Dudda, Tom L.; Hornuf, Lars

Working Paper

The Perks and Perils of Machine Learning in Business and Economic Research

CESifo Working Paper, No. 11721

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Dudda, Tom L.; Hornuf, Lars (2025) : The Perks and Perils of Machine Learning in Business and Economic Research, CESifo Working Paper, No. 11721, CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/314760

CESifo WORKING PAPERS

11721
2025

February 2025

The Perks and Perils of Machine Learning in Business and Economic Research

Tom L. Dudda, Lars Hornuf

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 11721*

The Perks and Perils of Machine Learning in Business and Economic Research

Abstract

We examine predictive machine learning studies from 50 top business and economic journals published between 2010 and 2023. We investigate their transparency regarding the predictive performance of machine learning models compared to less complex traditional statistical models that require fewer resources in terms of time and energy. We find that the adoption of machine learning varies by discipline, and is most frequently used in information systems, marketing, and operations research journals. Our analysis also reveals that 28% of studies do not benchmark the predictive performance of machine learning models against traditional statistical models. These studies receive fewer citations, arguably due to a less rigorous analysis. Studies including traditional statistical models as benchmarks typically report high outperformance for the best machine learning model. However, the performance improvement is substantially lower for the average reported machine learning model. We contend that, due to opaque reporting practices, it often remains unclear whether the predictive gains justify the increased costs of more complex models. We advocate for standardized, transparent model reporting that relates predictive gains to the efficiency of machine learning models compared to less-costly traditional statistical models.

JEL-Codes: C180, C400, C520.

Keywords: machine learning, predictive modelling, transparent model reporting.

Tom L. Dudda
Faculty of Business and Economics
Dresden University of Technology / Germany
tom_lukas.dudda@tu-dresden.de

Lars Hornuf
Faculty of Business and Economics
Dresden University of Technology / Germany
lars.hornuf@tu-dresden.de

This version: February 26, 2025
We are grateful to Dietmar Harhoff, Matthias Hanauer, Paul Hünermund, Kai Heinrich, Bernhard Lutz, Thomas Walther, Patrick Zschech, and participants of the Causal Data Science Meeting 2023, the DRUID Academy 2024, and research seminars at Katholische Universität Eichstätt-Ingolstadt, KU Leuven, Utrecht University, and Maastricht University for their valuable suggestions and comments. We are also grateful to the respondents of our survey for their participation and valuable comments. We thank Sofian Bekhtaoui, Jannik Holfert, Hannes Köhler, and Arnes Triemer for their excellent research assistance.

1. Introduction

Because of its non-linear and often more complex nature, machine learning (ML) offers significant theoretical advantages over traditional statistical models in addressing prediction problems in business and economics (Athey, 2018; Athey and Imbens, 2019; Bzdok et al., 2018; Goldstein et al., 2021; Mullainathan and Spiess, 2017; Varian, 2014). These advantages become particularly relevant if a meaningful prediction for an economic problem requires many potentially interacting variables that are non-linearly related to the outcome variable (Goldstein et al., 2021). ML models might uncover hidden non-linear relationships between variables that traditional statistical models would overlook (Choudhury et al., 2020). Additionally, ML methods are well-suited for working with "big data"[1] and effectively processing the high-dimensional data present in large volumes of unstructured text or images (Athey, 2018; Gentzkow et al., 2019; Goldstein et al., 2021; Mullainathan and Spiess, 2017; Varian, 2014). It is therefore not surprising that researchers increasingly apply ML models to address research questions in business and economics, as Figure 1 shows. An analysis by Currie et al. (2020) corroborates this, showing that mentions of ML in microeconomic articles have increased exponentially since 2010, making ML the most discussed new method in the field.

[Figure 1 about here]

At the same time, many problems are in reality rather simple and sometimes linear in nature (Athey and Imbens, 2019). Using ML models to approach these problems can be inefficient given the higher costs associated with ML models in terms of time and energy consumption compared to traditional regression models. Complex ML models require longer training periods on a higher number of GPUs and a larger amount of data. The use of additional computer resources results in higher financial costs in the form of hardware, electricity, and computing time in the cloud, as well as environmental costs in the form of CO2 emissions, all of which might economically outweigh marginal gains in predictive accuracy (see, e.g., Bender et al., 2021; Ebert et al., 2024; Lacoste et al., 2019; Luccioni et al., 2024; Schwartz et al., 2019; Strubell et al., 2019, 2020; Thompson et al., 2021).[2] Further ML model costs arise from the lower explainability and interpretability of the results compared to those of traditional econometric methods (Hünermund et al., 2022; Messeri and Crockett, 2024).

[1] Goldstein et al. (2021) provide a definition of the term "big data" in the context of finance research that translates well to other business and economic disciplines. They describe three characteristics of "big data": the pure size of the data set, high dimensionality (i.e., a high number of variables compared to the sample size), and complexity regarding the data structure. The latter refers to unstructured data sets, such as text, image, video, or audio data.
[2] Training one common natural language processing model on a GPU with parameter tuning and experiments is estimated to account for more than twice as much carbon emissions than the emissions of an average U.S. citizen per year. Training one big transformer model on a GPU, including neural architecture search, is estimated to emit more than 17 times as much carbon as the average U.S. citizen per year (Strubell et al., 2019).

Given the rapidly rising number of ML-related publications in business and economic research in recent years, we hypothesize that published articles in these disciplines increasingly employ ML models for predictive research problems that often do not have the content complexity that requires researchers to rely on more resource-intensive models. In such cases, these models will frequently perform only marginally better than conventional, less time- and energy-consuming methods. Researchers might refrain from reporting the predictive performance of less complex conventional methods if they yield similar results to ML models, given that marginal improvements might not justify the higher financial and environmental costs of using more complex models. Using novel ML methods might also attract more attention, leading to higher chances of publication and a higher number of expected citations (Leech et al., 2024).

This is likely the first explorative study investigating the use of ML across the 50 high-quality business and economic journals comprising the Financial Times Research Rank (FT50). Specifically, we focus on articles that apply ML models to solve predictive research problems. We investigate whether these articles are transparent about the performance improvement of using more costly ML models compared to less complex conventional statistical methods, such as linear or logistic regression. Provided that the studies report comparable results for both ML and established traditional statistical models, we compare the reported predictive performance of both model types. Finally, we examine whether the transparency about and the extent of the relative performance improvement through ML models is associated with the impact that an article generates, as measured by its citation count.

Results. Out of 56,262 articles published between 2010 and 2023 in journals of the FT50, we manually identified 1,211 articles that involve ML. The annual number of published articles applying ML models has increased significantly over recent years but differs between research disciplines. ML models are most frequently covered in information systems, marketing, and operations research journals. Measured as the share of the total number of publications, we identified considerably fewer ML-related studies in human relations, organization studies, economics, and accounting.

The sample for our main analysis consists of 203 studies that apply at least one ML model to predict a variable that is of central interest for answering one of the main research questions of the article. We find that more than a quarter of articles do not benchmark the predictive performance of the employed ML models against traditional statistical models. Neglecting to report comparable results for traditional statistical models makes it difficult to assess the true economic value of using a more complex and resource-intensive model. We also encounter substantial differences between research disciplines. For example, 94% of the articles published in finance journals report the performance of traditional statistical models, whereas only 69% in information systems and 62% in marketing journals do so. Studies that base their prediction on text, image, or video data are, on average, 13.5% less likely to report results for traditional statistical models. If we exclude studies using these often highly unstructured data sets, 22% of the articles do not contain results for traditional statistical models. We find no

evidence that the seniority of the authors or the size of the author team affects transparency about the performance of traditional statistical models.

Studies reporting results for traditional statistical models predominantly state a strong outperformance of the best-performing ML model over the best-performing traditional benchmark. However, the outperformance is, on average, reduced by 65%, often even turning negative, when we compare the average performance of all reported ML models in a study against the best traditional statistical model. The magnitude of the reported ML outperformance can be explained by the number of reported ML models and traditional statistical models. While the performance improvement of the best ML model over the best traditional statistical model increases with the number of reported ML models, the difference between the average ML model and the best traditional statistical model decreases with the number of reported traditional statistical models. This suggests that the effort of scholars to find an ML or traditional statistical model that is well-suited to their research problem will affect the reported performance difference between the two model types.[3] On average, larger author teams report noticeably more ML models.

We further find that the authors' total citation count prior to publication is positively related to the reported performance of ML models relative to traditional statistical models. When we measure the authors' seniority by the number of previous FT50 publications or the number of years since they obtained their PhD, we find a negative correlation between seniority and the reported outperformance of ML models. Overall, our results indicate that beating a well-established traditional statistical model in business and economics with ML might often require substantial effort in finding and training a powerful model, while the outperformance of the average ML model is comparably low.

Lastly, we find that the transparency of published articles about the relative performance of ML models compared to traditional statistical models is positively related to their citation count. Studies that report results for traditional benchmarks receive, on average, 2.7 to 6.8 more citations per year than studies that do not report such benchmark results. This effect is sizeable considering the 8.1 citations per year that the average article published in the journals of our sample garnered between 2018 and 2023. We argue that studies transparently reporting benchmark results are methodically more rigorous, which might also be an indicator of the general quality of the study, which ultimately results in more citations. We also identify other factors that positively correlate with citations, such as the authors' general ability to produce impactful research, or the use of innovative data sets, including texts and images. However, we find no convincing evidence that the reported performance improvement of ML over traditional statistical models is associated with an article's number of citations.

With regard to the increasing adoption of more complex ML models in applied empirical research, we expand on the claims of Hofman et al. (2017) supporting standardized reporting of predictive

[3] This relates to Menkveld et al. (2024), showing that the evidence-generating process chosen by researchers affects the outcome of a study and thereby adds uncertainty to the results.

performance in social sciences. Hofman et al. (2017) argue in favor of consistent evaluation of predictive performance, including comparison to the best-known models and clarification of the modeling choices researchers have made to arrive at their results. We add to their discussion and argue that the predictive performance of competing models should be evaluated and reported relative to model costs. A relative assessment of predictive performance taking model costs into account is particularly important in light of resource-intensive deep ML models competing with less costly traditional statistical models.

Indeed, the general ML literature points towards the importance of computational and energy efficiency alongside predictive performance. As criticized by García-Martín et al. (2019), Schwartz et al. (2019), and Strubell et al. (2020), ML researchers primarily aim to increase the accuracy of prediction models, often without imposing constraints on computation power or energy consumption. Strubell et al. (2020) stress that ML researchers should consider the accuracy of models in relation to their efficiency to keep track of their carbon footprint. Bender et al. (2021) suggest putting more focus on environmental and financial costs, particularly for large language models. Schwartz et al. (2019) and Dodge et al. (2019) discuss measures of model efficiency. A standardized, transparent reporting of relative model performance and efficiency enables fellow researchers and practitioners to understand the actual (i.e., not only statistical but also economical) comparative value of a more complex ML model in addressing a given research problem. If a significant advantage is evident, authors might also benefit from such reporting standards through higher visibility and impact of their research, as our results suggest.

