the variability of this prediction error. In this case, K-Fold cross validation is typically a more efficient estimator, although some properties of the es-

[11]Fundamentally, the holdout set needs to be large enough to capture the intricacies of the original data well. Therefore, for low dimensional data with limited variation a small holdout set might already suffice. Conversely, high dimensional data or clustered data might require considerably more observation in the holdout to accurately reflect the data of interest. The same rationale holds when selecting the size of the single evaluation fold in Monte Carlo cross validation.

timand are not yet completely understood [21].[12] The Holdout or External validation approach will give the most precise assessment of a model's ability to accurately predict the outcome of new observations, although it is clearly expensive as one has to completely set aside a part of the data for evaluation or collect a new dataset. These questions are less relevant when using predictions to improve our understanding of explanatory models, as is the focus in this paper.

The Three Virtues of Prediction

As outlined in the introduction, complementing traditional methods with prediction brings three key virtues to the social scientist's table. In what follows, I illustrate these in kind using examples of prior work and novel reproductions.[13]

Virtue I: prediction provides improved insight into model fit

In its most simple form, prediction provides a distinct way to assess model fit on the level of the actual outcome variable. Such a perspective provides a renewed focus on what one could call predictive consciousness: an understanding how well our models are actually able to fit the outcome variable of interest. In practice, a model's fit is often left undiscussed, leading to a broad lack of predictive consciousness in empirical work. For example, whether the models we estimate are able to accurately predict 0.1%, 1% or 50% of the variation in the outcome. Model fit, if it is discussed at all, is typically assessed at the aggregate level only.[14] We are often left guessing what elements in a model contribute most to its ability to predict well. Nor do we know whether a model is able to predict all of the data equally well, or just parts of it. A renewed appreciation for prediction will improve our assessment of model

[12] The general approach to estimating the predictive error of a model is to assess the error of all the N predictions made using cross validation against their observed values and reporting the mean and standard deviation. Interestingly, cross validation has been shown to consistently estimate the expected error of the model fit to a random dataset drawn from the same underlying distribution as the training set, and not the expected error of the estimated model. In addition, the approach can lead to overly narrow confidence intervals [38, 39]. The reason is that the errors are not independent, as each observation is used in both estimation and evaluation. This problem will be minor when the impact of omitting a specific fold from the estimation process on the estimated model is small – i.e., the model is stable across the omission of folds – but can have a serious impact otherwise. Some solutions have been suggested to correctly scale the CI's of predictive error for certain families of models, but research remains ongoing [38].

[13] All publicly available data and corresponding code underlying the reproductions are archived in a Zenodo repository accompanying this paper.

[14] Typically, the in-sample R^2 has been used for the purpose of evaluating explanatory power. The measure has various limitations in often encountered empirical setups, for instance when modeling ordered outcomes or estimating other non-linear models. In those cases, information criteria are typically reported although these tend to defy an intuitive interpretation of the models ability to fit the outcome.

fit, as predictions are made for every single observation in the evaluation set. As a consequence, a (disaggregated) assessment of fit on the scale of the outcome comes naturally.

Predictive performance is also intuitive to understand and will help the implementation of research findings in the real-world. Its intuitive nature promotes acceptance and understanding of research findings by both policymakers and the public. Academically, a general sense of predictive accuracy is equally important to further a research agenda: if predictive performance (strongly) underperforms expectations, it prompts reflection of whether we are actually missing important determinants or perhaps our preferred functional form is not able to capture the mechanisms in operation. Finally, making and evaluating predictions beyond the aggregate level can also provide additional transparency into the academic process. By reporting on model fit at lower levels of aggregation, consumers of empirical research become better able to critically assess what a model can and cannot do in terms of fitting the data.

As a first example of the above, consider the FFC which was introduced earlier [8]. The FFC challenged research teams to accurately predict life outcomes at age 15 based on a rich set of data using the Holdout approach – i.e., setting aside a partition of the data.[15] As part of the challenge, the organizers calculated a benchmark performance using models constructed by domain experts. As a consequence, the low predictive power of the models typically estimated in this domain already became quite obvious from the outset (Figure 2, panel A). For many of the life outcomes of interest measured at age 15, a model including the hand-picked variables and a lagged version of the variable at age 9 did not substantially improve predictive accuracy relative to a null model predicting the overall mean of the outcome. This is a sobering finding putting into perspective the supposedly large amount of understanding that had been generated regarding these outcomes [10].

Perhaps the outcomes of interest to the FFC are inherently noisy and difficult to predict, as the organizers also note [10, 11]. Regardless of the question why predictive ability was low within the FFC, the very insight and subsequent discussion it provoked are essential for the field to develop. The FFC illustrates how predictive consciousness can be crucial to instigate a critical reflection on the state of knowledge in a field and can spur important debate. Accordingly, the key take-away of the FFC is that the type of discussion which a predictive focus triggered should not have taken so long to materialize. Likely, a predictive consciousness is equally relevant for many other social science research fields. More consistent reporting of predictive accuracy would

[15]The Holdout approach was chosen due to the competitive nature of the challenge: organizers were interested in finding the best predictive model amongst the participants.

