BJM BRITISH JOURNAL of MANAGEMENT
BRITISH ACADEMY OF MANAGEMENT
British Journal of Management, Vol. 00, 1–18 (2022)
DOI: 10.1111/1467-8551.12678

Towards a Paradigm Shift: How Can Machine Learning Extend the Boundaries of Quantitative Management Scholarship?

Danat Valizade,[1] Felix Schulz[1] and Cezara Nicoara[2]
[1]Leeds University Business School, Maurice Keyworth Building, Woodhouse, Leeds, LS2 9JT, UK, [2]Newcastle University Business School, 5 Barrack Rd, Newcastle upon Tyne, NE1 4SE, UK
Corresponding author email: d.valizade@leeds.ac.uk

Management scholarship is beginning to grapple with the growing popularity of machine learning (ML) as an analytical tool. While quantitative research in our discipline remains heavily influenced by positivist thinking and statistical modelling underpinned by null hypothesis significance testing, ML is increasingly used to solve technical, computationally demanding problems. In this paper, we argue for a wider, more systematic adoption of the key tenets of ML in quantitative management scholarship, both in conjunction with and, where appropriate, as an alternative to canonical forms of statistical modelling. We discuss how ML can extend the boundaries of quantitative management scholarship, help management scholars to unpack complex phenomena, and improve the overall trustworthiness of quantitative research. The paper provides a representative review of the use of ML to date and uses a worked example to demonstrate the value of ML for management scholarship.

Introduction

In a seminal work on automation and technological change, Brynjolfsson and McAfee (2016) shared an almost flawless experience of testing a driverless car, which beat their wildest expectations of the cognitive tasks that computers can accomplish. Underlying this rapid advancement is the revolution in machine learning (ML), a multidisciplinary field that combines insights from computer science and statistical learning to build algorithms capable of learning patterns and associations from data without human supervision (Breiman, 2001a; Friedman, 2006). ML has spurred on cutting-edge methodological debates across the social sciences on how to utilize the better predictive accuracy and replicability of ML models relative to traditional methods of statistical analysis (Athey & Imbens, 2015, 2019; Grimmer, 2015; McFarland *et al*., 2016; Molina & Garip, 2019). In management scholarship, ML has been recognized and implemented for decades (Bennell *et al*., 2006), albeit mostly as a niche tool for solving technical problems (e.g., working with extremely large datasets, unstructured/text data). This study contributes to a nascent methodological debate of how ML can be leveraged to advance management scholarship (Leavitt *et al*., 2021; Choudhury *et al*., 2021; Shrestha *et al*., 2021).

We engage with ML holistically and argue that a wider adoption of its key principles can encourage new ways of thinking about the types of research questions that can be addressed with quantitative methods and increase the replicability and trustworthiness of research findings. The paper begins with a review of existing tensions in quantitative management scholarship, focusing on the prevalence of positivist thinking, null hypothesis significance testing (NHST), and statistical modelling as a tool to generate valid, replicableknowledge (Mingers, 2006; Scandura & Williams, 2000). We then place ML in the context

of ongoing methodological innovation and argue that it can advance management scholarship in three interrelated ways. First, as a methodology underpinned by a model-agnostic, algorithmic approach to data analysis, ML is capable of leveraging the explanatory power of observational data (Breiman, 2001a, p. 199). ML can ensure a more balanced methodological paradigm, where abductive and exploratory quantitative studies are on an equal footing with deductive, hypothesis-testing contributions. That will carve out space for theory-generating quantitative research and unexpected, counterintuitive findings that go against current norms and expectations (Ethiraj *et al*., 2016; Leavitt *et al*., 2021).

Second, ML rests on out-of-sample predictions whereby an algorithm is trained on a subset of data and evaluated depending on how well it performs on previously unseen data. Management scholarship can benefit from a greater use of out-of-sample predictions and concomitant procedures of cross-validation and regularization in addition to in-sample goodness of fit in canonical statistical models. That will contribute to the better generalizability and replicability of quantitative findings (Blockeel & Struyf, 2002; Sarstedt & Danks, 2021). Third, wider use of non-parametric, so-called 'black box', learning algorithms in conjunction with contemporary techniques of global and local interpretable ML will contribute new insights to theory testing and causal inference by allowing us to unpack complex, non-monotonous and non-linear effects that canonical statistical models overlook (Leavitt *et al*., 2021).

We provide a worked example of ML application with real-world data from the Eurofound Company Survey to demonstrate how these steps can work in practice. In the worked example, we focus on supervised ensemble algorithms and regression analysis, a class of methods frequently deployed by management scholars. Using firms' product and service innovation as an outcome variable, we show that: (a) additional focus on generalizability beyond directly observed data and predictive accuracy has implications for the interpretation of empirical results; (b) non-parametric algorithms are useful for predicting rare events (e.g., innovative firms in the worked example); and (c) techniques of interpretable ML can detect patterns that parametric regression analysis misses. In the concluding sections, we reflect on the limitations of ML and provide a note of caution in how it should be applied to avoid widespread abuses of algorithmic modelling.

Background

Entrenched tensions in quantitative management scholarship

Quantitative management scholarship is renowned for methodological conservatism. It continues to rely (often implicitly) on positivist thinking, where theoretical assumptions are tested by a posteriori knowledge derived from statistical modelling. That ensures universality of the graduate curriculum and the peer review process. However, recent decades have witnessed a growing critique of the current paradigm amid calls for innovation and methodological diversity. The critique proceeds along three interrelated strands.

First, a continuing reliance on positivist thinking and confirmatory hypothesis testing restricts the scope and depth of research questions that can be answered with quantitative methods. The current imperative, with historical roots in positivism and naïve realism (Mingers, 2006; Wicks & Freeman, 1998), reduces theoretical postulates to confirmatory, theory-driven statements. This prescribes that research hypotheses take a narrow form, for example: 'There is an association between *X* and *Y*' or '*X* positively affects *Y*'. Management research based on such hypotheses attempts to positively verify theoretical assumptions, while pushing data-driven and theory-generating studies to the fringes of quantitative management scholarship.

The second issue is a de facto use of NHST as a yardstick to draw theoretical conclusions about relationships between the variables of interest (Amrhein *et al*., 2019; Bzdok *et al*., 2018; Gigerenzer & Marewski, 2015; Ziliak & McCloskey, 2008). As an inductive process of quantifying the probability – under the null hypothesis – of finding the same or more extreme results than in the data at hand (Perezgonzalez, 2015; Szucs & Ioannidis, 2017), NHST is not a measure of the probability of the null or alternative hypotheses being true (Ziliak & McCloskey, 2008). Because NHST cannot confirm the null hypothesis, there is a high risk of a quantitative study turning into a quest for statistically significant results, leading to p-hacking, namely a process of selectively reporting significant results while omitting non-significant effects and

under-reporting low effect sizes (Killeen, 2005). P-hacking and publication bias are the main causes for the replication crisis, highlighting the failure of our discipline to reproduce the results of many prominent studies (Bergh *et al*., 2017; Christensen & Miguel, 2018; Duvendack *et al*., 2017; Goldfarb & King, 2016; Nelson *et al*., 2018; Pagell, 2020; Open Science Collaboration, 2015).

Third, the generation of valid, rigorous knowledge continues to be understood as a process of fitting an a priori model to observational data (Mingers, 2006). How well a model performs outside directly observed data is not a problem routinely addressed by canonical statistical modelling (Gigerenzer & Marewski, 2015). This becomes increasingly problematic because in management scholarship, statistical models are built predominantly on stringent distributional assumptions concerning an unknown data generation process (in statistical theory, real-world, unobserved processes that produce data). Such assumptions are unlikely to hold in inference with observational, non-experimental datasets used for studying complex, non-monotonous effects (e.g., Schulz *et al*., 2022).

Methodological innovation

A growing perception among scholars that the foregoing tensions impede methodological innovation has led to several developments. First, major steps have been undertaken to increase the transparency and trustworthiness of statistical modelling. These include meta-analysis as a means to arrive at generalizable scientific conclusions (Sharpe & Poets, 2020), alongside less drastic measures aimed at reducing p-hacking and taking greater care of type I and type II errors: for example, by paying close attention to the representativeness of sampling frames and reducing selection bias in sampling procedures; and by reporting effect size, confidence intervals, and statistical power (Pagell, 2020; Simmons *et al*., 2011). These initiatives are being implemented to ensure that where NHST is used, it is carried out appropriately and with caution (Simmons *et al*., 2011). More radical solutions assume dropping NHST altogether or turning to viable alternatives, for example Bayesian hypotheses testing (Andraszewicz *et al*., 2015).