Related Literature. Our study is directly related to the literature discussing and reviewing the use and applications of ML in business and economic research. These studies usually focus on specific disciplines, such as economics (Athey, 2018; Athey and Imbens, 2019; Mullainathan and Spiess, 2017), entrepreneurship (Lévesque et al., 2020), finance (Goldstein et al., 2021; Kelly and Xiu, 2023), organization research (Leavitt et al., 2021), operations management (Chou et al., 2023; Kraus et al., 2020), strategy and management research in general (Choudhury et al., 2020), or ML in the context of electronic markets (Janiesch et al., 2021; Bawack et al., 2022). The authors of these studies evaluate the importance, merits, limitations, and pitfalls of ML in the respective research discipline and discuss specific methods they consider particularly promising. For example, Mullainathan and Spiess (2017) warn against naively employing ML methods in research just because they are easy to implement with ready-to-use programming packages. Some articles also review existing ML studies such as in financial markets (Kelly and Xiu, 2023), operations management (Chou et al., 2023), electronic commerce (Bawack et al., 2022), or predictive healthcare (Heinrich and Keshavarzi, 2024). Often, the role of ML in the evidence-generating process beyond solving main predictive research problems is discussed, for example for theory building (Choudhury et al., 2020; Chou et al., 2023; Leavitt et al., 2021). Many studies emphasize that research can benefit if ML and traditional statistical models are used in combination (e.g., Athey, 2018; Choudhury et al., 2020; Leavitt et al., 2021). It does not always have to be an either/or decision but can be a question of when and for what tasks a method should be used at different

stages of a research project. For example, Choudhury et al. (2020) propose that ML can be applied to derive new hypotheses from data by capturing previously hidden patterns and then testing them using traditional statistical models that focus on statistical inference. However, this is not something we address in this article. Rather, we focus on studies in which ML and traditional statistical models compete to solve predictive research problems.[4]

Our paper is also connected to studies detecting methodological pitfalls of applying ML models to predictive research questions in other disciplines, such as biomedical science (Andaur Navarro et al., 2021), neuroscience (Arbabshirani et al., 2017; Rosenblatt et al., 2024), medicine (Roberts et al., 2021; Vandewiele et al., 2021; Varoquaux and Cheplygina, 2022), psychology (Hullman et al., 2022), and political science (Kapoor and Narayanan, 2023). Kapoor et al. (2024) review the literature that addresses issues with validity, reproducibility, and generalizability in ML-based research. Based on their review, the authors derive guidelines to ensure transparency and reproducibility of ML-based research in computer science, data science, mathematics, social sciences, and biomedical sciences. Gundersen and Kjensmo (2018), Beam et al. (2020), McDermott et al. (2021), and Pineau et al. (2021) also discuss the replicability of ML research.

Contribution. We contribute to the literature on the use of ML in business and economic research. Other than the studies above, we approach the use of ML in business and economic research from a different angle. First, we do not restrict the scope of our study but rather consider all research disciplines in business and economics, such as finance, management, organization studies, and accounting. Second, the primary goal of this study is not to provide guidelines regarding when or for what tasks ML in these disciplines can be *theoretically* useful. Instead, we focus on transparency about the performance of ML models relative to traditional statistical models in studies addressing predictive research questions. Our focus on the transparent comparison of ML to traditional statistical models also sets our study apart from Kapoor et al. (2024) and Leech et al. (2024), who discuss good and bad practices in ML research in general.

In addition to the literature directly related to our study, we add to the following research streams. Our study contributes to ongoing discussions on the rigor and transparency of studies from social science, in particular business and economics, and how transparency affects the quality, credibility, and

[4] Like our study, Pérez-Pons et al. (2022) conduct a comparative analysis of the predictive performance of ML models vs. traditional statistical models. However, our study differs significantly from theirs in all essential aspects. Pérez-Pons et al. (2022) use search strings in major academic databases to identify ML-related studies, identifying 48 relevant articles with a cut-off in June 2020 without further specifying the journals in which the articles are published. They descriptively analyze the prediction performances by presenting whether the traditional method or the ML method achieved the highest performance in each study for which they were able to draw such a conclusion. The greatest difference from our study is that we mainly focus on transparency about the performance of the different types of models. We also quantitatively compare the prediction performances and explain the differences in predictive performance. Finally, our study uses a hand-collected sample of papers, obtained by manually screening all FT50 articles between 2010 and 2023. Given that many relevant articles have been published in recent years, we have collected a significantly larger sample of articles published in top business and economic journals.

impact of research. While existing studies are concerned with the replicability of findings (e.g., Ankel-Peters et al., 2024; Bergh et al., 2017; Brodeur et al., 2020; Camerer et al., 2016; Christensen and Miguel, 2018; Fišar et al., 2024; Gundersen and Kjensmo, 2018; Pérignon et al., 2024; Serra-Garcia and Gneezy, 2021) and address transparency about the evidence-generating process in general (e.g., Christensen and Miguel, 2018; Maula and Stam, 2020; Miguel et al., 2014; Miguel 2021; Nosek et al., 2015), we specifically focus on transparency in reporting relative predictive performance when using novel, potentially more powerful, but also more costly prediction models.

The final part of our analysis adds to the literature on the drivers of citations of academic publications by showing that articles with more rigorous comparisons of prediction models receive higher citations. Previous studies have found differences in citation counts depending, for example, on team size and the extent of collaboration between authors (e.g., Bosquet and Combes, 2013; Adams et al., 2005; Franceschet and Costantini, 2010; Larivière et al., 2014; Wu et al., 2019), replicability of findings (Serra-Garcia and Gneezy, 2021), and writing style (Boghrati et al., 2023). In a broader sense, our study is also connected to the literature evaluating different empirical methods in business and economic research (e.g., Hoetker, 2007; Papies et al., 2023; Starr, 2012; Stone and Rasp, 1991). In this respect, we contribute to the literature by providing an overview of the use of predictive ML models and comparing the reported predictive performance between ML and traditional statistical methods.

The structure of this article is as follows. Section 2 derives our research questions. In Section 3, we describe the data and motivate the variables and methods we employ for our analysis. We present our results in Section 4. Section 5 provides a discussion of our results and implications for future research. Section 6 concludes.

2. Research Questions

This study examines four research questions related to the adoption and use of ML in business and economic research. The term *artificial intelligence* (AI) dates back to the Dartmouth conference in 1956 (McCarthy et al., 2006), and ML is one of its central subfields. AI and ML have since evolved through three major "eras" (Brynjolfsson and Li, 2024) or "waves" (Deng, 2018) from preset, hard-coded rules to multi-layered systems that autonomously learn patterns from vast amounts of data to make predictions. The advent of deep learning models in the recent era of AI, beginning in the early 2010s, led to a dramatic surge in AI applications across various fields and industries. Advances in computational resources and software have improved both the performance and accessibility of ML models, and enabled researchers to unlock new information from innovative data including texts and images (Brynjolfsson and Li, 2024; Deng, 2018; LeCun, 2015; Mullainathan and Spiess, 2017). As Figure 1 shows, the adoption of ML as a method in business and economic research began to grow slowly in the mid-2010s before rapidly accelerating in recent years. To establish a general overview of the status quo and importance of ML both as a topic and as a method for researchers in the realm of business and economics, we pose the following research question:

RQ1. *How widely is machine learning discussed and applied across various business and economic research disciplines?*

We then shift our focus to articles that employ ML models to solve predictive research problems. In our analysis, we distinguish between ML models, such as neural networks, and less complex, more traditional statistical models that have been used for decades in the literature to address predictive research questions. Examples of such traditional statistical models are linear or logistic regression.[5] ML models are not *per se* better suited to research questions in business and economics (Athey and Imbens, 2019). Finding and training an ML model that significantly outperforms traditional statistical models in these areas can therefore be challenging. Furthermore, marginal model improvements may not justify using these more complex and more resource-consuming models. In recent years, however, researchers might have benefited from a methodological bonus point in the publication process when using ML techniques, with the use of ML potentially increasing the probability of getting published. ML models might also attract more attention in the academic community, resulting in an increased number of expected citations. If ML models perform poorly, this can entice researchers to opt against the transparent reporting of results for a conventional statistical benchmark due to lower chances of publication (see, e.g., Sculley et al, 2018; Lin, 2018). Therefore, we are interested in researchers' transparency about the performance of ML models relative to traditional statistical models in studies that address predictive research questions in business and economics. We also examine which paper-, journal-, and author-specific variables are associated with the probability of reporting results for traditional statistical models. We thus pose the following research questions:

RQ2a. *Do authors that apply machine learning models to solve predictive research questions consistently compare their predictive performance against traditional statistical models?*

RQ2b. *Which factors can explain whether studies report the benchmark results of traditional statistical models?*

Third, given that ML studies contain benchmark results for traditional statistical models, we are interested in the reported performance difference between ML and traditional statistical models. If ML models are in fact increasingly being applied to research problems that do not warrant them and can be sufficiently addressed using simpler traditional methods, we would, on average, observe only minor outperformance in favor of the ML model. We are also interested in common factors, such as specific data types, seniority of authors, and collaborations between multiple authors, that can explain whether and how strongly an ML model is reported to outperform traditional statistical models across a wide

[5] In section 3.1.2, we explain in more detail how we differentiate between ML and traditional statistical models.

variety of predictive research questions in the field of business and economics.[6] Thus, we pose the research questions:

RQ3a. *How well do machine learning models perform relative to traditional statistical models in answering predictive research questions in business and economic research?*

RQ3b. *Which factors can explain the magnitude of the reported performance difference between machine learning models and traditional statistical models?*

Finally, we examine whether (1) transparency about the performance of traditional statistical models and (2) the extent of the performance improvement by ML models influence the impact that an article generates, as measured by its citation count. We argue that there are two main explanations for why transparency about the performance of traditional statistical models might be related to the citations a study is able to garner. First, comparing the predictive performance of an ML model not only to that of another ML but also to well-established traditional statistical models adds rigor to the analysis. It enables other researchers to assess the comparative value of using a more complex ML model over a traditional model, increasing the likelihood of them citing the paper. Reporting the performance of traditional models also encourages the application of the new ML methods to similar research questions by other researchers, as they can assess beforehand whether it is worth spending the extra effort in building and training a more complex ML model.

The second possible explanation involves an indirect effect of the transparency on the citation count: the article's transparency might be an indicator of the overall quality of the article. If a study thoroughly assesses the performance of ML models against established traditional statistical models, it might be more likely that the entire study was more rigorously conducted and had gone through a more thorough review process. In this case, an article's citation count might not be directly linked to the transparent reporting of the results of traditional statistical models, but rather from the general quality of the article, which may be correlated with the reporting of traditional statistical models' performance.