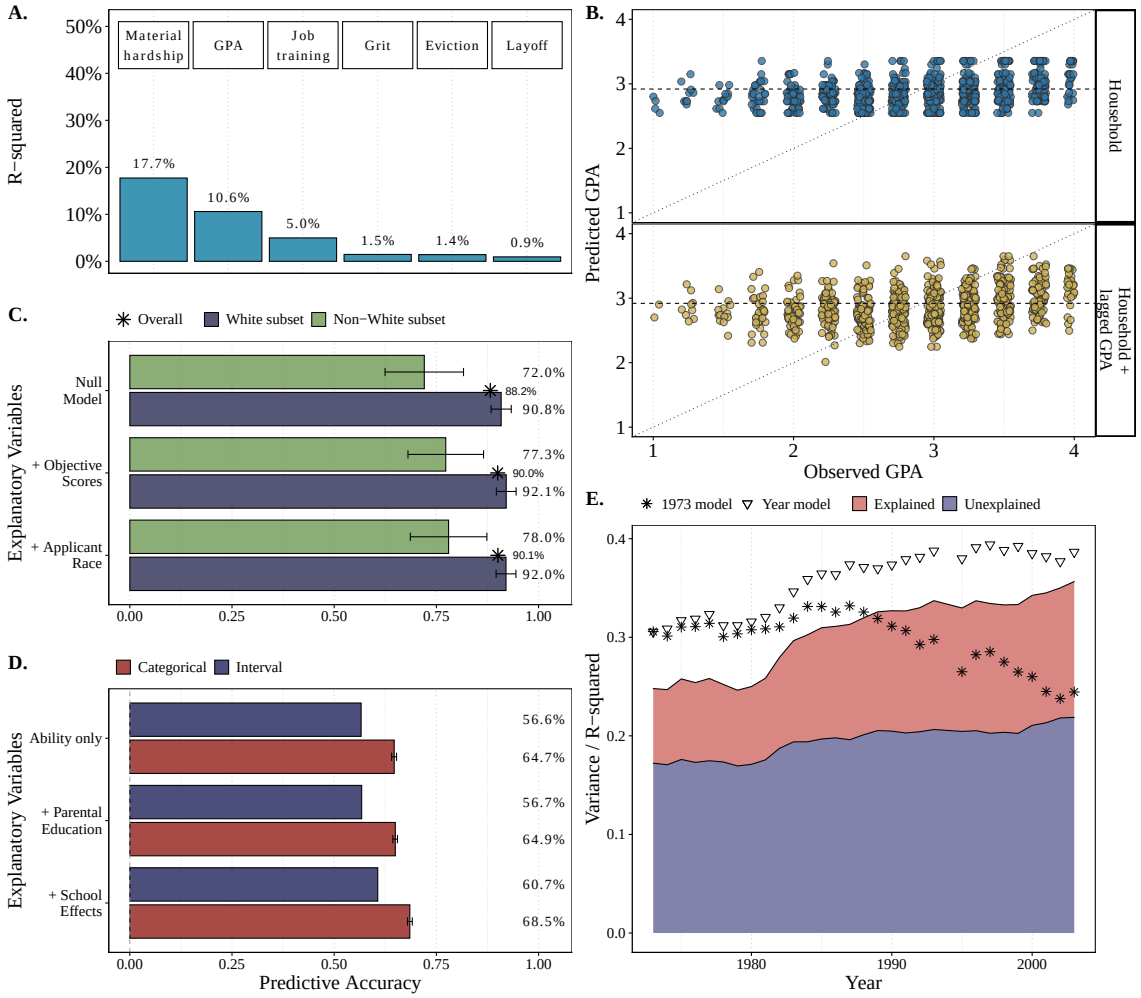

Figure 2. Panel A shows the predictive R^2 of the linear benchmarks chosen by domain experts relative to a null model for the FFC. Each model includes four explanatory variables and a lagged version of the outcome [10]. Panel B shows individual predictions of the GPA outcome in the FFC using only the domain expert variables (top) and when including the lagged version of the outcome (bottom). Panel C shows the performance of logistic regression models predicting whether a mortgage application was successful using various sets of explanatory variables. Performance is shown disaggregated for White (green) and Non-White (blue) applicants, showing that the performance is considerably lower for the latter. Panel D shows the performance of models predicting students' track levels using various sets of explanatory variables but different models. Performance of the categorical model is substantially higher than the linear model. The performance of both models strongly increases when including school variation. Panel E shows the performance of various models in explaining hourly wage in the US. Predictive power is assessed for each year using the same model but re-estimated to that year's data (triangles) or using the model estimated in the year 1973 (stars). Initially, the latter model performs well on the first couple of subsequent waves, but deteriorates from 1983 onwards. Confidence intervals, where present, reflect 95% confidence bounds of the estimated predictive accuracy across the various evaluation folds.

guarantee that the type of reflection resulting from the FFC becomes more likely to occur throughout the social sciences.

Predictions can also diagnose model performance at different levels of detail or aggregation. As an example, take the individual predictions rather than the overall predictive accuracy of the benchmark model for the GPA outcome in the FFC. We can visualize the predicted GPA against the observed GPA for each observation in the evaluation set and can do so for different models (Figure 2, panel B). Taking such a disaggregated approach to model fit, we observe that both models struggle to structure the outcome well, but this is especially so for students performing below average. A nuance which provides pointers for future avenues of research. In other words, prediction's ability to interrogate model fit on a disaggregated level provides a different vantage point than summary metrics of model fit.

In addition to individual predictions, they can also be assessed i) at the group level or ii) using completely different models and/or sets of explanatory variables. As an example, consider a reproduction of the influential 1996 study from the Federal Reserve Bank of Boston regarding discrimination in mortgage lending [40]. The authors – amongst other things – found evidence of discrimination against Non-Whites based on a logistic regression including a Race dummy and conditioning on various objective characteristics of the application. By complementing their analysis with a predictive perspective, additional nuances emerge (Figure 2, panel C).[16]

For example, most individuals are successful in their mortgage application and a null model already correctly predicts 88% of the data (by predicting successes for everyone). Including variables like objective score measures and household characteristics further increases the model's performance. However, aggregate fit is somewhat misleading, as there is a considerable gap in the model's ability to predict outcomes for Non-White applicants compared to White applicants (77% versus 92%). The inclusion of a Race indicator only marginally improves the gap. In other words, Non-Whites are modeled considerably less accurately than Whites. This could imply that additional sources of heterogeneity are present for the former – for example if bias is multimodal and depends on other factors like the employee reviewing the application – or some other reason is present why Non-Whites are modeled considerably less well.