Second, quantitative management scholarship has started to shift towards more robust approaches to inference and establishing causality. Among the key measures are a move away from primary surveys drawn from samples of convenience towards large, representative administrative and panel datasets that are less susceptible to omitted variable bias and endogeneity. Quasi-experimental design, Bayesian networks, and causal diagrams are among the initiatives aimed towards estimating treatment effects with greater robustness (e.g. Rizov *et al*., 2016).

Third, management scholars have started paying greater attention to exploratory quantitative studies, such as cluster analysis, latent class models, and sequence analysis (Anderson & Maxwell, 2017). More recently, scholars have called for an integration of predictive analysis into management scholarship as a necessary addition to the currently prevalent modelling aimed at producing statistically unbiased estimates (Hofman *et al*., 2020). This is in line with Friedman's assertion that 'a theory [that is] is realistic "enough" can be settled only by seeing whether it yields predictions that are good enough for the purpose in hand or that are better than predictions from alternative theories' (Friedman, 1953, p. 182).

In what follows, we introduce the core principles of ML and demonstrate how a wider adoption of such principles both chimes with the aforementioned pathways of methodological change and can foster further innovation by increasing the diversity of research problems that quantitative scholars can address, improving real-life impact and the replicability of quantitative studies.

Setting the agenda for a paradigm shift

Machine learning as a distinctive methodological paradigm. ML algorithms have been around for decades, but it is only recently that the rise in computer power has unleashed their innovative potential. Inspired by the initial success in pattern recognition, researchers have adopted algorithmic modelling in business practice to optimize delivery routes, forecast demand and performance, detect fraud, and automate hiring solutions (Choudhury *et al*., 2021). Driven by the wide applicability of ML software (i.e., IBM Machine Learning, Google Cloud AI, or Microsoft Azure), the business benefits of ML for management and strategic decision making become evident, as shown in the recent examples in Table 1. These examples outline the practical application of ML pattern

Table 1. A representative selection of machine learning (ML) applications in practice

Strategic objective	Company name and market	ML software used	How ML is implemented	ML results
Improve sales and customer satisfaction among loyalty card customers	Boots, UK	IBM Machine Learning	Boots, the UK's leading pharmacy-led health and beauty retailer, uses IBM Machine Learning to match transactions to individual loyalty card customers and sets personalized marketing goals for each person based on their unique purchasing histories and preferences. The IBM predictive models support the analytics-driven marketing processes at Boots, transforming its vast quantities of transactional data into sources of actionable insights. With IBM, the company deploys highly targeted marketing messages that reach customers via their preferred contact method.	The tangible impact of using ML included a 70% increase in the annual volume of tailored messages, boosts in incremental spend for loyalty card customers and engagement of customers with the Boots brand (IBM, 2022).
Improve customer experience and personalized recommendations	IKEA, global	Google Cloud AI	Furniture retail company IKEA employed Google Cloud AI in the development of more personalized consumer shopping and real-time recommendations to support business goals, including optimizing for conversion rate, click-through rate, and revenue. Looking for opportunities to improve the consumer journey, IKEA was able to retrieve high-quality quantitative consumer data and develop relevant recommendation solutions through personalization.	Through the Recommendation AI system, IKEA delivered greater personalized and real-time customer recommendations, leading to: an increase in the number of relevant recommendations displayed on a page by +400%; a +30% improvement in click-through rates; and an overall average order value increased by +2% globally (Google Cloud, 2022).
Cost optimization, reduction, and management	ASOS, UK	Microsoft Azure Machine Learning	British online retailer ASOS used Microsoft Azure Machine Learning during the COVID-19 pandemic to implement a cost optimization strategy focused on managing costs effectively and reducing the number of staff affected by furlough. ML allowed the firm to optimize the utilization of costly machines and large storage accounts, streamlining firm operations and providing visibility and a greater predictive power to better inform long-term decisions.	The financial impact was significant, leading to a 15–20 per cent reduction in annual spend, a return higher than the 5–10 per cent originally anticipated (Moln-Page, 2021).

Table 1. (Continued)

Strategic objective	Company name and market	ML software used	How ML is implemented	ML results
Enhance customer experience and combat decision fatigue	Discovery+, global	Amazon Personalize	Discovery+, a multinational mass media company, uses Amazon Personalize to understand its consumer audiences and help curate content that matches the specific interests of individual consumers. It also helps to support consumers in their streaming journey, by providing a solution to combat the decision fatigue resulting from too many choices that become overwhelming and can take away from the viewer experience. Rather than providing generic content, Discovery turned to Amazon Personalize to enable tailored content suggestions for their Discovery+ streaming platform users.	Amazon Personalize enabled the company to build a 'product with a very fast time to market and a flexible and scalable solution. Amazon Personalize was a really solid fit.' (Husain, 2021, p. 1).
Speed up customer service delivery; reduce contract delivery time; increase revenue	BT, UK	Adobe Document Cloud	British telecommunications multinational, BT, incorporated Adobe Document Cloud to support its strategic decision to eliminate the need to sign paper contracts and instead offer customers an integrated digital signing option. Adobe Document Cloud was used to improve the time consumers take to get started with BT by accelerating the contract process. It also enabled the redirection of staff from admin work to sales and customer service, thus leading to additional increases in revenue.	Using Adobe Document Cloud led to a 99% reduction in average contract turnaround, from 28 days to seven hours, as well as £630K savings annually in associated paper-related costs. The integrated, simple one-step contract sign service allowed for a further 98% decrease in time spent creating, chasing, and uploading contracts (Swan, 2022).

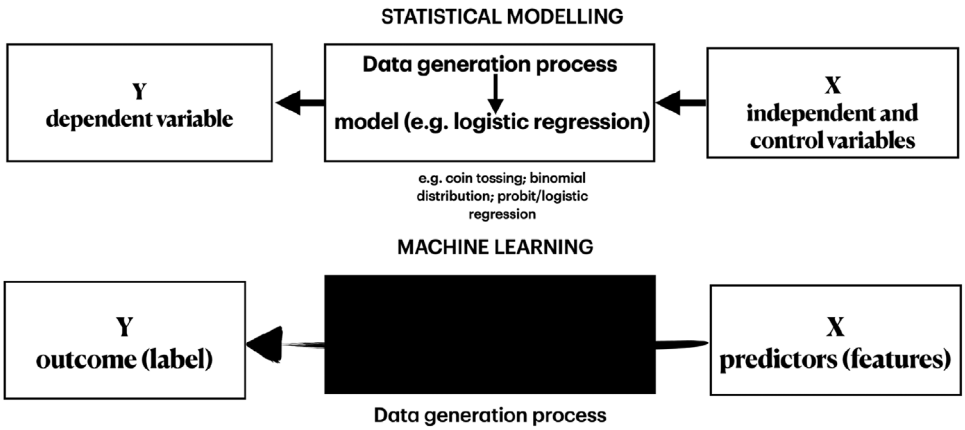


Figure 1. Statistical modelling and machine learning (adopted from Breiman, 2001a, p. 199)

recognition and predicting capabilities in the *formulation*, *implementation*, and *evaluation* of management issues of strategic importance, focusing on the tangible, performance-related outcomes of ML.

ML has started to proliferate in management research, where it is seen mostly as a tool to solve specific technical problems. A growing body of work equates ML and big data (e.g., Tonidandel *et al.*, 2018), while other works focus on the practicalities of big data analytics rather than on ML or algorithmic modelling in a methodological sense (Batistič & van der Laken, 2019). Still, a significant number of emerging research publications suggest that ML applications are prospering across various management areas (Garg *et al.*, 2021). A review of representative examples of ML applications already present in the management literature is outlined in the online Appendix, Table 1A. Common among these examples is a piecemeal adoption of ML to work with text data (e.g., Schmitt *et al.*, 2021), solve forecasting tasks (e.g., Hwang *et al.*, 2020), or establish subgroups of respondents in heterogeneous populations (e.g., Kellard & Sliwa, 2016).

Few studies in management journals recognize ML as a standalone methodological paradigm whose core principles can be adopted in quantitative research irrespective of a specific method or estimation strategy. Leavitt *et al.* (2021) place emphasis on the theory-building potential of ML, echoing similar arguments in other disciplines (Adner *et al.*, 2019; Tonidandel *et al.*, 2018). Two features make ML a distinctive methodological paradigm:

(1) we seldom directly observe the nature or process of the phenomenon and should therefore make no strict mathematical assumptions about it. Hence, the common reference to a black box in relation to ML (Breiman, 2001a, see Figure 1);
(2) there is a focus on generalizability by way of extending the algorithm to previously unseen data and maximizing predictive accuracy.