Alternatively, based on the findings of Serra-Garcia and Gneezy (2021), one might argue in favor of a negative link between the citations an article generates and the authors' transparency about the relative performance of the ML model. Serra-Garcia and Gneezy (2021) find that non-replicable studies gain more citations than replicable studies. The authors hypothesize that referees may be less stringent if the findings appear more interesting. Similarly, studies that are not transparent about the performance of traditional statistical models can be published without having to compare the performance of the ML model with traditional statistical models if their results are more interesting. In this case, the citation count of less transparent studies should exceed the citation count of studies that report benchmark results for traditional statistical models. We also investigate whether a study receives more citations if it

[6] Section 3.2.2 contains more details on the variables that we consider *ex ante* as potentially relevant factors to explain (1) the transparency about and (2) the magnitude of the performance improvement that ML models generate compared to traditional statistical methods.

reports a particularly high outperformance of ML models over traditional statistical models, provided that at least one traditional statistical model is included. We therefore formulate our last research questions as follows:

RQ4a. *Does the decision to report benchmark results for traditional statistical models explain differences in the citation count of machine learning articles?*

RQ4b. *Does the magnitude of the reported performance improvement by machine learning models explain differences in the citation count of machine learning articles?*

3. Data and Method

3.1. Data

3.1.1 Identification of relevant studies

We draw our hand-collected sample from all studies published in a print issue or as an online article in the 50 journals of the Financial Times Research Rank between January 2010 and June 2023, thereby restricting our sample to research appearing in high-quality refereed academic journals. Figure 2 illustrates our process for manually identifying the relevant articles for our study. The article classification process was divided into two major phases. *Phase I* involved the identification of articles that are related to ML in general, i.e., without a specific focus on whether the studies applied predictive ML models or not. Two researchers independently screened the titles, abstracts, and keywords of 56,262 FT50 articles. Together, they identified a sample of 1,542 articles potentially related to ML.

In *Phase II*, both researchers independently classified the potentially ML-related articles into four categories based on the full text: (1) articles that have been falsely identified as an ML-related study during *Phase I*; (2) articles that are about ML without applying ML models themselves; (3) articles that apply ML models for other purposes than articles of category (4); and (4) articles that apply at least one ML model to answer a predictive research question that is of central interest to the article.

Articles of category (2), which are about ML without applying ML models (like the present study), typically review the use or discuss the role and potential applications of AI and ML in, for example, organizational theory (Leavitt et al., 2021), entrepreneurship (Lévesque et al., 2020), or finance (Goldstein et al., 2021). This category also includes articles studying firm behavior in the age of ML.[7] Articles of category (3), which apply ML models but not to answer one of their main predictive research questions, include, among others, studies in which authors use ML models for feature extraction before addressing their research question using other non-ML techniques (e.g., Banerjee et al., 2023). Another example is given by studies that construct measures or variables via ML techniques and then use them

[7] For example, Cao et al. (2023) investigate whether linguistic tones in corporate disclosure have changed since the advent of natural language processing models such as Bidirectional Encoder Representations from Transformers. Hence, the study of Cao et al. (2023) is related to ML as a topic, but the authors do not apply an ML model to investigate their research question.

as input to predict their variable of interest using a traditional statistical method such as linear regression (e.g., Huang et al., 2021). Articles that apply ML models for causal instead of predictive evidence also fall under this category (e.g., Chernozhukov et al., 2015).

Articles of category (4), which use ML models to answer at least one of their main predictive research questions, are of central interest to our study, in particular to answer RQ2–RQ4. Henceforth, we refer to them as *predictive ML studies*. These studies can include, for example, predicting financial market variables (such as risk premia or returns), earnings, sales, fraud, as well as customer or employee behavior (e.g., Avramov et al., 2022; Bali et al., 2023; Cecchini et al., 2010; Chen et al., 2022; Choudhury et al., 2020; Cui et al., 2018; Gu et al., 2020; Ketzenberg et al., 2020; Matz et al., 2019; Xu et al., 2023).[8]

To assign the articles to one of the four categories based on the full text, we had to remove five articles to which we could not obtain full paper access, leaving us with 1,537 potentially ML-related studies. Cohen's Kappa of .70 indicates substantial interrater reliability for categorizing the 1,537 potentially ML-related articles into categories (1)–(4), coded independently by the two researchers. Only considering the classification of all screened FT50 articles as ML-related studies (category (2)), we achieved an interrater reliability of .94. After each researcher completed *Phase II*, we discussed the differences in the rating with a third researcher and finally arrived at a total of 1,211 articles in which ML plays an important role. In 1,058 of 1,211 articles, the authors apply ML models. We assigned 223 of them to category (4), meaning that the authors of these predictive ML studies use ML models to predict the variable of central importance to answer one of the main research questions of their article.

[Figure 2 about here]

3.1.2 Data extracted from relevant studies

We are interested in the out-of-sample predictive performance of all models used in the identified predictive ML studies. We only collect the main prediction results regarding the main predictive research question from each article. These results are usually found in the first table (or figure) of the results section. We do not consider any subsequent analyses like subsample tests or other robustness analyses. In case the results for multiple samples are displayed in the main results table, we use the results on the largest available sample. From each article, we extract all models and their reported predictive performance according to each measure the authors use to evaluate model performance. Only if the study reports multiple results for the same underlying model with various specifications (e.g., linear regression models with different predictor variables or neural networks with a different number of hidden layers), we consider them as one model and extract the minimum and maximum reported performance for that model. This task was again independently performed by two researchers. We regularly

[8] More applications can be found in Table A.1, where we group the prediction objectives of the identified predictive ML studies into overarching categories.

discussed differences in the extracted information and refined our coding method. The correlation of the number of extracted models and measures per paper of both researchers was, on average, at .92 and .96, respectively. Given that both researchers extracted the same models and measures, the predictive performance was extracted identically for 85% of the models per study, on average.[9]

After extracting the relevant data from the articles, we grouped the reported prediction models into two categories to investigate RQ2–RQ4: (1) traditional statistical models and (2) ML models. We refer to (1) as *traditional benchmark models*, as they can serve as a benchmark to the reported predictive performance of ML models in order to assess their performance improvement relative to the more conventional models that have been used for decades in business and economic research. As Gu et al. (2020) point out, there is no uniform definition of what is considered ML, as it often depends on the context at hand. Our context is the use of ML in applied business and economic research. For example, many ML textbooks begin with linear and logistic regressions, which we clearly consider traditional statistical models in our context. Thus, our classification must be guided by what the broad business and economic literature views as novel ML techniques or, conversely, as traditional statistical models. Our differentiation mainly draws on Mullainathan and Spiess (2017), Athey (2018), Athey and Imbens (2019), Choudhury et al. (2020), and Kelly and Xiu (2023), who analyze the use of ML in empirical economics, management, and finance research, while also distinguishing ML from traditional statistical models. In sum, our distinction of the overall model type (ML vs. traditional statistics) is based on the underlying approach to modeling, the assumptions of the model, the flexibility of its functional form, and the emphasis placed on the interpretability of the model's results versus predictive power.

Traditional statistical benchmark models. These models are rooted in classical statistical and econometric theory. They typically rely on strong parametric assumptions (e.g., normality of the residuals or linearity) and often prioritize inference and interpretability of model parameters to grasp the underlying relationships in the data. They usually involve explicit functional forms with an *a priori* defined relationship between input and outcome variables. The variables in models are often manually selected based on economic theory. Parameter estimation is usually conducted through such methods as ordinary least squares or maximum likelihood. Examples of traditional benchmark models include linear regression, logistic regression, probit regression, generalized linear models, partial least squares regression, principal component regression, and discriminant analysis.

Machine learning (ML) models. ML models emphasize out-of-sample predictive accuracy and adaptability to complex, non-linear relationships over interpretability. They are generally non-parametric or semi-parametric, with fewer assumptions about the data structure. Instead of imposing fixed functional forms, ML models are flexible in their capability to *learn* non-linear patterns from the data through training. They may be better suited for handling high-dimensional or large-scale datasets. These

[9] While this rate was initially at 82% for coding the first 20% of the studies of our sample, it increased to 97% for coding the last 20% of our sample.

models may involve non-linear decision boundaries (e.g., decision trees and support vector machines) or regularization (e.g., LASSO and ridge regression) to select the variables that maximize out-of-sample predictive power automatically. Examples of ML models consistent with our definition include neural networks, random forests, gradient boosting, bootstrap aggregation (bagging), elastic nets, support vector machines, decision trees, LASSO, and ridge regression. Mullainathan and Spiess (2017) and Athey and Imbens (2019) discuss ML methods they perceive as a valuable extension of the traditional statistical and econometric toolset for researchers in empirical economics. The models they deem ML (as opposed to those of traditional statistics or econometrics in the context of the empirical economic literature) are in line with our methods for categorizing ML.[10] We are nevertheless aware that opinions on whether to view specific models as ML or as traditional statistical models can differ. Hence, we asked leading researchers in the field their opinions, and discuss the robustness of our findings to an alternative classification of models in Section 4.5.

Figure 3 depicts the most used traditional benchmark models and ML models in our sample. Note that we summarize different models using their overall model type for graphical illustration. For example, the bar for neural networks can include different types such as feedforward, recursive, and convolutional neural networks or long short-term memory networks. Likewise, the bar for linear regression models also includes principal component regressions. For the analysis in the following sections, we consider them as separate models. More than two-thirds of the studies report the predictive performance of neural networks. Other ML models that are often applied include random forests, support vector machines, and boosting algorithms, especially gradient boosting. In contrast, less than one-third of the studies report results for logit regressions. Results for linear regression models are reported in every fourth article.[11]

[Figure 3 about here]

For our analysis of the performance improvement of ML models relative to traditional benchmark models (RQ3 and RQ4), we group the reported evaluation measures into seven groups. At least one evaluation measure derived from the *confusion matrix*, such as accuracy or precision, is reported in 47% of studies. This is followed by reporting the *area under the (receiver operating characteristics) curve*

[10] For example, Kelly and Xiu (2023), reviewing the use of ML in financial markets, follow Gu et al. (2020) in subsuming methods they consider to be ML under the following definition: "[…] (a) a diverse collection of high-dimensional models for statistical prediction, combined with (b) so-called 'regularization' methods for model selection and mitigation of overfit, and (c) efficient algorithms for searching among a vast number of potential model specifications" (p. 2,225). See section 1.4 in Kelly and Xiu (2023) for details. Our distinction between traditional statistics and ML models also largely aligns with the short article of Bzdok et al. (2018) on differences between statistics and ML in the context of biological systems. Though they acknowledge the vague boundary between both domains, they state: "Statistics requires us to choose a model that incorporates our knowledge of the system, and ML requires us to choose a predictive algorithm by relying on its empirical capabilities" (p. 4).

[11] Naive predictors include, for example, the mean or random walk forecasts. Under cluster analysis, we subsume methods from traditional multivariate statistics such as k-nearest neighbor, k-means, or hierarchical clustering.