Another illustration where prediction provides additional understanding of model fit is a recent study assessing teacher bias in educational tracking – the process of assigning students to ability levels – in the Netherlands. In the paper, prediction is explicitly

[16]Repeated K-Fold stratified cross validation was applied, ensuring similar proportions of Whites and Non-Whites in each fold, with $K = 5$ and $M = 100$. Predictive accuracy was thus estimated for a total of 500 folds.

applied to understand the relative importance of different sets of explanatory variables as well as modeling assumptions [41] (Figure 2, panel D).[17] This predictive perspective led to a number of important nuances to the existing knowledge on teacher bias. First, a predictive approach showed that commonly studied bias factors like parental education – although statistically significant – mattered little for the model's fit of the data, improving the predictive R^2 by a mere 0.1%. When allowing for separate intercepts per school – typically perceived as a control variable – the improvement on model fit was considerably more impactful, increasing the predictive R^2 by almost 3%. Second, using a non-linear categorical model strongly improved the model's fit of the data, when compared to the simpler linear model traditionally estimated in the field.[18]

Both nuances have important substantive implications for research on teacher bias, which were not picked up in pre-existing work that focused on traditional inference. For example, school effects were typically evaluated through the estimated variance term of the random intercept. They were not typically compared to the other variable in the model in their substantive ability to model the outcomes. As a consequence, a considerable source of variation in tracking had been neglected. Similarly, traditional fit metrics would only indicate an objective preference for the categorical model, but did not provide a normative reflection of the extent to which model fit improves. Importantly, changing to the categorical model also considerably affected the size of estimated biases in tracking [41].

A final advantage of using prediction as a measure of model fit is that it can be used as an approach to address questions of external validity. A recurring question in the social sciences is the persistence of research findings outside of the particular sample used to estimate a model. Prediction makes such assessments more natural than in-sample methods do. For example, consider the following puzzle in Labour Economics. A growing literature is studying the reasons underlying an increase in the amount of residual variance over time when explaining logged hourly wages using a similar set of explanatory variables: education, age, and their interactions [43]. This example lends itself well for an illustration of how the external validity of a model can be assessed from a predictive perspective. By estimating the model on one of the survey years – the first wave, 1973, in this illustration – the performance of the original model

[17] The authors applied stratified Monte Carlo cross validation with $M = 250$. The evaluation set represented a stratified 25% of the total data, ensuring similar proportions of students from each school in the estimation and evaluation sets. Predictive accuracy was thus estimated for a total of 250 folds.

[18] In practice, most researchers studying tracking in The Netherlands have assumed the outcome to be continuous and estimate a simple linear model [42]. As the authors point out, this is predominantly a convenience assumption, as it yields easer-to-interpret coefficient estimates [41].

in terms of fitting the data can be explicitly assessed for datasets collected at later waves.

As the results show (Figure 2, panel E), the 1973 model tracks the performance of models which are retrained to each year quite closely for the first 5 to 10 years, but then starts to deviate.[19] This is insightful for two reasons. First, it provides an indication of the stability of the findings from the 1973 model outside of that sample. Second, it points at a shift in estimated model coefficients from the year 1983 onward, possibly providing additional pointers into the original puzzle. As the models use many interactions, leading to more than 100 coefficients, these differences in the model's fit to separate datasets would be close to impossible to learn from studying the in-sample coefficients of each model in isolation.

Virtue II: prediction provides a benchmarking tool across modeling domains

Although social life is known to be complex to study, simple linear additive models are still the bread-and-butter methods used throughout the social sciences for this very purpose. The reason might be that we have grown accustomed to fitting such models for so long now, that we are reluctant to believe more complicated functional forms are appropriate. A more likely reason is that simple models allow for a more straightforward interpretation of results, which is usually not the case in complicated non-linear models even if they are objectively better at capturing reality. A key problem is that we often don't know whether our models are in fact too simple, prolonging the use of simplistic models in practice. Through benchmarking, prediction provides a way to assess whether the level of complexity in our models is appropriate, as predictive accuracy can be used as a holistic metric of model fit for any type of empirical model [22, 41]. Therefore, it can be used to compare parametrized models with flexible alternatives.

For example, model complexity was a key motivation of the FFC and many research teams heeded this call by innovating extensively on the methods applied.[20] In other words, the heterogeneity in modeling approaches was considerable. As a consequence, conventional model diagnostics would not have sufficed to allow comparisons of the

[19] All independent variables were normalized such that mean difference in average wages across time would not distort the predictive performance of the model fitted in 1973.

[20] Note that the data was mainly appropriate for methods exploiting some form of variable selection. Methods like neural nets should not be recommended as the FFC contained only 4,000 observations, but more than 12,000 variables. This means that the 'curse of dimensionality' would be a serious issue without variable selection or regularization techniques [44]. Similarly, limited N might be one of the most important reasons for the lack of predictive improvement observed in the FFC.

various modeling approaches chosen by the research teams.[21] Opting for prediction on a holdout set solved this problem. As an illustration, the predictive performance of every single submission to the FFC has been visualized in Figure 3 (Panel A).[22] For some outcomes, considerable improvements in predictive accuracy were obtained relative to the benchmark models – e.g., for the GPA outcome – although for most, improvements were negligible – e.g., for the Job Training outcome.

Many of the top performing submissions made use of complex, flexible models but prediction made them directly comparable to the linear benchmarks [45]. This is true at the level of the various submissions (Fig 3, Panel A), but a direct comparison can also be done for a single submission (Fig 3, Panel B). Here, the individual predictions of the top submission for the GPA outcome are visualized together with those from the benchmark model. As can be seen by the LOESS fit, the top submission is slightly better in predicting the extremes of the distribution correctly. That said, the plot also shows that both models still struggle to predict the low-end of the distribution well.

The role of benchmarking is arguably most important to identify functional form misspecification. Comparing the fit of a functional form hypothesized by a researcher with that of a flexible alternative fit to the same data provides an assessment whether the model might miss complexity. This can for example be used to to verify whether the assumption of linear additivity is reasonable. As a concrete example, consider the following simulation study assessing the predictive performances of various Mincerian wage equations [30]. In these models, the effect of additional years of schooling on wages is of central interest. These so-called 'returns to schooling' are typically estimated by relating information on years of education to observed wages while controlling for years of work experience [46].