In the absence of a model that can approximate the effect of X on Y, attention switches from confirming a priori deduced hypotheses (although that is not an uncommon task in ML, see Cornwall *et al.*, 2021) to building an algorithm to predict/explain the phenomenon in question accurately. As we remain agnostic about the nature of the data generation process, algorithms ought to be generalizable beyond the current set of data to retain predictive accuracy when exposed to previously unseen data. Underlying that is the principle whereby higher predictive accuracy can be achieved by learning to predict future outcomes from past experiences. Empirically, this is achieved by separating the data into three sets: training, test, and validation. The algorithm is trained to 'learn' patterns from the training data; the algorithm is then extended to the test set to evaluate the algorithm's predictive accuracy. Thereafter, the chosen algorithm is fine-tuned to increase predictive accuracy and verified once again on the validation set. Schematically, the process of ML is shown in Figure 2 (including data pre-processing and a post hoc explanation of the algorithm).

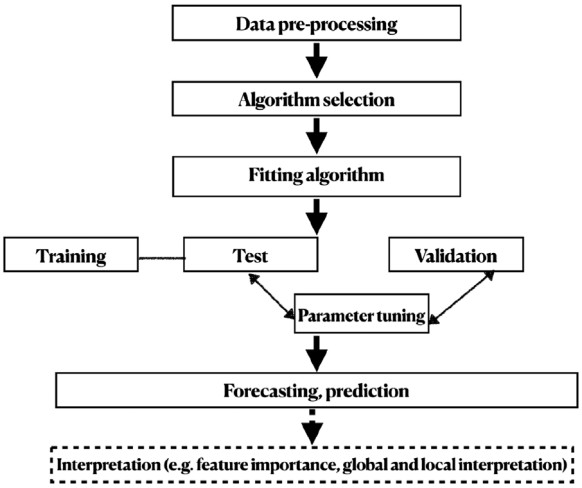


Figure 2. The process of machine learning

There are three interrelated ways in which ML can foster innovation in quantitative management scholarship. First, a wider adoption of ML will diversify the types of questions that can be addressed with quantitative methods. It is important to note that ML is perfectly commensurate with hypothesis-testing research, albeit not strictly in a confirmatory fashion such that multiple competing hypotheses and broader research questions can be explored simultaneoulsy (for example, how does organizational isomorphism influence product innovation?). Yet, in the first place, ML will carve out space for genuinely exploratory questions (e.g., which factors can explain employee attrition among minority ethnic professionals?) that will be perceived as equally valid relative to confirmatory hypothesis testing.

The main difference between ML applications and statistical modelling lies in their approach to theory. While the current approach to theory testing in management scholarship is largely positivist (theory – hypotheses – statistical modelling), ML assumes an abductive logic: it undertakes an iterative process between patterns and associations emerging from the data and provides plausible theoretical explanations for these patterns (Leavitt *et al*., 2021). ML cannot, in principle, positively verify a theory, as it seeks the most likely explanation at each step and accepts that such explanations might change as new data become available. Apart from expanding the scope of questions available to quantitative researchers, ML offers, as Leavitt *et al*. (2021, p. 754) note, a potential for 'serendipitous discovery', where unexpected, theory-defying results are not dismissed out of hand (although, as with any method, researchers should be weary of discovering spurious effects).

Overall, then, applying basic principles of ML learning in management research should not necessarily affect a pro forma academic paper: the theoretical background-questions (hypotheses)-findings structure can remain intact. Qualitatively, though, the language, inner logic, and an understanding of what is considered rigorous knowledge can differ significantly, reflecting the abductive reasoning behind ML.

Algorithmic learning: producing accurate predictions and generalizable knowledge

Which elements of the ML process depicted in Figures 1 and 2 can serve as a useful knowledge-generation tool in quantitative management scholarship? Achieving a balance between in-sample goodness of fit and out-of-sample predictive accuracy (the latter is seldom used in management research) is the first step (Sarstedt & Danks, 2021). Instead of relying almost exclusively on sample goodness of fit, statistical significance, and confidence estimates as the universal characteristics of model performance, management scholars should begin to consider how an algorithm performs on the training, test, and validation sets. By implication, the issues of overfitting and underfitting (Figure 3) ought to be given the foremost consideration. When overfitting occurs, the algorithm picks up too much noise from the training set and predicts it quite well; however, predictive accuracy drops substantially when it is applied to new data. In statistical modelling, scholars are often interested in how well the model fits the data, for instance, emphasizing how much variation in the outcome variable can be explained by the proposed model, commonly referred to by the coefficient of determination. High determination can, however, also be the result of overfitting, where additional variables add noise to the model instead of better describing actual relationships between variables (Hawkins, 2004).

In ML, out-of-sample predictive power and overfitting are addressed by validation, cross-validation and other similar procedures (Breiman, 2001a, 2001b; Beleites *et al*., 2005; Kim, 2009). In cross-validation, data are randomly split into a number (n) of groups with approximately the same

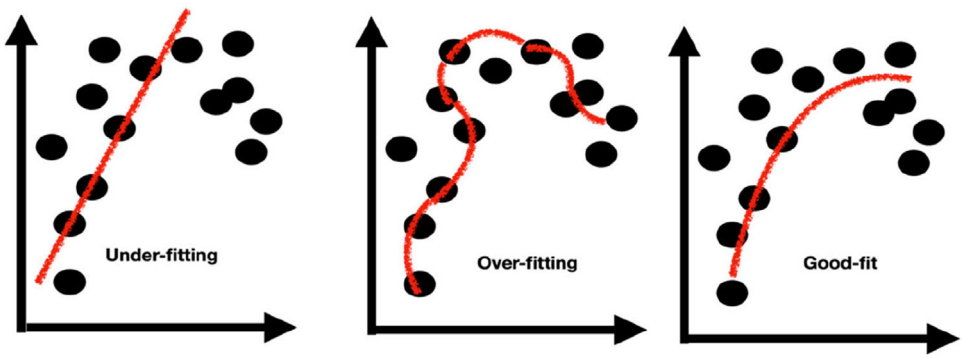


Figure 3. Under- and overfitting (Easterby-Smith et al.*, 2021, p. 391) [Colour figure can be viewed at wileyonlinelibrary.com]*

number of observations in each. The first group remains as the validation sample, while the algorithms are trained on the remaining $n-1$ groups. In this way, each sample is used both to train the model and as a hold-out sample (James *et al.*, 2017). To help solve the overfitting problem, ML applies regularization techniques that aim to minimize the instances of picking up noise from the training set by adding extra information to the algorithms at each step and penalizing the model for it, thus shrinking the importance of predictors with lower predictive power. Essentially, the algorithms are trained to detect automatically which features are relevant for the outcome and which are not (Cohen & Jensen, 1997).

None of the aforementioned procedures are alien to statistical modelling (see for instance Stoltzfus, 2011). In reliability analysis, test–retest techniques are widely used, and various resampling techniques (e.g., jackknife, bootstrapping) are commonplace. What sets ML apart from statistical modelling is the focus on predictive accuracy and performance on previously unseen data. Management scholars would benefit from greater attention to these issues as indicators of model performance without necessarily abandoning canonical estimation procedures in statistical modelling (e.g., OLS or maximum likelihood regression). That will mitigate the risks associated with potential abuses of NHST.

The explanatory potential of algorithmic learning

A more balanced approach between the focus on unbiased estimates and predictive accuracy does not, in itself, require further adoption of more complex, black box learning algorithms. Increasingly, however, management scholarship can benefit from the use of more complex models. For example, Muchlinski *et al.* (2016) demonstrate that in the analysis of rare events, canonical statistical models (e.g., logistic regression) fail to predict such events correctly in out-of-sample data, whereas non-parametric learning algorithms (e.g., random forest, RF) provide much higher predictive accuracy. This is a potentially important quality of algorithmic learning, in that more accurate predictions of the events and parameters that management scholars are interested in have greater theoretical value. As shown in previous research, a model that more accurately predicts the likelihood of employees with specific demographic and occupational characteristics quitting their jobs has significant consequences for the theoretical interpretation of empirical results (Sajjadiani *et al.*, 2019). Thus, closer attention needs to be paid to the predictive and explanatory potential of so-called black box ML algorithms.