(28%), *loss functions* such as the mean squared error or the mean absolute error (27%), the out-of-sample *R-squared* (12%), and *profit* measures such as average profits or annualized returns (3%). In sum, 95% of studies report at least one measure that can be assigned to one of the aforementioned groups. We summarize the remaining measures that cannot be assigned to one of these groups as *other* measures.

Next to the evaluated prediction models and their predictive performance, we gather the following data on predictive ML studies: title, authors, journal, journal impact factors, information on the data type that is used as an input for the prediction models (e.g., texts or images), and a description of the prediction objective (i.e., the thematic context of the main variable of interest that is predicted). Furthermore, we gather data from Google Scholar on the citations of the articles to answer RQ4, including the yearly and total citations of the authors, and the title, journal, and year of all publications that the authors are affiliated with, which we identified through the unique Google Scholar ID. The data from Google Scholar was current as of August 26, 2024. Finally, we obtained the year in which the authors of the studies in our sample obtained their PhD, according to their personal websites or academic CVs. In sum, we were able to obtain the data on all variables for 203 predictive ML studies, which constitute our final sample.[12]

3.2. Variables

3.2.1 Dependent variables

Whereas RQ1 is addressed by analyzing descriptive statistics, we define six dependent variables to answer RQ2–RQ4. For RQ2, we use the variable *traditional benchmark*, which is a dummy variable that equals one if the study reports the predictive performance for at least one traditional benchmark model and zero otherwise. A rigorous comparison of ML to traditional benchmark models would ideally involve multiple ML and multiple traditional benchmark models. Hence, as an alternative dependent variable for RQ2, we use *n traditional models* to denote the number of traditional benchmark models for which a study reported the predictive performance.

The major difficulty in measuring performance differences between ML and traditional benchmark models for RQ3 is to make the reported performance differences comparable, given the heterogeneous nature of the research questions in our sample. For some research questions, an increase in accuracy from 96% to 98% can be economically sizeable, while the increase from 52% to 54% might be economically insignificant for other research questions. A much lower level of predictive performance is considered good for some topics such as stock return predictions, while for other topics such as

[12] We could not extract numerical data on the predictive performance for 17 of the 223 identified predictive ML studies, either due to restricted access or because the performance was merely reported graphically. For three more studies and their authors, we were not able to attain citation data from Google Scholar. For two more studies, we were not able to attain data on the year in which the authors obtained their PhD. Because we only use the data on the year of PhD for robustness reasons, we run our baseline analysis based on 203 sample studies.

healthcare, more accurate predictions can or must be achieved to be economically meaningful. Another difficulty stems from the various measures that studies report to evaluate the models' predictive performance. Again, a 2% increase in the out-of-sample R-squared can tell a whole different story than a 2% increase in accuracy.

To address these issues and make the results comparable across the different prediction problems, we measure the performance improvement of a model relative to the improvement that is usually achieved by using a better-suited model for this specific prediction problem. We propose the following method to approximate the mean improvement in the predictive performance by using a better model. Let $\alpha_{i,j}^{(m)}$ denote the performance measure for model j in paper i and m the model type that can be either ML, $\alpha_{i,j}^{(ML)}$, or a traditional benchmark, $\alpha_{i,j}^{(B)}$. First, we order the predictive performances of all n_i models reported in paper i such that $\alpha_{i,1}^{(m)} \geq \alpha_{i,2}^{(m)} \geq \ ... \ \geq \alpha_{i,n_i}^{(m)}$. Now, we can compute the incremental performance increase from each model to the next-best model:

$$\Delta_i = \left[\left(\alpha_{i,1}^{(m)} - \alpha_{i,2}^{(m)}\right), \left(\alpha_{i,2}^{(m)} - \alpha_{i,3}^{(m)}\right), \ldots, \left(\alpha_{i,n-1}^{(m)} - \alpha_{i,n}^{(m)}\right)\right]' \tag{1}$$

By averaging across Δ_i, we get an indication of the mean incremental improvement in the predictive performance by using the next-better-suited model. Let $\alpha_{i,max}^{(ML)}$ ($\alpha_{i,max}^{(B)}$) denote the performance of the ML (traditional benchmark) model that achieves the highest predictive performance of all ML (traditional benchmark) models reported in paper i. We measure the difference of the best-performing ML and the best-performing traditional benchmark model relative to the mean incremental performance improvement as:

$$y_i = \frac{\alpha_{i,max}^{(ML)} - \alpha_{i,max}^{(B)}}{mean(\Delta_i)}. \tag{2}$$

For example, $y_i = 2$ implies that the performance improvement of the best-performing ML model relative to the best-performing traditional benchmark model in paper i is twice as large as the incremental improvement that is achieved on average by using the next-best model. We exclude all papers from our analysis of RQ3 that only report one ML and one traditional statistical model, since we cannot measure the mean incremental performance in this case and y_i would always be equal to one.

Because many studies evaluate model performance using several evaluation measures, we first average y_i across the measures within one category of evaluation measures (e.g., confusion matrix or loss functions) and then use only the category that is reported the most often across all studies of our sample (see Section 3.1.2). Thus, we would always use the average over measures derived from the confusion matrix, if available. If not, we would use the area under the curve, if available, and so on. Hence, to investigate RQ3, we use the dependent variable *best ML vs. best benchmark* measured as y_i averaged across the evaluation measures of the same category. The second dependent variable for RQ3

is given by *avg ML vs. best benchmark* using the difference of the average instead of the best ML model performance to the best traditional benchmark performance.

Finally, we address RQ4 using the *yearly citation count* and *total citation count* generated by each study as of August 26, 2024.

3.2.2 Explanatory and control variables

This section outlines the explanatory and control variables relevant to transparency about the relative performance of ML (RQ2) and the reported performance difference from traditional benchmark models (RQ3). We group the variables according to whether they pertain to the paper, its author(s), or the journal in which it was published. At the end of the section, we outline the variables used to explain differences in the citation count of ML studies (RQ4).

Paper-specific variables. Traditional statistical models might often be considered less useful than ML when using textual or visual data—such as images or videos—for prediction tasks. First, using textual or visual data often results in high-dimensional and unstructured data sets[13] that might be less suitable for traditional statistical models without preparing the unstructured data sets and reducing the dimensionality beforehand through regularization or manual feature selection. ML models and particularly deep learning models, such as transformer models or convolutional neural networks, are, in contrast, specifically designed to recognize patterns in high-dimensional text and image data by automatically detecting important features during the model's training, and subsequently using these patterns to generate predictions of the variable of interest (Bishop, 2006; Krizhevsky et al., 2012; Goodfellow et al., 2016; Gentzkow et al., 2019). Text and images can also involve highly non-linear or non-additive relationships between the individual features (e.g., words, phrases, sentences, pixels, coordinates) (Bishop, 2006; Gentzkow et al., 2019; Mullainathan and Spiess, 2017). ML models have a superior ability to capture these non-linear relationships, such as when predicting the sentiment from a complex text or using satellite images to predict economic variables (Mullainathan and Spiess, 2017).

These considerations can lead researchers and reviewers alike to assume ML models to be the more appropriate model choice for such data sets.[14] Hence, studies might be less likely to contain the results for traditional benchmark models (respectively, only for a very small number) if textual or visual data is involved. Furthermore, given that at least one traditional benchmark model is reported, the outperformance of the ML methods might be higher if ML models can better capture the patterns in textual or

[13] Text or words are, for example, often transformed into numerical high-dimensional vectors through word embeddings or transformer models including thousands of features (Gentzkow et al., 2019). As Athey and Imbens (2019) emphasize, ML models can be particularly superior to traditional statistical models when there is a large number of covariates by differentiating between relevant and less relevant variables.

[14] Mullainathan and Spiess (2017), who discuss the applications of ML in empirical economics, state that "ML can deal with unconventional data that is too high-dimensional for standard estimation methods, including image and language information that we conventionally had not even thought of as data we can work with, let alone include in a regression."

visual data sets. The variable *textual/visual data* is a dummy variable that equals one if a study uses textual or visual data as input for predictions and zero otherwise.

As a second paper-specific variable, we include *n ML models* representing the number of ML models reported in a study. For RQ2, we examine whether a higher number of ML models indicates a more comprehensive model comparison overall, which may increase the likelihood of reporting traditional benchmark models or lead to a greater number of reported traditional benchmark models. To explain the performance differences (RQ3), we consider *n ML models* and *n traditional models* for two reasons. First, reporting more models of one group might increase the probability of finding a better-performing model. For example, if researchers try to address their research question using a substantially higher number of ML models than traditional statistical models, it may increase the likelihood of identifying an ML model that significantly outperforms the best benchmark or *vice versa*, resulting in a greater performance difference. Second, we control for the effect the number of reported models has on our performance comparison via the mean incremental performance increase (see Eq. (2)).[15]

Paper-specific control variables for RQ2 and RQ3 include fixed effects for the publication year. For RQ3, we add fixed effects for the type of measure that was used to evaluate the models' predictive performance (confusion matrix, area under the curve, loss functions, out-of-sample R-squared, profit, other) and fixed effects for the overall context of the prediction to account for differences in the performance of ML and traditional statistical models in different fields of application.[16]

Author-specific variables. The rigor of a study, which should include the comparative value of the best prediction model against several other ML and traditional models, could be related to the seniority level of the authors. We argue that seniority should be positively correlated with total citations and high-quality research output. Hence, the variable *average citation count* denotes the average number of total citations garnered by the authors in the years before the publication year of the predictive ML study.[17] Because publications of lower quality can also attract many citations, we consider *average*

[15] Consider a prediction problem for which all ML models achieve similarly high accuracy, and all traditional benchmark models achieve similarly low accuracy. Because the performance differences within one group of models are small, the mean incremental performance increase would be negatively related to the number of reported models. Hence, if two articles study the same prediction problem and obtain identical results, we would measure a higher outperformance of ML compared to traditional benchmark models for the study that reports the results for more models. If all ML models achieve similarly high accuracy and all traditional benchmark models achieve low accuracy, while the difference within the group of traditional benchmark (ML) models is high (low), mean incremental performance increase would be positively (negatively) related to the number of reported traditional benchmark (ML) models. We control for these effects by including the number of reported ML and traditional benchmark models in our regressions.

[16] We list the groups of research questions with a similar overall context in Table A.1, together with a short description and examples. We cross-validated our categorization of the predictive ML studies into groups of similar research questions with a second independent researcher. Cohen's Kappa was .71, indicating already substantial initial agreement. We discussed differences in the coding and were able to iron out most discrepancies. After mutually agreeing on a category wherever possible, Cohen's Kappa increased to .92. Our results remain qualitatively unchanged if we use the differing categorizations of the second researcher.

[17] We removed any citations the predictive ML study attained pre-publication from the authors' total citation count.