The Mincerian wage equation is an interesting case because the complexity of the functional form has been innovated upon over the past decades. The functional form started out as a simple linear additive model where log wages were regressed on years of work experience and years of education. In the 80s and 90s, higher order terms on the effect of work experience were proposed and, more recently, a step function in the effect of education has been included into the functional form [46, 47]. In other words, the functional relationship between outcome and explanatory variables was found to

[21] Traditional information criteria and goodness-of-fit measures are often dependent on pre-defined functional forms to correct for the degrees of freedom in the model. When including models from different paradigms of modeling, information criteria lack comparability.

[22] By default, teams submitted the predictions of a null model unless they submitted their own predictions for an outcome, which is why there is considerable density around an improvement of 0% as many teams choose to focus on a select number of outcomes.

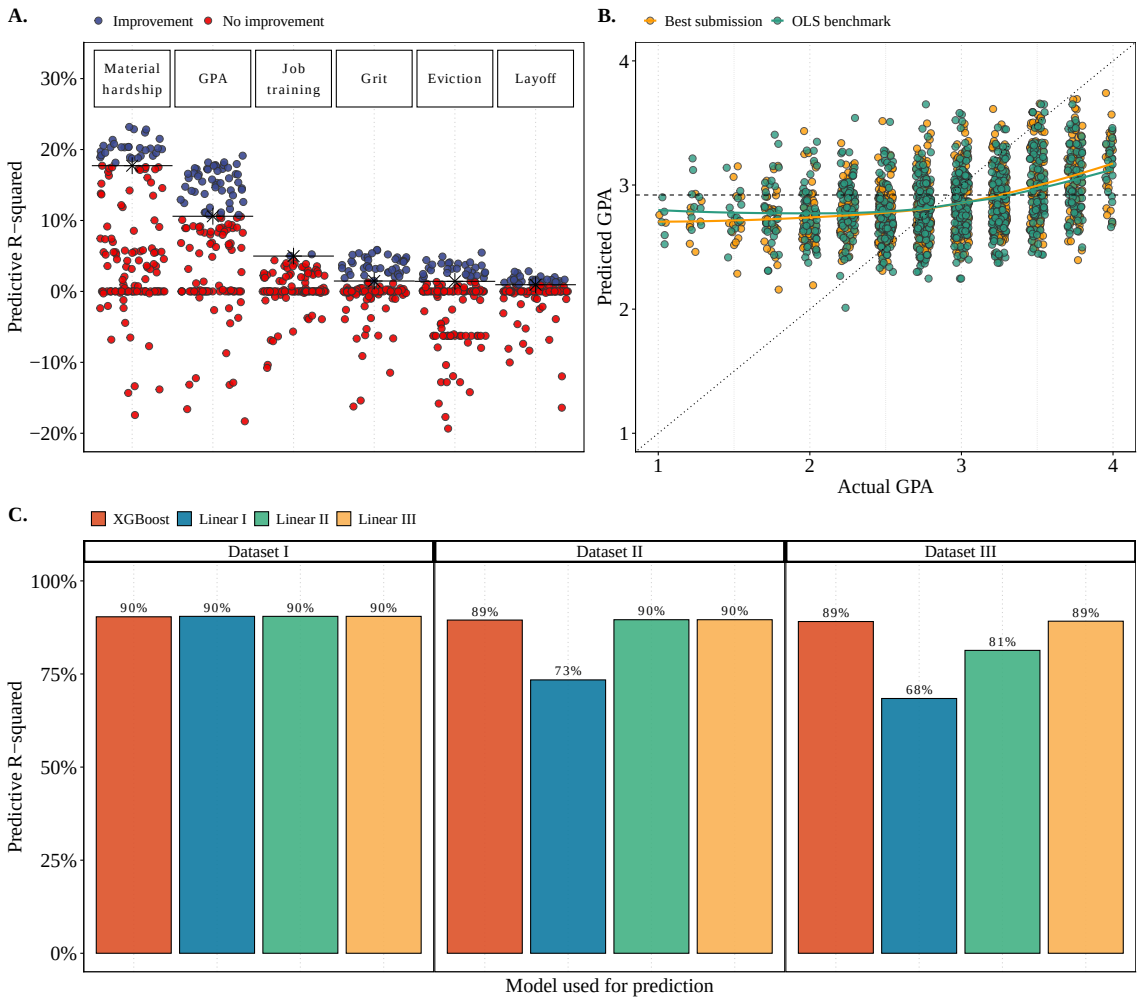

Figure 3. Panel A shows the predictive performance of all submissions to the FFC. The horizontal bars reflect the performance of the benchmark. Blue submissions outperformed the benchmark. Predictive accuracies of 0% indicate that a team did not submit a submission for that specific outcome [10]. Panel B shows the individual predictions of the best performing submission to the FFC for the GPA outcome (yellow) versus the individual predictions of the OLS benchmark (green). GPA was observed in 0.25 point intervals, and have been spread horizontally for illustrative purposes. The dashed horizontal line indicates the mean of the outcome and the dotted diagonal line indicates perfect predictions. The R^2 of the best submission was nearly double (0.19) that of the benchmark (0.11). Panel C shows the predictive accuracy for three simulated datasets using three pre-specified linear functional forms and a flexible non-linear algorithm (XGBoost). The flexible algorithm converges on the true functional form for all three datasets, whereas only the Linear II and Linear III model had the appropriate complexity to fit Dataset II well, and only Linear III had the appropriate complexity to fit Dataset III well [30].

be underspecified and lacking in complexity. Benchmarking can help identify such lack of complexity by comparing a model's performance to that of a flexible alternative which does not constrain the functional form in a particular way. If a flexible model using the exact same covariates strongly outperforms a linear additive model, there is likely a lack of functional form complexity [21, 30].[23]