There are several reasons why such algorithms can more adequately capture the phenomena commonly studied by management scholars. First, black box algorithms tend to work well with complex problems that are difficult to force into the stringent assumptions of statistical modelling. Algorithms generally show better predictive accuracy, and they can also unpack non-monotonous effects by identifying crucial inflection points (Grimmer, 2015). Second, such algorithms are better at dealing with the variance–bias trade-off, the aspect of choosing a model in the training data that best fits the validation data rather than the training dataset itself. ML ensemble methods such as RF (an algorithm based on fitting multiple decision trees generated by bootstrapping or bagging)

have been shown to work well in achieving this by resampling the training data and dealing with overfitting by being robust to noise and outliers (Breiman, 2001b).

Strictly speaking, black box is a misnomer in that it implies a lack of transparency in how an algorithm arrives at certain results. That is not entirely correct, as the mathematical and computational rationale for many ML algorithms is well established and can be worked through (with the exception of some deep learning algorithms). The term black box refers to agnosticism in relation to data generation processes and corresponding distributions in the population, whereas a general use of the term emphasizes a lack of easily derivable estimates from the model akin to the regression coefficients and confidence intervals in canonical statistical analysis (Breiman, 2001a; Svetnik *et al*., 2003). The field of interpretable ML is rapidly growing. In ensemble algorithms, an early solution was to utilize feature importance scores proxied by impurity-based metrics (the likelihood that a randomly chosen feature predicts the outcome incorrectly) and permutation importance scores (estimating the effect on predictive accuracy by randomly reshuffling predictors) (Choudhury *et al*., 2021; Svetnik *et al*., 2003). Another way of peeping inside the black box was to extract a single easily interpretable predictive model (for instance, one tree from all the trees in RF) (Choudhury *et al*., 2021). However, these solutions fell short of the 'elegance' of confidence estimates afforded by more traditional regression models. That has changed dramatically in the past decade, with the emergence of various techniques of global and local interpretation of learning algorithms (ElShawi *et al*., 2020; Slack *et al*., 2020).

The idea behind *global* interpretive ML is to extract effects for an average, representative case in a data sample in a manner similar to that used to extract marginal effects in the regression framework. In a seminal article, Friedman (2001) developed partial dependence plots (PDPs), displaying average predicted non-linear effects between the predicted response and one or more predictors; PDPs were recently succeeded by unbiased accumulated local effects (ALE) plots that relax the assumption of independence of predictors. Schulz *et al*. (2022) employed RF models in addition to traditional linear random intercept multilevel models to investigate the relationship between intra-workplace pay inequality and employee trust in managers. While the random intercept model with a polynomial term did not yield any significant results, PDPs indicated a clear and theoretically relevant average non-linear, inverse U-shaped relationship. Earlier work by Somers and Casal (2009) deployed neural networks to reveal the non-linear relationship between job satisfaction and job performance missed by traditional ordinary least squares regression.

Global interpretive techniques can identify conceptually relevant inflection points. However, they might mask heterogeneous effects in a data sample. Local methods of interpretive learning address this. Goldstein *et al*. (2015) developed individual conditional expectation (ICE) plots as a refinement to PDPs; these graphically visualize the partial relationship for each individual observation in the data. By combining PDPs and ICE plots, researchers can uncover variable, multidirectional effects in the data (see Schulz *et al*., 2022). More recently, ALE plots have been developed as an unbiased alternative to PDPs.

Developments in local interpretable learning include model-agnostic explanation, which stick closer to the algorithmic black box and provide interpretable solutions that maintain flexibility in the choice of models (Ribeiro *et al*., 2016). Examples of such solutions include local surrogate models (a local interpretable model-agnostic explanation, or LIME, is an example of black box predictions expressed in a linear form for every case in a dataset) and SHapley Additive exPlanations (SHAP), where game theory is leveraged to explain the outputs of ML (see Slack *et al*., 2020 for more recent derivatives of these models). SHAP is a particularly useful approach, as apart from a local explanation, where each observation is given a unique set of values describing the effects pertaining to it, it can be extended to a global case to show the contribution of each predictor to the outcome. SHAP values are directly extracted from tree-based models (e.g., RF models) rather than by narrowing the model down to a linear equivalent (Lundberg & Lee, 2017).

Overall, then, the distinction between global and local interpretation, as well as the ability of ML to detect non-linear, non-monotonous patterns, can impact theory development and the testing of mid-range theories in heterogeneous data with different layers of complexity (Leavitt *et al*., 2021). It is worth emphasising that the notion of global and local fit is present in statistical modelling too, although mostly in relation to estimates of how an

a priori theoretical model fits observational data. The meaning of global and local interpretation in ML learning is different, as shown above. The same concerns detection on non-linear patterns which are, of course, available in canonical statistical models. However, ML is generally superior where non-linear patterns do not obey strict mathematical functions (e.g., a quadratic term). These aspects of ML can contribute to empirical and theory-driven studies, as we demonstrate next in the worked example.

Worked example

Data and method

To demonstrate the innovative potential of ML, we used publicly available data from Eurofound's 2019 Company Survey – a management questionnaire with 21,869 observations from 28 European countries – to look at a strategically important phenomenon in management: product and service innovation. This technical exercise aims to show what the basic application of ML can achieve.

The outcome variable was coded as a dummy that took the value '1' if the firm had engaged in product and service innovation in the past three years, and '0' otherwise. As with many quantitative studies in management scholarship, we included a range of contextual variables as predictors in the model: engagement in other types of innovation (process and marketing), engagement in e-commerce, use of customized applications, use of data analytics, use of robots, company size, industry sector, type of establishment, change in management, employee task autonomy, pay determination, levels of hierarchy, use of collective labour agreements, and degree of market competition (further description of the variables can be found in the online Appendix, Table 2A). After omitting missing variables, the final sample consisted of 17,208 observations.

Given the binary categorical nature of the outcome variable – innovative versus non-innovative – the standard choice of a canonical statistical model is binomial logistic regression, where $\text{logit}(\pi_i)$, the log-odds of the underlying probability, is a linear function of the predictors: $\text{logit}(\pi_i) = x'_i\beta$. The non-parametric black box counterpart to logistic regression is a classification supervised learning algorithm that can take the form of ensemble methods (e.g., RF, gradient boosting), deep learning (e.g., artificial neural network), and other algorithms (e.g., support vector machine). The choice of the optimal algorithm is itself a learning process. By way of example, we demonstrate the performance of RF. Following the process of ML outlined in Figure 2, we divided the data into training data (70 per cent, i.e. 12,044 observations) and test data (30 per cent, i.e. 5164 observations) and deployed tenfold cross-validation. We let the algorithm learn patterns using the training data, applied it to the test data, and thereafter fine-tuned the algorithm to improve predictive accuracy. We did so with logistic regression first, to demonstrate how a model commonly used in canonical statistical modelling can be applied in line with ML principles. We then deployed a 'black box' algorithm (RF) and used 524 global and local methods of interpretable learning.

We now present the outcomes of logistic regression next to RF, beginning with traditional reported statistics for the logistic regression model and followed by the assessment of predictive accuracy and interpretation of empirical results (complete outputs are reported in online Appendix 2). For the latter, we zoom in on one predictor: the level of organizational hierarchy (indicating the number of levels in the internal organizational hierarchy).

Traditional model fit indices

The most commonly used model fit indices for logistic regression models are log-likelihood ratios, the Akaike information criterion, and pseudo-R^2. All three indices focus on how well the model fits the data at hand. The former two are relative performance measures that compare the fit of different regression models to establish the superiority of one model over another in terms of unexplained observations. The latter states the level of variation in the outcome variable, here product and service innovation, that is explained by the underlying regression model. Values for all three indices for the complete model are shown in Table 2 below. Yet, none of these indices suggest how accurately the model predicts product and service innovation in the current data or how well the model performs when being confronted with unseen data.

Predictive accuracy

We report predictive accuracy using confusion matrices. A confusion matrix visualizes the predictive

Table 2. Traditional model fit indices for logistic regression models

	Values
Pseudo-R2	0.264
Log-likelihood	–8340.13
Aike information criterion	16,746.25

Table 3. Confusion matrix: Logit model (full data set)

	Predicted '0'	Predicted '1'	Total	Error rate
Actual '0'	9448	1404	10,852	0.129
Actual '1'	2476	3880	6356	0.390
Total	11,924	5284	17,208	0.226

Table 4. Confusion matrix: Tuned random forest model (test data set)

	Predicted '0'	Predicted '1'	Total	Error rate
Actual '0'	2717	487	3204	0.152
Actual '1'	556	1404	1960	0.284
Total	3273	1891	5164	0.202

performance of a model by comparing the number of actual cases in a class (rows) with the number of predicted cases in the class (columns). The error rate refers to the percentage of cases not predicted accurately in each class. Table 3 corresponds to the logit model (model assumption tests can be found in online Appendix 2). The binominal logistics model is quite reliable at predicting if companies are non-innovative (error rate 0.13), but has a three times higher error rate at predicting innovative establishments (error rate 0.39). This is problematic, because we are naturally more interested in the latter. We also performed cross-validation on the logistic regression model, which yielded similar error rates and increased the robustness and generalizability of the results (see Appendix 2).