FT50 publications as an additional measure of seniority. This variable represents the average number of studies the authors published in FT50 journals prior to the publication year. Publications in FT50 journals can be assumed to be high-quality studies that are likely to be based on rigorous analyses. Both seniority measures convey different information that is potentially important to our analysis. As a third alternative measure of seniority, we include the *average years since PhD*, which we define as the average number of years between the publication year of the predictive ML study and the year in which the authors of the study obtained their PhD.

Alongside the authors' seniority, we investigate whether the *team size*, i.e., the number of authors, is positively associated with the transparency and the reported relative performance of ML to traditional statistical models. Franceschet and Costantini (2010) show that the quality of research can increase with the collaboration of multiple researchers due to increased resources, greater aggregated expertise, and more shared knowledge. If diversity of knowledge and expertise increases with team size, larger teams of authors might produce more comprehensive analyses in our context and are therefore more likely also to report benchmark results for traditional statistical models.[18] Likewise, more knowledge and expertise might increase the probability of training a well-performing ML model, resulting in a higher performance difference from the traditional statistical models.

Journal-specific variables. We use journal-specific variables as control variables. We include fixed effects for the journal's research discipline to account for differences across the various business and economic research disciplines. We assign one of fourteen different research categories to each of the FT50 journals, as listed in Table A.2. We then proxy for the quality of the journal and the refereeing process using *journal impact factor* from Clarivate (2024). Journals having more rigorous review processes might increase the probability that the predictive ML studies they publish benchmark the predictive performance against traditional statistical models. We use the impact factor of the year in which a study was published, calculated based on the citations of all studies that were published in the journal over the previous two years.[19] We note, however, that journal impact factor might not have a large effect on any of our dependent variables because we already restrict our sample to high-quality journals.

Citation analysis. To examine potential differences in the citation count of articles with respect to their transparency about the relative performance of ML models and the size of the reported performance improvement (RQ4), we use four main explanatory variables in separate regression models. We use the dummy variable *traditional benchmark*, which takes the value one if an article contains results for a traditional benchmark model and zero otherwise. We explore the effect of the number of reported traditional benchmark models using the variable *n traditional models*. We also use the performance difference between the best ML and the best traditional benchmark model (*best ML vs. best benchmark*)

[18] Researchers with a strong focus on ML may be less likely to report a large number of traditional benchmark models, while statisticians and economists may be more likely to advocate for the inclusion of traditional statistical approaches alongside ML models.
[19] See https://clarivate.com/webofsciencegroup/essays/impact-factor/.

as well as between the average ML and the best traditional benchmark model (*avg ML vs. best benchmark*).

To consider other factors potentially influencing the citation count of the study, we include the following control variables, which can be grouped into paper-, author-, and journal-specific variables, as above. The number of reported ML models (*n ML models*) is included to analyze whether we observe a similar effect on citations as for the reporting of traditional statistical models. If so, this would suggest that the total number of reported models is generally decisive for citations and that it does not make a difference whether these include mostly ML models or also traditional statistical models. We also include the dummy variable *textual/visual data* because the use of textual and visual data sets can be considered innovative in business and economic research. In recent years, a rapidly growing body of empirical research has seized upon the increasing amount of textual data and the proliferation of powerful methods for its analysis (Gentzkow et al., 2019). Consequently, results of these studies may be of particular interest to the academic community, potentially yielding higher citations.

To measure the general ability of the authors to generate impactful research, we use the sum of total citations the team of authors garnered prior to the publication year of the predictive ML study, excluding pre-publication citations of the focal study. The variable *total citation count* also captures the seniority of the authors combined with their ability to produce highly cited research, meaning that the academic community is generally interested in their research output.[20] The variable captures other author-specific variables that might affect citations, such as gender or writing style (Boghrati et al., 2023). The researchers' general ability to generate impactful research is likely positively linked to future publications. The total citation count moreover captures the potential positive effects on the authors' publications generated by their networks (Bosquet and Combes, 2013) and their efforts in marketing their research output, such as presenting it at conferences.

Although the total citation count should also account for the positive effect that the size of the team of authors has on citations (Adams et al., 2005; Larivière et al., 2014; Wu et al., 2019), we additionally follow Serra-Garcia and Gneezy (2021) in controlling for the number of authors (*n authors*). We further control for the *years since publication* (Serra-Garcia and Gneezy, 2021), the *journal impact factor*, fixed effects for the publication year, the research discipline of the journal (RQ4a), the overall context of the prediction, and the type of performance measure (RQ4b). All dependent and explanatory variables described in this section are summarized in Table A.3.

3.2.3 Summary statistics

Table 1 provides summary statistics for all variables used in the regression models discussed in Sections 4.2–4.4. The majority of studies in our sample were published after 2018. On average, they

[20] Hence, an additional variable for the seniority of the authors, such as the number of FT50 publications, becomes obsolete. We measure the impact of an author's seniority on citations more directly by just using the total previous citations of the author. If an author has many FT50 publications but only a few citations, this might indicate a high level of seniority with apparently little influence on his or her citations.

received 105 citations in total and 17 citations per year as of August 2024. The average number of co-authors per study is between three and four. Prior to the publication year of a predictive ML study, the author teams of half of the studies had accumulated at least 1,840 citations, 5.5 FT50 publications on average, and 4,659 citations in total. The high number of average citations and FT50 publications suggest a high seniority among the researchers who produced the studies on which our analysis is based. On average, the authors obtained their PhD 12.67 years before the publication year of a predictive ML study, which further indicates the high seniority of the researchers in our sample. The average journal impact factor at the time of publication was 5.5, indicating a generally high quality of papers. In addition to the variables, Table 1 presents our categorization of the studies according to the general research discipline of the journal, the thematic context of their predictive research question, and the type of measure used to evaluate predictive performance. Figure 5, Table A.1, and Table A.2 provide details. Pairwise correlations of the variables presented in Table A.4 are mostly low, suggesting that our regression models do not suffer from multicollinearity. For variables with moderate correlations, we report variance inflation factors when discussing the results of the respective model.

[Table 1 about here]

3.3 Method

First, we show how widely ML is covered and applied in the various business and economic research disciplines by analyzing descriptive statistics (RQ1). Likewise, we use descriptive statistics to explore the transparency of predictive ML studies about the performance of traditional benchmark models (RQ2a) and the difference between the reported performance of ML models and traditional benchmark models (RQ3a). Second, we use various regressions to explore the factors that are related to the transparency of the model reporting (RQ2b) and the magnitude of the reported performance improvement of ML over traditional statistical models (RQ3b). Lastly, we test whether the transparency of the model reporting and the magnitude of the outperformance correlate with a study's citation count.

For RQ2b, we estimate the following probit model to identify factors that are associated with the probability of reporting the results for traditional benchmark models:

$$\Pr(\textit{traditional benchmark}_i = 1) = \Phi(\beta_0 + X_i^{Paper}\beta_1 + X_i^{Author}\beta_2 + X_i^{Journal}\beta_3 + \delta_i^{Journal\ discipline} + \delta_i^{Publication\ year}), \quad (3)$$

where i denotes the predictive ML study, $X_i^{(\cdot)}$ respectively denotes row-vectors of paper-, author-, or journal-specific variables, and δ_i denotes fixed effects. We also regress the variable *traditional benchmark* in a linear probability model on the same variables and fixed effects as in Eq. (3).

Second, we estimate a Poisson regression model for the number of traditional benchmark models (*n traditional models*) for which a predictive ML study reports the predictive performance as the dependent variable. We use the same variables and fixed effects as in Eq. (3). To test whether the results

are specific to the transparency about the reporting of traditional statistical models, we estimate the same regression with the number of reported ML models as the dependent variable.

Third, we investigate whether common factors can explain the magnitude of the outperformance of ML models relative to traditional benchmark models across many different prediction problems (RQ3b). We estimate the following linear regression model to address RQ3b:

$$\begin{aligned} \textit{best ML vs. best benchmark}_i = {} & \beta_0 + X_i^{Paper}\beta_1 + X_i^{Author}\beta_2 + X_i^{Journal}\beta_3 + \delta_i^{Prediction\ context} \\ & + \delta_i^{Performance\ measure} + \delta_i^{Publication\ year} + \epsilon_i. \end{aligned} \tag{4}$$

Additionally, we estimate Eq. (4) using *avg ML vs. best benchmark* as the dependent variable.

Finally, we model the yearly citation count of a study in a negative binomial regression using a random effects estimator (RQ4):

$$\begin{aligned} E\left(\textit{yearly citation count}_{i,t}\right) = \exp(\beta_0 + \textit{traditional benchmark}_i\beta_1 + \textit{controls}_i\beta_2 \\ + \textit{years since publication}_{i,t}\beta_3 + \delta_i^{Journal\ discipline} \\ + \delta_i^{Publication\ year} + u_i), \end{aligned} \tag{5}$$

where $\textit{controls}_i$ are the time-invariant paper-, author-, and journal-specific control variables, and u_i is the random effect for study i. In a second analysis, we measure the transparency about the performance of traditional benchmark models with the number of reported benchmark models (*n traditional models)*. Thereafter, we replace our explanatory variable indicating transparency about the model performance with the performance difference of the best or, alternatively, average ML and the best traditional benchmark model given that a traditional statistical model is reported. Our regression model (5) is closely related to the analysis of Serra-Garcia and Gneezy (2021), who examine the link between the replicability of scientific studies and their citation count. Unlike Serra-Garcia and Gneezy (2021), we estimate a negative binomial instead of a Poisson regression model, given that citation data is usually highly overdispersed (see Table 1 and, e.g., Boghrati et al., 2023), but we report Poisson regression estimates for robustness. As in Serra-Garcia and Gneezy (2021), we also report the results for regressing the *total citation count* on the time-invariant explanatory and control variables from Eq. (5).

4. Results

4.1. Descriptive Analysis

We begin this section by descriptively analyzing our data set to examine how widely ML is discussed and applied across business and economic research disciplines (RQ1), how transparent predictive ML studies are about the performance of traditional benchmark models (RQ2a), and how the reported performance of ML and traditional statistical models differs (RQ3a).

Figure 4 categorizes the identified 1,211 ML-related articles published in FT50 journals between 2010 and 2023 according to the overall research disciplines of the journals (RQ1). In terms of the proportion of all publications within each discipline, ML is most frequently addressed in information

systems, marketing, and operations management/research journals. The majority of the publications use ML as a method, irrespective of the research discipline. Only a small proportion of studies discuss ML as a topic without applying any ML models. The share of ML-related publications in human relations, organization studies, psychology, economics, and entrepreneurship is small compared to management, finance, and consumer research journals. Despite the wide range of applications due to the abundance of text data, we also identify surprisingly few ML-related publications published in accounting journals.

[Figure 4 about here.]