This rationale is illustrated in Figure 3 (Panel C). Three datasets were simulated that include the same explanatory variables on age, years of education and years of work experience. However, the outcome variable – log wages – is simulated according to a different functional form for each dataset. Specifically, the outcome of the first dataset follows a linear additive function (Linear I), the second a linear additive function including a squared term for the effect of work experience (Linear II), and the third further includes a step function for the effect of education (Linear III).[24] All three outcomes included white noise proportional to about 10% of the total variation – thus capping the potential R^2 at 0.9. These three datasets reflect the functional form development of the Mincerian wage equation observed over the past decades. Four models were fit to each of the datasets, the first being a linear additive model, the second allowing a squared effect for work experience, and the third included the step function. In other words, all three models have the appropriate complexity to fit the first dataset, but only the second and third have sufficient complexity to model the second dataset well, and only the third model can fit the final dataset appropriately. The fourth model was a vanilla XGBoost algorithm, a highly flexible tree-based Machine Learning algorithm.[25]

All four models were used to make predictions on a holdout set of the data.[26] The results show that all four models fit the first dataset well – as should be expected. For the second dataset including the squared term, the first functional form strongly underperforms the alternatives since it cannot model the full complexity in the data. The same holds for the first two functional forms when fitted to the final dataset,

[23]The matter-of-fact comment by Efron and Hastie regarding the use of Random Forests – a flexible machine learning technique – in their 2016 handbook 'Computer Age Statistical Inference' is instructive here: 'if the Random Forest does much better [than a traditional parametrized model], you probably have some work to do' ([21], p. 347).

[24]The models were estimated using a synthetic dataset of 50,000 observations based on the observed age, work experience, and schooling in the 2018 General Social Survey. For the construction of the synthetic sample and exact functional forms underlying the three datasets, see the original study from which this example has been taken which included a fourth functional form where each coefficient varied by sex [30].

[25]The XGBoost algorithm iteratively estimates shallow decision trees to the data, giving more weight to less accurately fit observations after each iteration. Decision trees are non-linear by design, making the XGBoost model able to fit complicated patterns, whilst requiring no a-priori specified functional form [20].

[26]To generate predictions, a holdout set was partitioned off equal to 20% of the total dataset.

which included the step function. Importantly, the flexible algorithm converges on the 'true' model's performance in each of these cases. It thus identifies the need for additional model complexity without requiring the researcher to formulate a pre-specified functional form. Prediction is the key benchmarking metric that allows this comparison of fit across modeling domains.[27,28]

Virtue III: Prediction can help generate insights into complicated models

Perhaps the most important reason why prediction is traditionally underutilized in the social sciences is its supposed lack of explanatory ability. Often, predictions are merely assumed to be useful at the descriptive level at best. However, prediction can actually be used as a method to improve the interpretability of models. First, as a way to make coefficient estimates in non-linear models as interpretable as those from standard linear models by intervening on observed variables and comparing the effects of such an intervention on the predictions. This is especially relevant when dealing with categorical outcome models [31, 32, 41]. Second, a predictive analysis is amenable to more substantial interventional reasoning. For example, to assess the impact of changing a coefficient – as mentioned above – but also to compare the effect of estimating separate models on subsets of the data. Assessing these types of differences by comparing in-sample coefficient alone is practically unfeasible.[29]

To illustrate the first point, consider the mortgage application introduced in Figure 2 (Panel C). Logistic regression models were estimated, making the interpretation of the coefficient estimates less straightforward than a standard linear model which would simply reflect the increase in the value of the outcome when increasing the covariate by one point. This ease-of-interpretation is not available for the logistic regression model. However, prediction provides a way to obtain a similarly intuitive effect size. This is illustrated for the Race coefficient in the mortgage example by comparing the predicted probabilities of success when intervening on the observed Race variable – i.e., changing the observed Race from White to Non-White or vice-versa (Figure 4, Panel A).[30] As can be seen, the average probability of success decreases by 8.4% for

[27] To understand the type of complexity which is missing, recent developments in the field of Explainable A.I. can be used [48]. For the example of the Mincerian Wage Equation shown here, such methods accurately recovered the underlying functional forms used to generate the data [30].

[28] Akin to benchmarking, prediction provides a common metric for researchers to align their research efforts [11, 12]. The astonishing improvements in machine learning methods can in part be attributed to such an alignment on a common goal: improving the predictive accuracy on benchmark sets.

[29] Note that prediction is understood here as a tool to help interpret estimated model coefficients. Prediction does not make model results causally interpretable. Causal interpretation is encoded in the research design and model estimation, whereas prediction is a tool to assess the model post-estimation [49].

[30] A single 5-fold cross validation run was used to generate a full set of predictions.