We now turn to the outcomes of hyper-tuned RF. While the overall error rate is similar to that of the logit model, the error rate in predicting innovative firms, the aspect we are interested in, has decreased to 0.28 (see Table 4). This amounts to an average 15 per cent reduction in the error rate in predicting innovative firms accurately when taking the logit model as a benchmark.

Figure 4 shows the variable importance score for each predictor. 'MeanDecreaseAccuracy' indicates the average loss in the model's accuracy in predicting the outcome variable if the predictor is omitted from the model. For example, removing 'Process innovation' from the model would result in an increased misclassification of more than 250 cases.

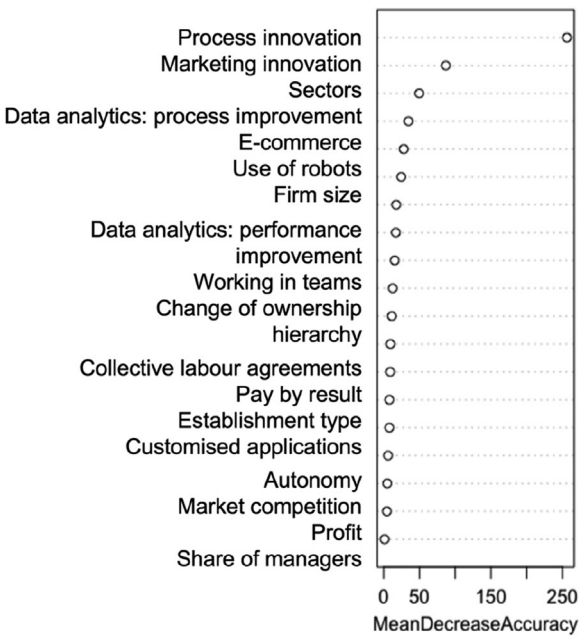


Figure 4. Variable importance scores

To show how ML can identify non-linear, non-monotonous relationships, Figures 5 and 6 visualize the relationship between the levels of hierarchy and product and service innovation. The PDP (Figure 5) suggests a non-linear, approximately inverse U-shaped relationship that would be undetected in logistic regression (see marginal effects

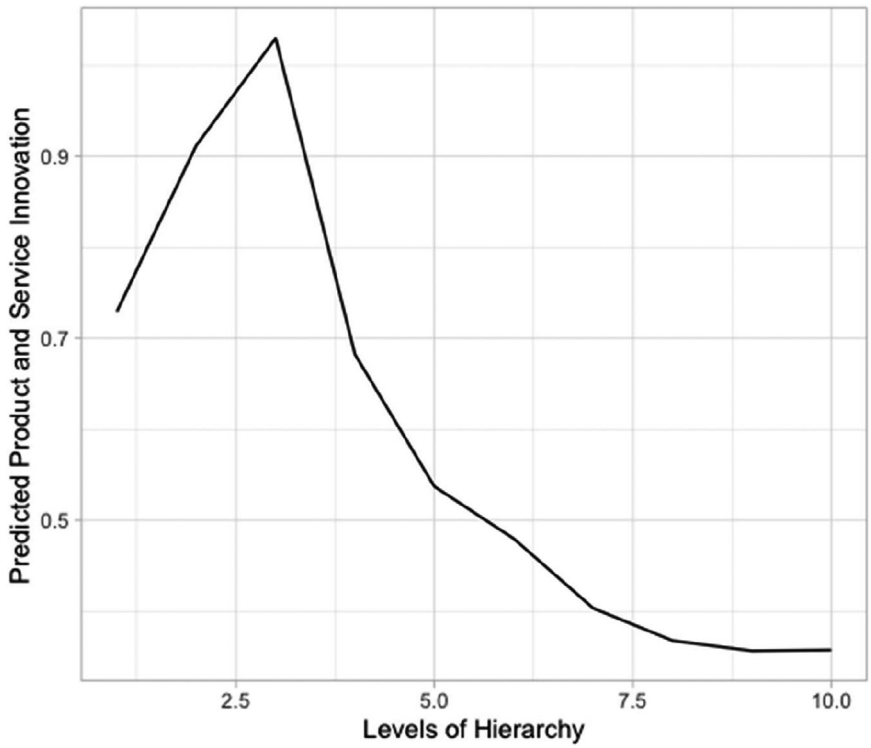


Figure 5. Partial dependence plot

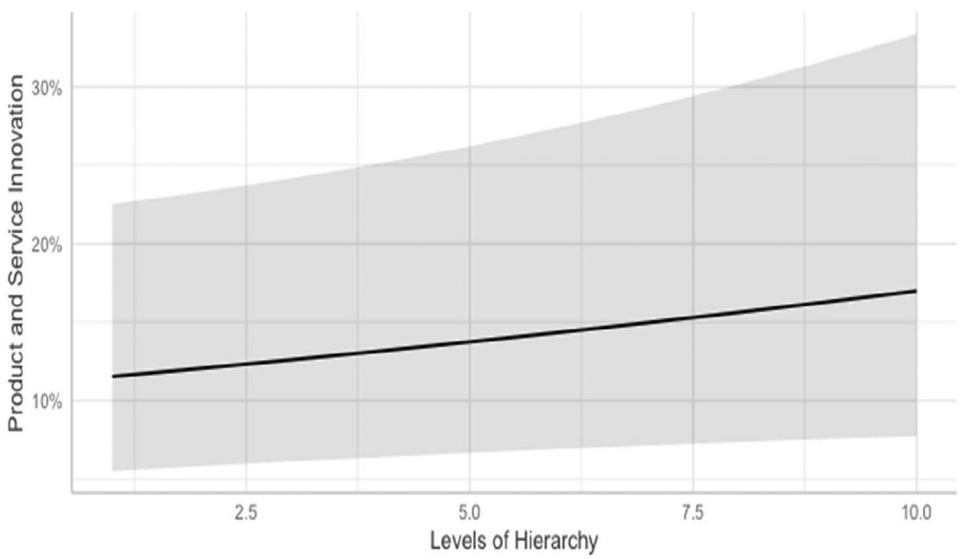
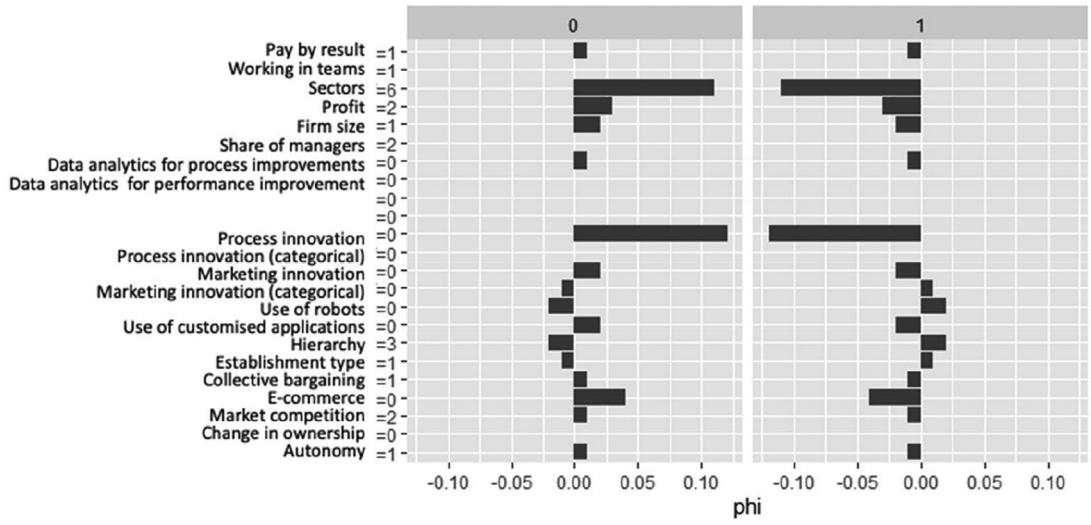


Figure 6. Marginal effects plot

derived by logistic regression in Figure 6). Based on the findings from the RF model, we tested for non-linearity in the logistic regression model using a quadratic term. The regression coefficients were non-significant, that is, the found non-linearity was not detected even after testing for it (see the output table in online Appendix, Table 2D).