We now shift our focus to predictive ML studies. As illustrated in Figure 5, almost two out of three predictive ML studies in our final sample stem from information systems, operations management/research, and general management journals. This is followed by publications in marketing, finance, and accounting journals. Studies from psychology, entrepreneurship, economics, consumer research, ethics, and organization science constitute only a small portion of the sample. Our sample does not entail predictive ML studies from journals focusing on topics from human relations.

[Figure 5 about here.]

Next, we examine the transparency of predictive ML studies in reporting benchmark results of traditional statistical models to address RQ2a. As Figure 6 shows, we find that 57 out of 203 articles (28%) do not report results for traditional benchmark models, making it difficult to assess the comparative advantage of using more complex and resource-intensive ML models.[21] This finding cannot be entirely attributed to the use of alternative types of unstructured data, such as text or images. While many ML prediction models can inherently handle these data types, extensive pre-processing is often required to use unstructured data with conventional prediction models. As a result, researchers might assume a significant advantage in using ML over traditional statistical models *ex ante* and choose not to compare their results with those of conventional methods. Indeed, 54% of the studies that do not report traditional statistical models leverage textual or visual data to address their predictive research questions. Conversely, two-thirds of the articles using textual or visual data also include at least one traditional statistical model to compare the predictive performance of ML models. If we exclude studies involving textual or visual data, we find that still 22% of the remaining studies do not report the benchmark performance of traditional statistical models. Figure 7 suggests notable differences between research disciplines in reporting traditional benchmark results. For example, 94% of articles published in finance journals compare the ML model performance with traditional statistical models, while a mere 69% of articles in information systems journals and 62% in marketing journals report such comparisons.

[21] Excluding naive prediction models (e.g., random walk or mean forecasts) that often serve as benchmarks for more sophisticated traditional statistical models, we find that 31% of the studies do not report the predictive performance of (sophisticated) traditional statistical models.

[Figure 6 about here]

[Figure 7 about here]

The number of reported ML models and traditional statistical models per study, grouped by research disciplines, are presented in Figure 8. Except for economics, predictive ML studies in all other disciplines report substantially more ML models than traditional models. On average, 3.2 ML models are reported per study, compared with 1.1 traditional statistical models. This difference is particularly pronounced in information systems and marketing journals. Studies that include at least one traditional benchmark model report on average 2.5 times as many ML models as traditional models.

[Figure 8 about here]

Figure 9 illustrates the performance improvement of ML models compared to traditional benchmarks (RQ3a). We find that, on average, the difference in the predictive performance of the best-performing ML and the best-performing traditional benchmark model is twice as large as the mean incremental performance improvement (see Eq. (2)) across all models whose prediction results are reported. In 87% of studies, the best-performing ML model outperforms the best-performing traditional benchmark model. However, based on the mean predictive performance of all ML models reported in a study, ML outperforms the best traditional statistical models in only 69% of studies. The average ML outperformance is then reduced to .7 times the mean incremental performance improvement. As Figure 9 shows, the distribution of the reported performance differences then becomes left-skewed towards a very small or even negative ML outperformance. In less than half of the studies, the worst-performing ML model achieves higher predictive performance than the best-performing traditional statistical model and underperforms the traditional benchmark model on average.

Our analysis indicates that outperforming established traditional statistical models with ML models is not straightforward for many research questions in business and economics. Furthermore, researchers seem to experiment with many more ML models than traditional benchmarks to identify the best-performing model (Figure 8). This imbalance may indicate a potentially unfair comparison between the two model types. Moreover, it is not unreasonable to assume that our findings may be subject to a publication bias (e.g., Christensen and Miguel, 2018; Stanley, 2005). It is likely that using one or more ML models might often be one of the main contributions of a predictive ML study given that almost 90% of the predictive ML studies report an outperformance in favor of the best-performing ML model. Hence, predictive studies in which ML models do not improve upon traditional statistical models might not be published in the first place or do not report results of inferior ML models. On the other hand, authors might not report benchmark results of traditional statistical models in case of marginal improvements by the ML model. Hence, these results cannot be included in our performance comparison.

Many studies across different scientific disciplines have also shown that studies employing ML prediction models often overstate their predictive performance due to methodological flaws or comparison against weak benchmarks (Andaur Navarro et al., 2021; Arbabshirani et al., 2017; Kapoor and Narayanan, 2023; Kapoor et al. (2024); Lin, 2018; Rosenblatt et al., 2024; Sculley, 2018; Vandewiele et al., 2021).[22] Leech et al. (2024) discuss questionable practices in ML research and disparage the "cherrypicking" of results, which involves the unintentional or deliberate selection and reporting of weak benchmarks. Also, the large number of "researcher degrees of freedom" when applying ML models can yield an overstated outperformance of ML over traditional statistical models (Leech et al. 2024; Simmons et al., 2011).

In summary, all of this could cause an overestimation of the actual performance improvement that ML models attain on average compared to traditional statistical models. Our results are therefore conservative estimates. Given that the mean predictive performance of ML models is only slightly better than the best traditional benchmark model (compared to the performance improvement generated by the best ML model), we conclude from our findings that a substantial effort in terms of time and energy is often required to find and train an ML model that yields a significant outperformance over established traditional statistical models.

[Figure 9 about here]

4.2 Reporting benchmark results for traditional statistical models

In this section, we examine whether we can explain the transparency of papers regarding the results of traditional statistical models with paper-, author-, and journal-specific variables (RQ2b). Column (2) of Table 2 presents the results of the probit regression model of Eq. (3), using a binary dependent variable indicating whether the study contains results for at least one traditional statistical model or not. Column (1) refers to the same regression using a linear probability model (LPM). Our results suggest that papers using textual or visual data sets for prediction are less likely to report traditional benchmark results. This is consistent with our argumentation in Section 3.2.2. Using textual or visual data is associated with a 13.5% lower probability of reporting traditional benchmark results. The coefficient is significant at the 5% (10%) level in the probit (linear probability) model. We do not find any evidence that the seniority of the authors, the number of authors, or the journal impact factor are related to the probability of reporting traditional benchmark results. Our findings do not change if we use the average

[22] Kapoor and Narayanan (2023), Kapoor et al. (2024), and Rosenblatt et al. (2024) point out the problem of data leakage, which is evident in many published ML studies across various disciplines. Data leakage refers to when a prediction model receives "out-of-sample" information on the data in the training phase, which can ultimately result in exaggerated "out-of-sample" performance. For example, Kapoor and Narayanan (2023) review twelve papers on civil war prediction in top political science journals. Their replication study reveals that all papers suggesting that ML models outperform logit regressions (which were traditionally used in the field) suffer from data leakage. Correcting the models for these errors suggests that ML models do not achieve a meaningful outperformance over traditionally used logit models for civil war prediction.

years since the authors obtained their PhD as an alternative seniority measure. Studies that report more ML models are not more likely to also include at least one traditional benchmark model.[23]

[Table 2 about here]

In a second analysis, we replace the binary outcome variable with the number of reported traditional statistical models. The results in column (3) confirm the negative link between textual or visual data usage and the number of reported traditional statistical models. The total number of reported traditional statistical models in a study positively correlates with the number of reported ML models, which is significant at the 1% level. However, given the small effect size and the absence of a significant relation with the probability of reporting at least one traditional statistical model, we conclude that the number of reported ML models has a weak correlation with transparency about the performance of traditional statistical models. In column (4) we present the results for regressing the number of reported ML models on the paper-, author-, and journal-specific variables. Studies employing more traditional statistical models also tend to employ more ML models. While using textual or visual data negatively correlates with the number of reported benchmark models, this type of data is associated with an increase in reported ML models. Interestingly, the number of authors is positively related to the number of different ML models the researchers apply to address their research question. This finding is consistent with Franceschet and Costantini (2010), who argue that expertise and resources increase with the collaboration of researchers. Larger author teams are potentially associated with more knowledge about different models and more resources to apply them. However, we do not observe a link between team size and the reporting of traditional benchmark models. Consequently, we cannot confirm that a greater number of authors leads to more comprehensive and rigorous research regarding the variety of examined traditional *and* ML models.

4.3 Performance improvement of ML models relative to traditional statistical models

Table 3 contains the results of linear regressions regarding RQ3b, whether common paper-, author-, and journal-specific factors can explain the magnitude of the reported outperformance of ML over traditional statistical models. Columns (1) and (2) refer to a performance comparison of the best ML to the best traditional statistical model within each study, as stated in Eq. (4). Columns (3) and (4) present results for the performance difference between the average ML and the best traditional statistical model as the dependent variable.[24]

[Table 3 about here]

[23] The variance inflation factors of the two seniority measures (correlation of .66) are around 2.1 and between 1.2 to 1.3 for *textual/visual data* and *n traditional models* (correlation of -.21), respectively.
[24] The variance inflation factors of the seniority measures *average citation count* and *average FT50 publications* (*average years since PhD*) are 3.6 and 2.9 (2.1 and 2.6), respectively.

The relatively high adjusted R-squared of .38 to .40 suggests that the model explains a substantial portion of the reported performance difference. We do not find evidence that using textual or visual data is *per se* related to a higher outperformance of ML compared to traditional statistical models. However, our results regarding the effect of textual or visual data might suffer from a potential selection bias. As discussed above, many studies relying on such data sets do not report benchmark results for traditional statistical models, as they might anticipate high outperformance. Thus, these studies are not included in the regression models of this section.

We find that reporting more ML models is associated with a higher outperformance of the best-performing ML model. As outlined in Section 3.2.2, this can be due primarily to two reasons. First, trying out the predictive performance of multiple ML models may increase the likelihood of finding a very accurate model compared to traditional statistical models, particularly if more ML models than traditional statistical models are reported, as suggested by Figure 8. Second, reporting more ML models may decrease the mean incremental performance improvement between the models under the assumption that ML models generally generate a similarly high outperformance relative to reported traditional statistical models. Measuring the performance difference between ML and traditional statistical models relative to the mean incremental performance improvement would then increase our dependent variable for a rather technical reason. We cannot entirely disentangle these two effects. However, if the latter were true, we would assume to observe the same positive effect for the number of reported ML models on the outperformance of the *average* ML model. Columns (3) and (4) show that we cannot establish this effect for the outperformance of the average ML model. In contrast, the outperformance of the average ML over the best traditional statistical model is significantly negatively related to the number of reported traditional statistical models. This supports our first explanation that the more models of one type (ML or traditional) are employed, the more likely it is that more powerful models will be found.

We also find that the seniority of the authors, as measured by their average citation count, is positively related to the outperformance of both the best and the average ML model. Assuming that authors with a higher citation count produce higher-quality research, the positive link between the outperformance and the citation count might indicate that these authors are also more skilled in training performant ML prediction models. However, this conclusion should be taken with a grain of salt as the authors' seniority measured by the average amount of FT50 publications is negatively associated with the magnitude of the ML outperformance. Frequently publishing in FT50 journals might signal that these authors conduct very rigorous empirical analysis, leading to a smaller reported performance difference between ML models and traditional statistical models. Although the coefficient for the average number of FT50 publications is only weakly significant at the 10% level, we obtain the same finding using the average years since the authors obtained their PhD as explanatory variable, for which the estimated effect is significant at the 5% level.