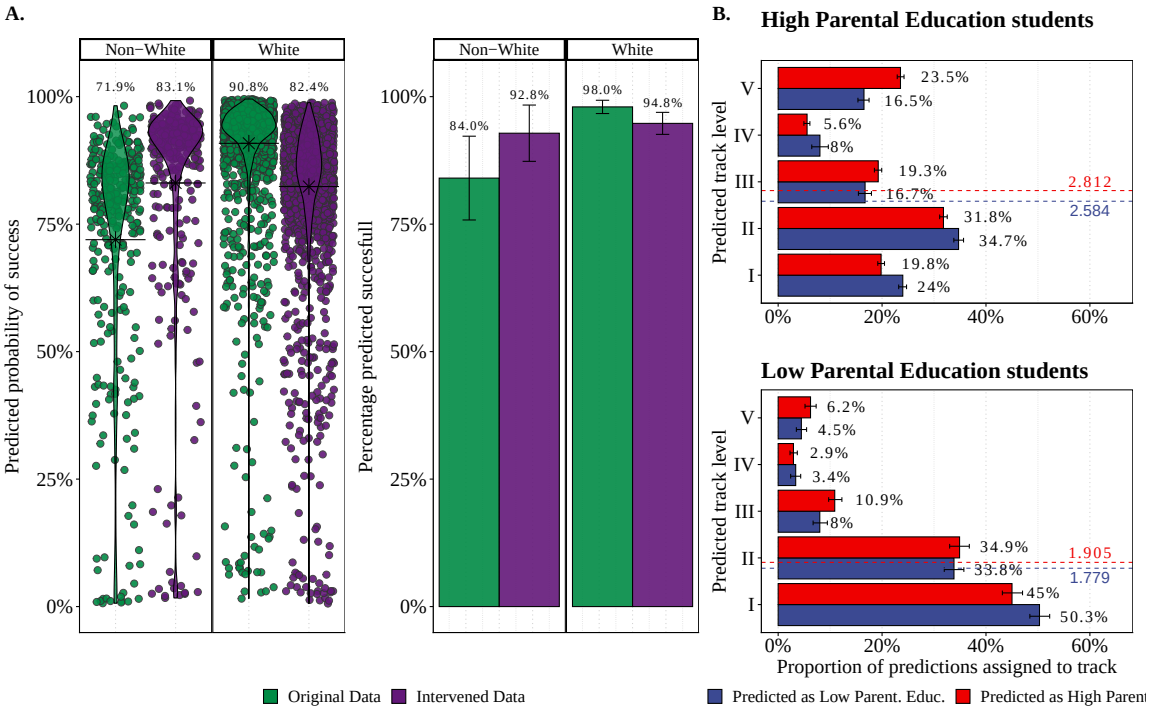

Figure 4. Panel A shows predicted probabilities (left) and outcomes (right) of successful mortgage applications for Whites and Non-Whites in the data. Green values indicate predictions on the original data, whereas purple values show predictions when intervening on the Race variable: i.e. changing the observed Race from Non-White to White and vice-versa. The average predicted probability of success increases by 11.2% for Non-Whites when changing the observed Race, and decreases by 8.4% for Whites when doing so. The predicted number of successes increases from 84% to 92.8% for Non-Whites and decreases from 98% to 94.8% for Whites. Panel B shows the predicted proportion of students in each of the five track levels for students of High Parental Education (top) and Low Parental Education (bottom) when using a model fit to students of High Parental Education (red) or Low Parental Education (blue). The average predicted track level is depicted using the dashed red line when predicting using the High Parental Education model, and by the dashed blue line when predicting using the Low Parental Education model. Confidence intervals, where present, reflect 95% bounds based on variation in the predictive accuracy or number of predicted classes across the evaluated folds.

Whites and increases by 11.2% for Non-Whites when intervening on the Race variable. This type of do-style reasoning is easy to implement when making predictions, and improves on the common practice of simulating predictions by setting other observables to their mean or median values. A predictive approach uses the actual data which is considerably more informative.

Instead of using predicted probabilities, it is also possible to assess the effect of a variable in terms of the actual outcome. An increase in the probability of success does not automatically reflect a similar increase in the expected number of successful applications. Probabilities will yield predictions anywhere between 0 and 1, whereas predicted outcomes will always consists of either 0 (failure) or 1 (success). The outcome-focused equivalent of the mortgage example – where the predicted probabilities are rounded – is given in Figure 4, Panel B.[31] The results illustrate how intervening on the Race variable would increase the share of predicted successes from 84% to 92.8% for Non-Whites, while decreasing the number of predicted successes from 98% to 94.8% for Whites.[32] The differences in the number of actual successes reflect that most Whites already have a high predicted probability of success prior to intervening on the Race variable. Whether to use predicted outcomes or the underlying predicted probabilities will typically depend on the particular use case.

To illustrate the use of prediction to ask more complex interventional questions, I return to the study introduced earlier concerning teacher bias in tracking (Figure 2, Panel D). Pre-existing work chose to model the outcome – track levels – as a continuous variable, allowing for a straightforward interpretation of the estimated coefficients. The study illustrated earlier used an hierarchical Ordered Probit model (H-OPM) leading to difficult to interpret coefficient estimates. However, by using the same interventional reasoning as outlined above, an intuitive assessment of the impact of bias features like Parental Education could easily be generated [41]. This reasoning was then taken a step further by assessing whether the effect of Parental Education could reasonably be captured by the coefficient on a dummy-coded variable, or might manifest itself through the entire model – i.e. whether students of Low Parental Education are assessed differently on observables.

To this effect, separate models were estimated for students with Low and High

[31] The same Repeated Cross Validation routine from Figure 2 (Panel C) was used for the predicted outcomes.

[32] The confidence intervals show that, depending on the fold used for evaluation, the observed percentage of successes in the fold can be higher or lower than the overall mean. This follows from the fact that the data size was limited – especially for Non-White applicants – leading to more variability in the baseline [40]. Note that the intervals should not be assessed to reflect the statistical significance of the race coefficient, but rather variability in the observed pre-intervention probability of the White or Non-White applicants in a specific fold.

Parental Education. This led to one model fitted to students of Low Parental Education, which could be used to make predictions for observations 'as if they were Low Parental Education students' and another model which could do the equivalent but then for High Parental Education students. By making predictions using both models – i.e. predicting outcomes twice – for both the Low and High Parental Education subsets in the data, the implied difference between the two models could be assessed (Figure 4, Panel C).[33] As can be seen, high parental education students on average obtain a track level of 2.81 when assessed as high parental education students. However, when assessed as if they were low parental education students, this average track level drops by about 0.25. Conversely, low parental education students gain about 0.12 track levels when assessed as high parental education students. Determining these differences by comparing the two estimated models would not have been trivial, as multiple coefficients would have to be taken into account jointly including random effects and the cutoff points which are estimated as part of the H-OPM.