A further step in the analysis can be the use of local agnostic explanations to estimate the effects of predictors across individual respondents



Note: top figure for Construction and bottom figure for Mining and Quarrying sector

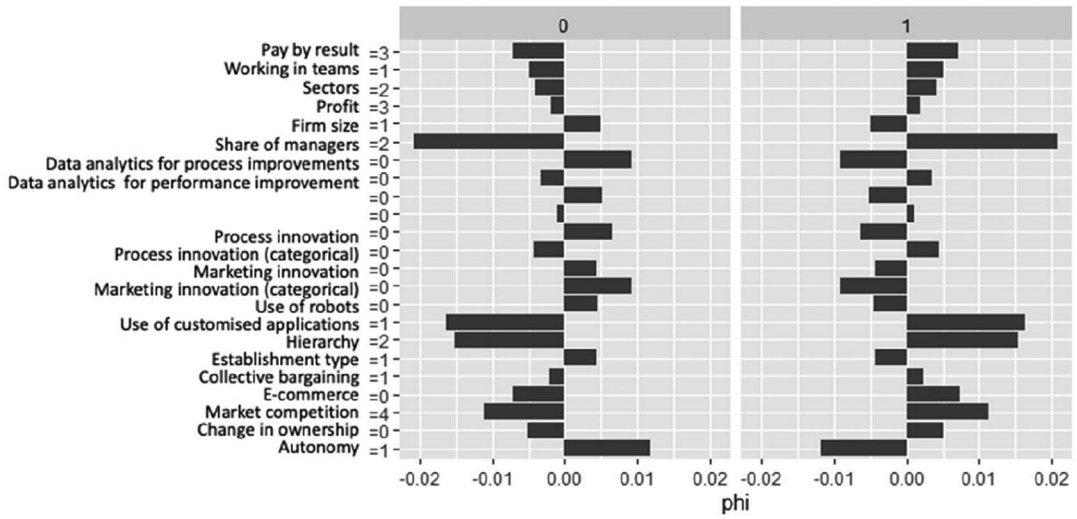


Figure 7. SHapley Additive exPlanations (SHAP) analysis
Note: top figure for Construction and bottom figure for Mining and Quarrying sector

or different subgroups in the data. In this example, we demonstrate the application of one such technique: SHAP analysis. SHAP analysis leverages game theory to assess every possible combination of predictors to determine each predictor's effect on the outcome for every individual respondent or group of respondents. Figure 7 visualizes SHAP analysis for two different sectors in the data (as seen in the code for the 'Sectors' variable). The values on the graph quantify the extent to which a given variable explains the difference in predictions between the two sectors by affecting the likelihood of being less innovative (the left-hand side of the graph, coded '0') or more innovative (the right-hand side, coded '1'). For example, the variable capturing hierarchy considerably increases the likelihood of innovation only in the lower graph (firms concentrated in a specific sector). This is an algorithmic alternative to moderation analysis.

This worked example of predicting product innovation acts as a simple demonstration of the inner workings and capabilities of ML. In canonical statistical modelling, we would most likely have been constrained by a confirmatory theoretical hypothesis and would have used significance levels to assess the effects of our variable of interest on the outcome. ML is not contingent on such assumptions, while even when using logistic regression we paid attention to predictive accuracy and what it might mean for the theoretical interpretation of our findings. We further showed that a non-parametric ensemble algorithm was more accurate at predicting rare events, as demonstrated here by firms' product and service innovation. Interpretation and visualization tools, such as PDP and SHAP analysis, helped to uncover interesting patterns that withstood the test of cross-validation but were unnoticed in logistic regression.

Addressing the limitations of machine learning

While the predictive and forecasting capabilities of ML are recognized (Bennell *et al*., 2006; Sarstedt & Danks, 2021; Yarkoni & Westfall, 2017), it is important to address the key limitations of ML as a methodological paradigm. The fact that greater predictability comes with poorer interpretability is a limitation often attributed to ML, which relies on data-driven estimates. As we have pointed out throughout the paper and shown in the worked example, black box is a misnomer, and there are innovative methods of interpretable ML that can be useful for management scholars.

Where ML is met with scepticism, this is explained by the lack of causal inference underpinning algorithmic modelling. Pearl (2018, p. 2) critiqued the focus of ML on predictions: 'no matter how skilfully you manipulate the data and what you read into the data when you manipulate it, it's still a curve-fitting exercise, albeit complex and nontrivial.' While this critique has merit, an ML algorithm focused on predictive accuracy and generalizability beyond the sample is not, ceteris paribus, inferior to a statistical model premised on the search for unbiased confidence estimates. In general, statistical models are weak indicators of causality unless they explicitly assume an intervention and scrutinize counterfactual scenarios, as is the case with quasi-experimental methods (see Rizov *et al*., 2016, for an example of a difference-in-differences design). Correcting statistical estimates for omitted variable bias, unobserved heterogeneity, and hierarchical data structure (all common approaches in management studies) does not solve the problem of causality. In both paradigms – statistical modelling and ML – causality stems from the researcher's interpretation: ex ante in statistical modelling by specifying a conceptual model to be tested; ex post in ML by referring back to relevant concepts and theories to make sense of emerging findings.

Recent developments in ML include the extension to incorporate treatment effects as in a quasi-experimental setting (Wager & Athey, 2018; Zhu *et al*., 2019). These algorithms can assess heterogeneous treatment effects that vary across different groups of respondents (Athey & Imbens, 2015; Grimmer *et al*., 2017; Wager & Athey, 2018); this can be invaluable for testing prominent management theories, such as the contingency theory of human resource management (Boselie *et al*., 2005).

While we delineate the role of a researcher in ML, however, we do not contend that algorithms are neutral or value-free. Issues of algorithmic bias and related ethical implications are important considerations that cannot be ignored (Mittelstadt *et al*., 2016). Encoded language and autonomous decision-making algorithms may be oppressive or discriminatory towards marginalized groups (c.f. Schroeder, 2021). As the power of ML rests on the quality of data the algorithm is trained on (Saltelli and Funtowicz, 2014) and the set of decision rules can be impacted by the boundaries of human assumptions, robust theoretical and ethical frameworks are important in the development of ML applications (Akter *et al*., 2022).

There are additional ethical implications for researchers in terms of the consequences of implementing ML. Regarding the application of ML to business practice, as well as its use in research to derive managerial implications, the case for improving efficiency, bringing down costs, or streamlining operations needs to be balanced with a regard for wider business and societal agendas, whereby ML is not positioned as a universal, all-encompassing solution. With this in mind, we subscribe to the view that to think outside the black box of algorithmic modelling requires critical, causal thinking *and* ethics.

Concluding remarks

In this paper, we engaged with ML as a standalone methodological paradigm. In doing so, we aimed to articulate a viable approach that can address some of the most entrenched methodological tensions in management scholarship. As management scholars increasingly utilize learning algorithms, we felt the need to provide a more systematic account of the core tenets of ML and to outline their innovative potential for management scholarship.

We recognize that statistical modelling and hypothesis significance testing have been subject to ongoing criticism. Gigerenzer and Marewski (2015) have written perhaps the most compelling rebuttal of the seemingly unshakable devotion of management scholars to statistical models. We argue that ML can deliver methodological innovation and we sought to demonstrate this with the worked example. Having addressed some prominent limitations of ML, we argue for an incremental change, with the adoption of algorithmic learning where it clearly benefits the research agenda. A useful initial step for researchers would be to articulate the empirical nature of their research and make a case for using either conventional statistical models or algorithmic modelling, or to use the principles of ML alongside statistical models.

A piecemeal adoption of ML is unlikely to leverage its innovative potential; this requires wider social processes. First, the graduate curriculum will have to change to incorporate ML as a stand-alone approach to data analysis. It is our responsibility to expose graduate students to cutting-edge methodological innovations (see Easterby-Smith *et al.*, 2021). In the era of open-source software packages, applying ML requires only basic training (for an overview of packages and tutorials for both R and Python, see Tonidandel *et al.*, 2018). That will stimulate master's and doctoral students to utilize ML algorithms. Second, we urge empirical researchers to adopt the solutions recommended in this paper. Increasing the volume of journal submissions based on the principles of ML (alone or in conjunction with statistical modelling) is a critically important step. Lastly, we call on journal editors and reviewers to encourage authors to use ML where studies lend themselves to algorithmic modelling and to highlight ML's potential to lead management scholarship to an era of new theoretical and practical impacts.