Lastly, our results indicate that collaborations between a greater number of researchers lead to a higher performance of ML models. The estimated effect is only weakly significant at the 10% level when looking merely at team size.

4.4 Citation count

The last part of our analysis examines RQ4, which asks whether transparency about the performance of traditional statistical models and the magnitude of the performance improvement through the use of ML is related to the number of citations an article is able to garner. First, we look at the correlation of the citation count with the transparency of the model reporting (RQ4a). Similar to Franceschet and Costantini (2010), who investigate the link between collaboration cardinality and citation count, Figure 10 depicts the articles' average number of citations per year dependent on the number of reported traditional statistical models (left panel) and the number of reported ML models (right panel).[25] The graphical analysis suggests a higher number of citations if an article is more transparent about the predictive performance of traditional statistical models. On average, articles reporting no traditional benchmark gather 16 citations per year, while articles reporting the results for one (more than one) traditional benchmark model pick up 25 (34) citations per year. In contrast, a positive relationship between the number of reported ML models and the citations of an article is not immediately apparent. Articles reporting one ML model or between four and five models generate 25 to 26 citations per year on average. Articles reporting two to three ML models garner 20 citations per year, whereas only articles with more than five ML models have a clearly higher mean yearly citation count of 39.

[Figure 10 about here]

Table 4 indicates a statistically significant and positive relation between transparency about the predictive performance of traditional statistical models and the impact an article generates when controlling for other factors that influence citations.[26] The left panel of Table 4 shows the results for the total citation count as the dependent variable. Articles reporting the results for at least one traditional benchmark model have accumulated, on average, 49 more citations than articles that entirely omit traditional statistical models, as suggested by the negative binomial regression model. We find that reporting one more traditional statistical model is associated with an overall increase of around 28 citations. The substantially lower AIC for the negative binomial regression compared to the Poisson model confirms that the negative binomial model is the more appropriate choice given the overdispersion in the

[25] The number of papers reporting no benchmark and the number of papers reporting only one ML model are coincidentally identical. A mere 20 articles that do not report a traditional benchmark model also present results for only one ML model.

[26] Although the number of ML models and whether a study employs textual or visual data sets are significantly related to the number of reported traditional benchmark models, variance inflation factors are well below a level of 2, suggesting no multicollinearity issues with our model.

citation count data. However, the findings are also robust to using a Poisson regression model if we exclude articles with more than 1,000 citations.[27]

[Table 4 about here]

The right panel of Table 4 provides the results regarding the yearly citation count using panel data as described in Eq. (5). Depending on the regression model that we use, studies that report traditional benchmark results have, on average, 2.7 to 6.8 citations more per year than studies that do not report traditional benchmark results. Reporting one more traditional statistical model is associated with 1.8 to 4 more yearly citations. The estimated average marginal effects are significant at the 1% or 5% level. In addition, we present the 95% confidence intervals of the average marginal effects of the negative binomial regressions in Figure 11. Studies reporting at least one traditional statistical model (one more traditional statistical model) have, on average, between 0.05 and 5.4 (0.53 and 3.09) more citations per year. According to the Poisson regressions, as depicted in Figure B.1, the yearly difference in citations is between 2.52 and 11.05 (0.96 and 6.96) compared to less transparent studies. The weighted average five-year journal impact factor in our sample as of 2023 is 8.1, meaning that the average article published in the journals in our sample gathered 8.1 citations per year between 2018 and 2023. Thus, the reported effect sizes in the negative binomial regression are considerable and, on average, comparable to the positive effect that an additional year after publication has on the yearly citation count. In contrast, the number of reported ML models per study is not related to the number of citations.

[Figure 11 about here]

As outlined in Section 2, our findings may have several explanations. The citation analysis might suggest that the academic community values the rigor of the analysis, which we argue is stronger if the predictive performance of novel ML models is benchmarked against well-established traditional statistical models. Apart from that, if a published article benchmarks the ML performance against that of traditional statistical models, it might be easier for researchers to gauge the relative advantage of using the more complex ML model for their own future research, resulting in higher citations of the article proposing the model. On the other hand, reporting traditional benchmark models might positively correlate with a study's quality, which ultimately drives its citation count. Our analysis finds no evidence in support of the opposite hypothesis that we derived from the results of Serra-Garcia and Gneezy (2021). If studies that report only ML and no traditional statistical models were more interesting and

[27] If we exclude (two) articles with more than 1,000 citations, we find that articles reporting traditional benchmark results gain 46 more citations according to the negative binomial model (significant at the 1% level) or, respectively, 28 more citations according to the Poisson regression (significant at the 5% level). Reporting one more traditional statistical model is then associated with an increase of 22 (significant at the 1% level) and 19 (significant at the 5% level) citations according to the negative binomial and Poisson regression, respectively.

hence weathered the review process without having to add benchmark results for traditional statistical models, we should have observed a higher citation count for non-reporting articles.

We also find that studies using textual or visual data have a significantly higher citation count than others. The difference in the total (yearly) citation count is estimated at 11 to 82 (0.61 to 6.13) citations, according to the 95% confidence intervals from the negative binomial regression. The average difference in citations between studies that use textual or visual data and those that do not is similar to the difference in citations between studies that report benchmark results for traditional statistical models and those that do not. The main reason for the positive relation might be that the handling of large, unstructured text or image and video data sets became more effective and convenient—or even possible at all—when modern ML techniques found their way into business and economic research (Gentzkow et al., 2019; Mullainathan and Spiess, 2017). As discussed by Gentzkow et al. (2019) and Mullainathan and Spiess (2017), text and other unconventional data sources such as images have gained increasing importance in business and economic research. Hence, it is likely that these studies and their citation count benefit from this trend by contributing novel methods and interesting findings. As expected, the ability of the authors to produce impactful research, as measured by the total citation count before the publication of the predictive ML study, is also positively correlated with the citation count of the ML study. However, estimated effect sizes are tiny. An author team with 10,000 more citations prior to publication receives, on average, less than one additional citation per year or, respectively, around ten more citations in total. Our analysis does not confirm the positive effect of the size of the research teams on citation count as established in prior research (e.g., Adams et al., 2005; Larivière et al., 2014; Wu et al., 2019). The effect of the team size on the citation count is primarily captured by our citation measure.

As Table 5 shows, we do not find compelling evidence that the extent of the performance improvement of ML over traditional statistical models is related to a higher citation count. Also, the coefficients for the number of traditional statistical models and the binary variable indicating the use of textual or visual data are positive but often insignificant, in contrast to the results presented in Table 4. Note that by including the performance comparison of traditional and ML models, we lose more than 40% of our observations associated with papers that do not report benchmark models and those that report only one traditional model and one ML model. For example, one-third of articles involving textual or visual data are omitted from this sample as they do not report traditional benchmark models. The considerably smaller sample size might be one of the main reasons for observing statistically weaker effects in Table 5.

[Table 5 about here]

4.5 Robustness

While we have taken a theoretical approach to classifying prediction methods into traditional and ML models, the assessment of what constitutes a traditional or ML model may vary among researchers.

For this reason, we have asked leading empirical researchers how they would classify the models. A total of 31 authors, who have either published the articles included in our meta-analysis or seminal articles that discuss the use of ML for business and economic research, classified the models in our sample. We find that for most models there is agreement about how they should be classified (Figure C.1). For example, all researchers classify artificial neural networks, random forests, and decision trees as ML, while more than 90% classify linear and logistic regression models as traditional methods. However, researchers disagreed about the classification of Bayesian networks, elastic nets, general additive models, cluster analysis, discriminant analysis, LASSO, and ridge regression.[28] As a robustness test, we focused on the two extremes. We classified all models on which the respondents' opinions diverged as traditional models. Then, we classified these models as ML models. Lastly, we classified all of the ambiguous models in the opposite direction compared to our initial classification; for example, LASSO as traditional and discriminant analysis as ML.

We present the results of our robustness tests in Appendix C. Our findings on transparency and the relative performance of ML models remain qualitatively unchanged by this reclassification. If all ambiguous models are classified as traditional, then still 21% of the studies do not report benchmark results for traditional statistical models. If we consider all ambiguous models as ML, 33% of the studies do not report traditional benchmark results. Finally, if we classify the models in the opposite direction from our initial classification, 24% do not report traditional benchmark results. Similar to our main analysis, the ML outperformance over the best traditional model is reduced by 66 to 69% for all three robustness tests if we consider the average instead of the best ML model performance. Importantly, our main conclusions from the regressions in Sections 4.2–4.4 hold, independent of the classification of ambiguous models. Overall, we conclude that the models with a rather ambiguous classification do not drive our results.

5. Discussion

5.1 Transparency cost trade-off

One question that should be asked is whether the ML models with the best predictive performance are actually more expensive than others. We cannot test this conjecture to the cent or kilowatt hour for the papers in our sample, as the authors themselves most likely often do not know the exact costs and we would have to expect considerable measurement error. However, we can get an indication of whether cost-intensive models drive the relative predictive performance in ML studies by looking at the type of the best-performing models per paper in our sample.

Figure 12 depicts the most used traditional statistical and ML models in the papers of our sample and how often they achieve the highest predictive performance in a paper relative to the total number

[28] For these models, less than two-thirds of the researchers classified a model as either ML or traditional, i.e., more than one-third either classified the model in the opposite direction or did not classify the model.

of papers in which these models are used. In 80% of studies that use ensemble learning techniques, an ensemble learning model also achieves the highest predictive performance of all reported models in these studies. Transformer models achieve the highest performance in 71% of the studies in which they are employed, while neural networks and boosting algorithms are the best-performing models in 60% and 48% of studies, respectively. Measured in absolute terms, neural networks are reported as the best-performing model in 49 studies, followed by random forests (40 studies) and boosting algorithms (27 studies). In general, due to their complexity, large number of parameters, and the need for extensive training data, we can assume that transformer models, neural networks, and ensemble learning techniques also have, on average, the highest energy consumption among the models used in our sample. Boosting algorithms and random forests can also be computationally expensive, depending on the number of trees. On the other hand, models that are generally less complex and computationally demanding, such as LASSO, ridge, linear, and logistic regressions, are only rarely reported as the best-performing model in the studies of our sample.[29]

While we cannot control for the actual energy consumption of the models that are used in the ML studies of our sample, the analysis from Figure 12 suggests that models with high energy consumption are also most often reported as the best-performing models. Therefore, the relationship between the reported relative outperformance of ML models in the studies of our sample and their energy consumption is most likely positive.