Taking Stock

This paper set out to change the underutilization of prediction in the social sciences, where prediction barely features in empirical work. This underutilization occurs for the wrong reasons. Many social scientists confuse the general concept of prediction with more narrow applications, like forecasting, or using predictive accuracy as an optimization measure. Yet, prediction is a much broader and simpler analytical perspective of evaluating models in terms of their ability to accurately fit the outcome of interest. Viewed in this manner, prediction becomes a logical complement to and enrichment of the methods we have grown accustomed to using throughout the social sciences. Importantly, there is absolutely no need to sacrifice traditional models when including prediction in empirical work – contrary to the sometimes dogmatic nature of the philosophical discussion concerning prediction and explanation. Both explanatory and predictive perspectives to analysis can and should be combined.

The benefits that prediction can bring when incorporated into the typical empirical workflow of social scientists are plenty and this paper illustrates but a few. For instance: how basic predictive consciousness can spur important debate in a research field (Figure 2, Panel A), but also how predictions allow for a more detailed assessment of model fit. For example by assessing the fit of individual predictions (Figure 2, Panel B), comparing predictive performance by subsets of the data (Figure 2, Panel

[33] This approach is similar in intuition to decomposition methods that decompose overall group differences in some outcome into compositional differences and effect differences [50]. However, composition methods are often used within the typical linear additive framework which can be restrictive. Exploiting prediction allows a wider variety of modeling approaches to be applied.

C), using different models (Figure 2, Panel D), or testing our models on completely new data (Figure 2, Panel E). Prediction also provides a measure allowing social scientists to compare the fit of wildly varying methodological approaches (Figure 3, Panels A-B). This includes models from different paradigms – like flexible machine learning models – which opens the way to benchmarking our models against flexible alternatives. This provides social scientists with a way to assess whether the models we estimate have the appropriate complexity to fit the data well (Figure 3, Panel C). Finally, prediction is amenable to do-style reasoning and allows us to obtain intuitive associations between variables in non-linear models (Figure 4, Panels A-B), but also to take this interventional reasoning a step further and compare how models estimated to subsets of the data differ in modeling the outcome of interest (Figure 4, Panel C).

In practice, the way in which prediction is applied fundamentally depends on the case at hand. There will be empirical settings where interest in the ability to predict an outcome is less natural than for example in the case of the FFC.[34] Generally speaking, this paper identifies three broad virtues were identified: i) using prediction as a descriptive tool to improve our understanding into the fit of a model, ii) using prediction to normatively compare different models, and iii) to help generate understanding into (complex) model behavior by interventional reasoning. These benefits help address criticisms plaguing the social sciences, like a lack of appreciation for the real-world relevance of research findings, and the use of overly simplistic models to study social life. At the same time, the cynic might question what we exactly gain from adding prediction to empirical work. Can't we identify predictive ability by measures like the R^2? Or use fit metrics like the AIC or BIC to compare models? Aren't there specification tests to diagnose serious misspecification, and can't we identify associations in non-linear models, if we really tried?

The answer to all of these questions is: yes, although there are various nuances that make prediction preferred. For example, existing fit metrics are in-sample and can suffer from overfitting. The R^2 measure does not work well in every design and gives no insights into heterogeneity in a model's fit. Non-parametric regression techniques and non-linear models are relatively complex to estimate and, without a way to illustrate their necessity, most empirical work will remain wedded to the simple linear additive model. There are many more subtle nuances. The broader

[34]Examples could include stylized lab experiments, or research designs like Difference-in-Differences, Regression discontinuity or Matching setups, as either the outcome is less intuitive or interest is fundamentally into an estimated coefficient. Nonetheless, many of the key benefits of prediction which have been illustrated in this paper will be equally relevant to understand the extent to which the model fits the data, irrespective of whether the outcome itself is of central substantive interest or this is less the case.

point is that predictions are disarmingly simple to understand and generate, and can serve multiple goals at the same time as the examples in this paper illustrate. Predictions also improve transparency by inviting a more rigorous assessment of a model's ability to fit the data than most aggregated in-sample metrics. Requiring researchers to make predictions is a much better way to diagnose model limitations than allowing researchers to (cherry-)pick their own robustness checks or descriptives. Finally, prediction paves the way for exciting methods from other domains, like that of machine learning, into the workflow of social scientists. Methods that should become complementary to the social sciences.

Hopefully, this paper can help social scientists decouple prediction from some of the field's most intriguing and sometimes heated discussions. For example whether explanation should imply prediction or what the role of machine learning should be in the social sciences. Although interesting, they are ultimately a distraction from what prediction as an analytical tool has to offer the social sciences. In sum, prediction's complete lack of complexity, transparency, intuitive nature, and flexibility to build on the methods we have used for decades – rather than forcing researchers into new techniques – are all substantial assets, that come at virtually no price to include into our work. In other words, prediction is truly one of those illustrious free lunch buffets which social scientists continue to ignore at their own peril.

References

1. Shmueli, G. *et al.* To explain or to predict? *Statistical Science* **25,** 289–310 (2010).
2. Miech, R., Pampel, F., Kim, J. & Rogers, R. G. The enduring association between education and mortality: the role of widening and narrowing disparities. *American Sociological Review* **76,** 913–934 (2011).
3. Athey, S. & Imbens, G. W. Machine learning methods that economists should know about. *Annual Review of Economics* **11,** 685–725 (2019).
4. Watts, D. J. Common sense and sociological explanations. *American Journal of Sociology* **120,** 313–351 (2014).
5. Cederman, L.-E. & Weidmann, N. B. Predicting armed conflict: Time to adjust our expectations? *Science* **355,** 474–476 (2017).
6. Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M. & Leskovec, J. *Can cascades be predicted?* in *Proceedings of the 23rd international conference on World wide web* 925–936 (2014).
7. Booth, H. Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting* **22,** 547–581 (2006).
8. Salganik, M. J., Lundberg, I., Kindel, A. T. & McLanahan, S. Introduction to the special collection on the fragile families challenge. *Socius* **5,** 2378023119871580 (2019).
9. Donoho, D. 50 years of data science. *Journal of Computational and Graphical Statistics* **26,** 745–766 (2017).
10. Salganik, M. J. *et al.* Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences* **117,** 8398–8403 (2020).
11. Hofman, J. M., Sharma, A. & Watts, D. J. Prediction and explanation in social systems. *Science* **355,** 486–488 (2017).
12. Cranmer, S. J. & Desmarais, B. A. What can we learn from predictive modeling? *Political Analysis* **25,** 145–166 (2017).
13. Greenhill, B., Ward, M. D. & Sacks, A. The separation plot: A new visual method for evaluating the fit of binary models. *American Journal of Political Science* **55,** 991–1002 (2011).