References

Adner, R., P. Puranam and F. Zhu (2019). 'What is different about digital strategy? From quantitative to qualitative change', *Strategy Science*, **4**, pp. 253–261.

Amrhein, V., S. Greenland and B. McShane (2019). 'Scientists rise up against statistical significance', *Nature*, **567**, pp. 305–307.

Anderson, S. F. and S. E. Maxwell (2017). 'Addressing the 'Replication crisis': using original studies to design replication studies with appropriate statistical power', *Multivariate Behavioral Research*, **52**, pp. 305–324.

Andraszewicz, S., B. Scheibehenne, J. Rieskamp, R. Grasman, J. Verhagen and E.J. Wagenmakers (2015). 'An introduction to Bayesian hypothesis testing for management research', *Journal of Management*, **41**, pp. 521–543.

Athey, S. and G. W. Imbens (2015). 'Machine learning methods for estimating heterogenous causal effects', *stat*, **1050**, pp. 1–26.

Athey, S. and G. W. Imbens (2019). 'Machine learning methods that economists should know about', *Annual Review of Economics*, **11**, pp. 685–725.

Ban, G. Y., N. El Karoui and A. E. Lim (2018). 'Machine learning and portfolio optimization', *Management Science*, **64**, pp. 1136–1154.

Batistič, S. and P. Van Der Laken (2019). 'History, evolution and future of big data and analytics: a bibliometric analysis of its relationship to performance in organizations', *British Journal of Management*, **30**, pp. 229–251.

Beleites, C., R. Baumgartner, C. Bowman, R. Samorjai, G. Steiner, R. Salzer and M.G. Sowa (2005). 'Variance reduction in estimating classification error using sparse datasets', *Chemometrics and Intelligent Laboratory Systems*, **79**, pp. 91–100.

Bellstam, G., S. Bhagat and J. A. Cookson (2021). 'A text-based analysis of corporate innovation', *Management Science*, **67**, pp. 4004–4031.

Bennell, J. A., D. Crabbe, S. Thomas and O. Ap Gwilym (2006). 'Modelling sovereign credit ratings: neural networks versus ordered probit', *Expert Systems with Applications*, **30**, pp. 415–425.

Bergh, D. D., B. M. Sharp, H. Aguinis and M. Li (2017). 'Is there a credibility crisis in strategic management research? Evidence on the reproducibility of study findings', *Strategic Organization*, **15**, pp. 423–436.

Blockeel, H. and J. Struyf (2002). 'Efficient algorithms for decision tree cross-validation', *Journal of Machine Learning Research*, **3**, pp. 621–650.

Boselie, P., G. Dietz and C. Boon (2005). 'Commonalities and contradictions in HRM and performance research', *Human Resource Management Journal*, **15**, pp. 67–94.

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

Breiman, L. (2001b). 'Random forests', *Machine Learning*, **45**, pp. 5–32.

Choi, J., Menon, A. and Tabakovic, H. (2021). 'Using machine learning to revisit the diversification–performance relationship', *Strategic Management Journal*, **42**, pp. 1632–1661.

Choudhury, P., D. Wang, N. A. Carlson and T. Khanna (2019). 'Machine learning approaches to facial and text analysis:

Discovering CEO oral communication styles', *Strategic Management Journal*, **40**, pp. 1705–1732.
Choudhury, P., R. T. Allen and M.G. Endres (2021). 'Machine learning for pattern discovery in management research', *Strategic Management Journal*, **42**, pp. 30–57.
Bzdok, D., N. Altman and M. Krzywinski (2018). 'Statistics versus machine learning', *Nature Methods*, **15**, p. 233.
Christensen, G. and E. Miguel (2018). 'Transparency, reproducibility and the credibility of economics research', *Journal of Economic Literature*, **56**, pp. 920–980.
Cohen P.R. and D. Jensen (1997). 'Overfitting Explained'. In D. Madigan and P. Smyth (eds), *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, pp. 115–122. Fort Lauderdale: PMLR.
Cornwall, G., J. Chen and B. Sauley (2021). 'Standing on the shoulders of machine learning: Can we improve hypothesis testing?', arXiv preprint arXiv:2103.01368.
Doshi-Velez, F. and B. Kim (2017). 'Towards a rigorous science of interpretable machine learning', arXiv preprint arXiv:1702.08608.
Duvendack, M., R. Palmer-Jones and W. R. Reed (2017). 'What is meant by 'replication' and why does it encounter resistance in economics?', *American Economic Review*, **107**, pp. 46–51.
Easterby-Smith, M., L. J. Jaspersen, R. Thorpe and D. Valizade (2021). *Management and Business Research*. London: Sage Publications.
ElShawi, R., Y. Sherif, M. Al-Mallah and S. Sakr (2020). 'Interpretability in healthcare: a comparative study of local machine learning interpretability techniques', *Computational Intelligence*, **37**, pp. 1633–1650.
Ethiraj, S. K., A. Gambardella and C.E. Helfat (2016). 'Replication in strategic management', *Strategic Management Journal*, **37**, pp. 2191–2192.
Friedman, J. H. (2001). 'Greedy function approximation: A gradient boosting machine', *The Annals of Statistics*, **29**, pp. 1189–1232.
Friedman, J. H. (2006). 'Recent advances in predictive (machine) learning', *Journal of Classification*, **23**, pp. 175–197.
Friedman, M. (1953). *The Methodology of Positive Economics*. Chicago, IL: University of Chicago Press.
Garg, S., S. Sinha, A. K. Kar and M. Mani (2021). 'A review of machine learning applications in human resource management', *International Journal of Productivity and Performance Management*, **71**, pp. 1590–1610.
Gentner, D., B. Stelzer, B. Ramosaj and L. Brecht (2018). 'Strategic foresight of future b2b customer opportunities through machine learning', *Technology Innovation Management Review*, **8**, pp. 5–17.
Gigerenzer, G. and J. N. Marewski (2015). 'Surrogate science: The idol of a universal method for scientific inference', *Journal of Management*, **41**, pp. 421–440.
Goldfarb, B. and A. A. King (2016). 'Scientific apophenia in strategic management research: significance tests & mistaken inference', *Strategic Management Journal*, **37**, pp. 167–176.
Goldstein, A., A. Kapelner, J. Bleich and E. Pitkin (2015). 'Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation', *Journal of Computational and Graphical Statistics*, **24**, pp. 44–65.
Google Cloud. (2022, April 14). 'IKEA Retail (Ingka Group) increases Global Average Order Value for eCommerce by 2% with Recommendations AI'. Retrieved from: https://cloud.google.com/blog/products/ai-machine-learning/ikea-uses-google-cloud-recommendations-ai.
Grimmer, J. (2015). 'We are all social scientists now: how big data, machine learning, and causal inference work together', *PS: Political Science & Politics*, **48**, pp. 80–83.
Grimmer J., S. Messing and S. J. Westwood (2017). 'Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods', *Political Analysis*, **25**, pp. 414–434.
Hawkins, D. M. (2004). 'The problem of overfitting', *Journal of Chemical Information and Computer Sciences.*, **44**, pp. 1–12.
Hofman, J. M., D. J. Watts, S. Athey, F. Garip, T. L. Griffiths, J. Kleinberg, H. Margetts, S. Mullainathan, M. J. Salganik, S. Vazire and A. Vespignani (2020). 'Integrating explanation and prediction in computational social science', *Nature*, **595**, pp. 181–188.
Husain, H (2021, April 14). 'Discovery enhances customer experience using Amazon Personalize'. Retrieved from: https://aws.amazon.com/solutions/case-studies/discovery-inc-case-study/.
Hwang, S., Kim, J., Park, E. and Kwon, S. J. (2020). 'Who will be your next customer: a machine learning approach to customer return visits in airline services', *Journal of Business Research*, **121**, pp. 121–126.
IBM. (2022, April 14). 'Boots UK significantly lifting incremental spend with tailored promotions for loyalty card customers'. Retrieved from: https://www.ibm.com/case-studies/boots-uk.
James, G., D. Witten, T. Hastie and R. Tibshirani (2017). *An Introduction to Statistical Learning – with Applications in R*. New York, NY: Springer.
Kellard, N. M. and M. Śliwa (2016). 'Business and management impact assessment in research excellence framework 2014: analysis and reflection', *British Journal of Management*, **27**, pp. 693–711.
Killeen, P. R. (2005). 'An alternative to null-hypothesis significance tests', *Psychological Science*, **16**, pp. 345–353.
Kim, J. H. (2009). 'Estimating classification error rate; repeated cross validation, repeated hold-out bootstrap', *Computational Statistics and Data Analysis*, **53**, pp. 3735–3745.
Leavitt, K., K. Schabram, P. Hariharan and C. M. Barnes (2021). 'Ghost in the machine: On organizational theory in the age of machine learning', *Academy of Management Review*, **46**, pp. 750–777.
Lundberg, S. M. and S. I. Lee (2017). 'A unified approach to interpreting model predictions'. U. von Luxburg, I. Guyon, S. Bengio, H. Wallach and R. Fergus (eds), *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768–4777. Long Beach: NIPS.
McFarland, D. A., K. Lewis and A. Goldberg (2016). 'Sociology in the era of big data: the ascent of forensic social science', *The American Sociologist*, **47**, pp. 12–35.
Mingers, J. (2006). 'A critique of statistical modelling in management science from a critical realist perspective: its role within multimethodology', *Journal of the Operational Research Society*, **57**, pp. 202–219.
Molina, M. and F. Garip (2019). 'Machine learning for sociology', *Annual Review of Sociology*, **45**, pp. 27–45.
Moln-Page, K (2021, April 14). 'ASOS implements cost optimization to fashion innovation for the future'. Retrieved from: https://customers.microsoft.com/en-us/story/1375958406314056214-asos-retailers-azure.
Muchlinski, D., D. Siroky, J. He and M. Koche (2016). 'Comparing random forest with logistic regression for predicting