14. Rahal, C., Verhagen, M. D. & Kirk, D. Machine Learning in the Social Sciences: Amara's Law and an inclination across Foster's S-Curve? *SocArXiv.* doi:10.31235/osf.io/gydve (2021).

15. Hempel, C. G. & Oppenheim, P. Studies in the Logic of Explanation. *Philosophy of Science* **15,** 135–175 (1948).

16. Freedman, D. A. Statistical models and shoe leather. *Sociological Methodology,* 291–313 (1991).

17. Hedström, P. & Ylikoski, P. Causal mechanisms in the social sciences. *Annual Review of Sociology* **36,** 49–67 (2010).

18. Hofman, J. M. *et al.* Integrating explanation and prediction in computational social science. *Nature* **595,** 181–188 (2021).

19. Breiman, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science* **16,** 199–231 (2001).

20. Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning: data mining, inference, and prediction* (Springer Science & Business Media, 2009).

21. Efron, B. & Hastie, T. *Computer age statistical inference* (Cambridge University Press, 2016).

22. Hindman, M. Building better models: Prediction, replication, and machine learning in the social sciences. *The ANNALS of the American Academy of Political and Social Science* **659,** 48–62 (2015).

23. Salganik, M. *Bit by bit: Social research in the digital age* (Princeton University Press, 2019).

24. Mullainathan, S. & Spiess, J. Machine learning: an applied econometric approach. *Journal of Economic Perspectives* **31,** 87–106 (2017).

25. Molina, M. & Garip, F. Machine learning for sociology. *Annual Review of Sociology* (2019).

26. Grimmer, J., Roberts, M. E. & Stewart, B. M. Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science* **24,** 395–419 (2021).

27. Athey, S. Beyond prediction: Using big data for policy problems. *Science* **355,** 483–485 (2017).

28. Ribeiro, M. T., Singh, S. & Guestrin, C. *"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* 1135–1144 (2016).

29. Lundberg, S. M. & Lee, S.-I. *A unified approach to interpreting model predictions* in *Proceedings of the 31st international conference on neural information processing systems* 4768–4777 (2017).

30. Verhagen, M. D. Identifying and Improving Functional Form Complexity: A Machine Learning Framework. *SocArXiv.* doi:10.31235/osf.io/bka76 (2021).

31. Breen, R., Karlson, K. B. & Holm, A. Interpreting and understanding logits, probits, and other nonlinear probability models. *Annual Review of Sociology* **44,** 39–54 (2018).

32. Hanmer, M. J. & Ozan Kalkan, K. Behind the curve: Clarifying the best approach to calculating predicted probabilities and marginal effects from limited dependent variable models. *American Journal of Political Science* **57,** 263–277 (2013).

33. Yousef, W. A. A Leisurely Look at Versions and Variants of the Cross Validation Estimator. *arXiv.* doi:arXiv:1907.13413v3 (2020).

34. Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)* **36,** 111–133 (1974).

35. Simonsohn, U., Simmons, J. P. & Nelson, L. D. Specification curve analysis. *Nature Human Behaviour* **4,** 1208–1214 (2020).

36. Young, C. Model uncertainty and the crisis in science. *Socius* **4** (2018).

37. Kohavi, R. *et al. A study of cross-validation and bootstrap for accuracy estimation and model selection* in *International Joint Conference on Artificial Intelligence* 2, 1137–1145 (1995).

38. Bates, S., Hastie, T. & Tibshirani, R. Cross-validation: what does it estimate and how well does it do it? *arXiv.* doi:arXiv:2104.00673 (2021).

39. Bengio, Y. & Grandvalet, Y. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research* **5,** 1089–1105 (2004).

40. Munnell, A. H., Tootell, G. M., Browne, L. E. & McEneaney, J. Mortgage lending in Boston: Interpreting HMDA data. *The American Economic Review,* 25–53 (1996).

41. Verhagen, M. D. To predict and explain. How prediction improves our understanding of models: an application to the study of teacher bias. *SocArXiv.* doi:10.31235/osf.io/y6mnb (2021).

42. Van Leest, A., Hornstra, L., van Tartwijk, J. & van de Pol, J. Test-or judgement-based school track recommendations: Equal opportunities for students with different socio-economic backgrounds? *British Journal of Educational Psychology* **91,** 193–216 (2021).

43. Lemieux, T. Increasing residual wage inequality: Composition effects, noisy data, or rising demand for skill? *American Economic Review* **96,** 461–498 (2006).

44. Bishop, C. M. *Pattern recognition and machine learning* (Springer, 2006).

45. Rigobon, D. E. *et al.* Winning Models for Grade Point Average, Grit, and Layoff in the Fragile Families Challenge. *Socius* **5** (2019).

46. Lemieux, T. *The "Mincer equation" thirty years after schooling, experience, and earnings* in *Jacob Mincer a pioneer of modern labor economics* 127–145 (Springer, 2006).

47. Heckman, J. J., Humphries, J. E. & Veramendi, G. Returns to education: The causal effects of education on earnings, health, and smoking. *Journal of Political Economy* **126,** S197–S246 (2018).

48. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & Müller, K.-R. *Explainable AI: interpreting, explaining and visualizing deep learning* (Springer Nature, 2019).

49. Pearl, J. *Causality: models, reasoning, and inference* (Cambridge University Press, 2000).

50. Fortin, N., Lemieux, T. & Firpo, S. *Decomposition methods in economics* in *Handbook of Labor Economics* 1–102 (Elsevier, 2011).