class-imbalanced civil war onset data', *Political Analysis*, **24**, pp. 87–103.

Nelson, L. D., J. Simmons and U. Simonsohn (2018). 'Psychology's renaissance', *Annual Review of Psychology*, **69**, 17.1–17.24.

Open Science Collaboration. (2015). 'Estimating the reproducibility in psychological science', *Science*, **349**, aac4716.

Pagell, M. (2020). 'Replication without repeating ourselves: addressing the replication crisis in operations and supply chain management research', *Journal of Operational Management*, **67**, pp. 105–115.

Pampouktsi, P., S. Avdimiotis, M. Maragoudakis and M. Avlonitis (2021). 'Applied machine learning techniques on selection and positioning of human resources in the public sector'. *Open Journal of Business and Management*, **9**, pp. 536–556.

Pearl, J. (2018). 'To build truly intelligent machines, teach them cause and effect', *Quantamagzine*. Retrieved from: https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/#

Perezgonzalez, J. D. (2015). 'Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing', *Frontiers in Psychology*, **6**, p. 223.

Ribeiro, M. T., S. Singh and C. Guestrin (2016). 'Model-agnostic interpretability of machine learning', arXiv preprint arXiv:1606.05386.

Rizov, M., R. Croucher and T. Lange (2016). 'The UK national minimum wage's impact on productivity', *British Journal of Management*, **27**, pp. 819–835.

Sajjadiani, S., A. J. Sojourner, J. D. Kammeyer-Mueller and E. Mykerezi (2019). 'Using machine learning to translate applicant work history into predictors of performance and turnover', *Journal of Applied Psychology*, **104**, p. 1207.

Sarstedt, M. and N. P. Danks (2022). 'Prediction in HRM research – A gap between rhetoric and reality', *Human Resource Management Journal*, **32**, pp. 485–513.

Scandura, T. A. and E.A. Williams (2000). 'Research methodology in management: current practices, trends, and implications for future research', *Academy of Management Journal*, **43**, pp. 1248–1264.

Schmitt, B., J. J. Brakus and A. Biraglia (2022). 'Consumption ideology', *Journal of Consumer Research*, **49**, pp. 74–95.

Schulz, F., D. Valizade and A. Charlwood (2022). 'The effect of intra-workplace pay inequality on employee trust in managers: assessing a multilevel moderated mediation effect model', *Human Relations*, **75**, pp. 705–733.

Sharpe, D. and S. Poets (2020). 'Meta-analysis as a response to the replication crisis', *Canadian Psychology/Psychologie Canadienne*, **61**, pp. 377–387.

Shrestha, Y. R., V. F. He, P. Puranam and G. von Krogh (2021). 'Algorithm supported induction for building theory: How can we use prediction models to theorize?', *Organization Science*, **32**, pp. 856–880.

Simmons, J. P., Nelson, L. D. and U. Simonsohn (2011). 'False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant', *Psychological Science*, **22**, pp. 1359–1366.

Slack, D., S. Hilgard, E. Jia, S. Singh and H. Lakkaraju (2020). 'Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods'. A. Markham, J. Powles, T. Walsh, A. Washington (eds), *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186. New York: AIES.

Somers, M. J. and J. C. Casal (2009). 'Using artificial neural networks to model nonlinearity: the case of the job satisfaction-job performance relationship', *Organizational Research Methods*, **12**, pp. 403–417.

Stoltzfus, J. C. (2011). 'Logistic regression: a brief primer', *Academic Emergency Medical Journal, Research Methods and Statistics*, **18**, pp. 1099–1104.

Svetnik, V., A. Liaw, C. Tong, C. Culberson, R. P. Sheridan and B. P. Feuston (2003). 'Random forest: a classification and regression tool for compound classification and QSAR modeling', *Journal of Chemical Information and Computer Sciences*, **43**, pp. 1947–1958.

Swan, A. (2022, April 14). 'Speeding customer service delivery. BT streamlines its contract process to improve experiences for customers and employees with Adobe Sign'. Retrieved from: http://business.adobe.com/ro/customer-success-stories/bt-sign-case-study.html.

Szucs, D. and J. Ioannidis (2017). 'When null hypothesis significance testing is unsuitable for research: a reassessment', *Frontiers in Human Neuroscience*, **11**, p. 390.

Tidhar, R. and K. M. Eisenhardt (2020). 'Get rich or die trying… finding revenue model fit using machine learning and multiple cases', *Strategic Management Journal*, **41**, pp. 1245–1273.

Tonidandel, S., E. B. King and J. M. Cortina (2018). 'Big data methods: leveraging modern data analytic techniques to build organizational science', *Organizational Research Methods*, **21**, pp. 525–547.

Wager, S. and S. Athey (2018). 'Estimation and inference of heterogeneous treatment effects using random forests', *Journal of the American Statistical Association*, **113**, pp. 1228–1242.

Wang, X. and J. Zhi (2021). 'A machine learning-based analytical framework for employee turnover prediction', *Journal of Management Analytics*, **8**, pp. 351–370.

Wicks, A.C. and Freeman, R.E. (1998). 'Organization studies and the new pragmatism: positivism, anti-positivism, and the search for ethics', *Organization Science*, **9**, pp. 123–140.

Yarkoni, T. and J. Westfall (2017). 'Choosing prediction over explanation in psychology: lessons from machine learning', *Perspectives on Psychological Science*, **12**, pp. 1100–1122.

Zhu, S., I. Ng and Z. Chen (2019). 'Causal discovery with reinforcement learning', arXiv preprint arXiv:1906.04477.

Ziliak, S. and D. N. McCloskey (2008). *The Cult of Statistical Significance: How the Standard Error Costs us Jobs, Justice, and Lives*. Ann Arbor, MI: University of Michigan Press.

Danat Valizade is an Associate Professor in Quantitative Methods at Leeds University Business School. With a background in economics and nearly a decade-long career in trade unions, his research interests coalesce around the changing nature of work with a specific focus on labour market polarisation, and disparities in the quality of work and their effect on performance and wellbeing. He is particularly interested in the application of data science, an interdisciplinary field with the use of advanced statistical techniques and machine learning algorithms to draw meaningful insights from raw data at its heart. As data-driven research takes hold, Danat's research fosters a better understanding of the causal mechanisms underpinning contemporary tendencies in work and employment.

Felix Schulz is a Post-Doctoral Research Fellow at the Digital Futures at Work Research Centre (digit) and the Work and Employment Relations Division at the University of Leeds. In his research, Felix investigates the intersections of work, pay inequality, and environmental sustainability, focusing on, amongst other things, the role of digital technologies. Theoretically his research is interdisciplinary, drawing on sociology, social psychology, and labour economics.

Cezara Nicoara is a Lecturer in Marketing at Newcastle University Business School. She obtained a PhD from Leeds University Business School. Cezara is an expert in quantitative marketing research. Her research interests focus on ethics in marketing and corporate social responsibility in multinational enterprise subsidiaries.

Supporting Information

Additional supporting information can be found online in the Supporting Information section at the end of the article.