[Figure 12 about here]

5.2 Implications for research

Provided that the prediction of an outcome variable is more important for answering a given research question than knowledge about the underlying relationships between the variables of interest, it is well known that ML models *can* be superior to traditional statistical models in achieving this task. However, whether complex ML models are particularly likely to outperform simpler traditional statistical models depends on knowledge about functional forms and the data set that is available to the researcher. The model that is selected for a study is always a discretionary decision of the researchers conducting the study. We encourage researchers to think about the suitability of ML models for addressing a given research problem in light of the potential higher costs in terms of money, time, energy, and less explainability. Also, presenting a convincing motivation for why ML is *ex-ante* likely to perform better than simpler traditional statistical models and providing a transparent comparison between

[29] Of course, the actual energy consumption of the models can vary depending on their specific characteristics and the data sets used. Even a support vector machine with non-linear kernels can be computationally demanding with larger and more complex data sets. However, as we cannot obtain the data on the energy consumption of each model used in the studies of our sample, we assume the following energy consumption that these types of models generally have on average, as depicted in Figure 12: neural networks, transformer models (*very high*); boosting, ensemble learning (*high*); random forests, Bayesian nets (*moderate to high*); support vector machines, decision trees, bagging (*moderate*); LASSO, ridge regression (*moderate to low*); linear regression and logistic regression (*low*).

these model types can prevent publications from recommending ML models that are potentially less useful relative to their traditional alternatives. Similarly, Kapoor et al. (2024) recommend that ML studies should clearly articulate why ML models are used to approach the research question at hand. In this section, we summarize some guiding principles—among others discussed in detail in Athey and Imbens (2019), Kelly and Xiu (2023), Mullainathan and Spiess (2017), and Varian (2014)—regarding the circumstances under which ML models are promising. If these conditions are not met, it is all the more important to report the best possible traditional benchmarks or even forego cost-intensive ML models altogether. We also address the reporting of meaningful traditional benchmarks at the end of this section.

First, as stated by Kelly and Xiu (2023), "machine learning methods are explicitly designed to approximate unknown data generating functions" (pp. 5–6). If it is reasonable to assume that there is a large number of potentially relevant covariates with complex hidden relationships in the data (e.g., nonlinearities and many interactions) that we cannot explicitly know and model, an ML-based data-driven approach to uncover these hidden patterns can be beneficial. In contrast, if we have a clear idea of the functional form, including nonlinear relationships and interactions between the variables, ML methods are less likely to add value over traditional statistical models. Athey and Imbens (2019) illustrate the absence of ML advantages with the example of earning predictions for individuals, where we can plausibly expect linear relationships, and where unknown higher-order interactions or nonlinearities are less likely to be of significant importance. Another example of ML failing to confer an advantage is when we can expect a specific nonlinear relationship such as the inverted U-shape between tax rates and tax revenues, as modeled by the Laffer curve.

The second condition that affects the probability of an ML model outperforming traditional statistical models relates to the data that is available to the researcher. ML models are well suited to handle high-dimensional data combining different potentially unconventional data types such as unstructured textual data together with classic numerical time series data. In theory, many different data sources, for which ML offers promising approaches, can be relevant to predict the variable of interest. On the other hand, complex ML models such as neural networks are generally also reliant on large-scale data sets to be less prone to overfitting and to deliver good out-of-sample performance, whereas the available data sets for many business and economic research questions are relatively small (Athey and Imbens, 2019; Kelly and Xiu, 2023). Note that the two major conditions—unknown functional forms and large, high-dimensional data sets—are often mutually dependent. The data that is available to the researcher determines whether unknown hidden relationships between variables are likely to affect the predictions and how many potentially relevant covariates are available. The nature of the research problem, in turn, determines which data sets can be theoretically useful to form predictions.

If it is likely that an ML model will outperform traditional statistical models given the research question and data of a study, the researchers should still transparently report benchmark results for traditional statistical models. For example, Chen et al. (2024) examine the impact of model design

choices on the performance of ML-based stock return predictions—an *ex ante* promising area for ML applications given potentially large-scale data sets that can be useful to form predictions of returns of financial assets, and *a priori* unknown functional forms (Gu et al. 2020; Kelly and Xiu, 2023). However, Chen et al. (2024) find that one-third of the tested composite ML models for stock return prediction end up with lower economic gains (in terms of portfolio returns) than comparable OLS models with identical design choices. To evaluate the true relative advantage of an ML model, the strongest well-established traditional statistical models should be chosen as benchmarks to avoid "cherrypicking" or "benchmark hacking," which would overstate the relative ML performance, as emphasized by Leech et al. (2024). Reporting a meaningful traditional benchmark also implies that the traditional model generates predictions based on the same data as the ML model.

If ML models are trained on richer data sets that have more predictive value than the data that is fed into the traditional statistical model, the actual ML outperformance might be overstated.[30] For example, if the data set is too high-dimensional to work well with a simple linear regression model, the covariates can be based on principal components (see, e.g., Gu et al., 2020) instead of a very limited number of preselected predictor variables. The study by Chen et al. (2024) can serve as a best-practice example of transparent comparison between the predictive performance of ML and that of traditional statistical models. The authors compare ML models with their OLS counterparts based on the exact same design choices. They also transparently derive from this comparison the conditions under which ML models are likely to outperform the traditional statistical benchmark in stock return prediction, and those where the traditional benchmark is likely to perform better. Likewise, Leech et al. (2024) underline the importance of not only reporting the performance of a "single run" of an ML model, as this does not allow for assessing the uncertainty underlying the ML model's performance. Kapoor et al. (2024) provide further guidance on how to choose appropriate baselines in ML-based scientific research and quantify the uncertainty underlying the reported ML model performance.

Alongside performance-wise comparison, we recommend evaluating performance relative to model costs in order to achieve an even more transparent evaluation of economically meaningful improvements through the use of more costly ML models over simpler traditional statistical models. This would ideally include the measurement of energy consumption, environmental costs, and the loss of explainability or interpretability of the results. We briefly address existing frameworks and efforts to capture the different types of model costs in Section 5.3.

5.3 Limitations and future research

Ideally, the performance of ML and traditional statistical models should be compared separately for individual research questions by including studies that use only traditional statistical models in the absence of ML. While meta studies with a focus on the comparison between ML and traditional

[30] We thank one of the survey participants for this comment.

statistical models within a homogeneous set of research question present an interesting avenue for future research, such a study is beyond the scope of the present article, given the variety of heterogeneous research questions. As an alternative to meta studies, we propose that future studies can comprehensively analyze the predictive performance of ML and traditional statistical models for individual research questions to investigate the conditions under which ML outperforms and when traditional statistical models are superior (e.g., Chen et al., 2024). For example, Bianchi et al. (2024) find that many recent studies propose predicting equity risk premia using novel flexible statistical techniques (including ML), but these studies rarely benchmark the model performance against well-established economically motivated regression models with a restricted set of predictors. Consequently, Bianchi et al. (2024) compare the predictive performance of both model types and find that economically motivated regression models outperform various flexible ML techniques such as LASSO and random forests. Such analyses will enable a clearer view of the true advantages afforded by ML models, depending on the research topic.

Future research is also needed to develop standardized frameworks for reporting the economic benefits of models in comparison to model costs. There is increasing demand in the broader ML literature for standards that frame the performance gains of ML models in terms of their efficiency (Strubell et al., 2020). For reporting frameworks that consider model costs, researchers would need to be aware of the energy consumption of the models they use. While the precise calculation of energy use and environmental impact is not the main focus of this study, Cai et al. (2017), Dodge et al. (2022), García-Martín et al. (2019), Lacoste et al. (2019), and Luccioni et al. (2024) provide interesting approaches. Alternatively, reporting computing resources and the time to run the code for the empirical analysis—ideally broken down for each of the employed models—can give an indication of associated energy consumption and financial costs. Kapoor et al. (2024) also recommend transparently describing the hardware, software, and computational resources available to the researchers when using ML models. Reporting details about the data and code to journals is already requested by the *Data and Code Availability Standard*,[31] which is endorsed by many economic journals, such as those of the American Economic Association. The focus of these reporting standards, however, is more on the replicability of a study's results than on a comparison of model performance and costs.

Along with financial and environmental costs, model costs in terms of sacrifices in explainability and interpretability in exchange of higher predictive performance should be considered. The degree of explainability is particularly relevant for many research problems regarding managerial decision-making, which requires causal knowledge of the underlying relationships between variables (Hünermund et al., 2022). An approach to evaluating the predictive performance of models relative to an interpretability score is provided in Kruschel et al. (2024).

[31] The *Data and Code Availability Standard* and the list of endorsing journals is accessible on the website https://datacodestandard.org./

6. Conclusion

Our study investigates the adoption of predictive ML models in high-quality business and economic research. We find that papers do not consistently report benchmark results for less complex conventional statistical models that have been traditionally used in the literature to answer predictive research questions. However, considering traditional statistical models can be particularly important for assessing whether a statistical improvement in predictive accuracy through the use of a more complex ML model translates into economically significant insights. Overall, our results indicate that, for the average ML model, there is often relatively little gain in predictive performance over traditional benchmark models. Time and energy seem to be required to achieve considerably improved predictions with well-trained ML models. Whether this improvement is always economically significant enough to warrant higher model costs in terms of increased effort, time, and energy consumption remains an open question. Future research needs to develop standardized frameworks in business and economic research for reporting predictive gains of ML models relative to their efficiency. The relative model performance should be compared against well-established and usually less expensive traditional statistical models in that field to evaluate the true economic benefits and relative advantages of more complex, resource hungry models.

References

Adams, J. D., Black, G. C., Clemmons, J. R., & Stephan, P. E. (2005) Scientific teams and institutional collaborations: Evidence from US universities, 1981–1999. *Research Policy* 34(3):259–285. doi: 10.1016/j.respol.2005.01.014.

Andaur Navarro, C. L., Damen, J. A., Takada, T., Nijman, S. W., Dhiman, P., Ma, J., Collins, G. S., Bajpai, R., Riley, R. D., Moons, K. G. M., & Hooft, L. (2021) Risk of bias in studies on prediction models developed using supervised machine learning techniques: Systematic review. *BMJ* 375(2281). doi: 10.1136/bmj.n2281.

Ankel-Peters, J., Fiala, N., & Neubauer, F. (2024) Is economics self-correcting? Replications in the American Economic Review. *Economic Inquiry*. doi: 10.1111/ecin.13222.

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017) Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage* 145:137–165. doi: 10.1016/j.neuroimage.2016.02.079.

Athey, S. (2018) The impact of machine learning on economics. In A. Agrawal, J. Gans, & A. Goldfarb (eds.), *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press), 507–547. doi: 10.7208/9780226613475-023.

Athey, S., & Imbens, G. W. (2019) Machine learning methods that economists should know about. *Annual Review of Economics* 11(1):685–725. doi: 10.1146/annurev-economics-080217-053433.

Avramov, D., Cheng, S., & Metzker, L. (2023) Machine learning vs. economic restrictions: Evidence from stock return predictability. *Management Science* 69(5):2587–2619. doi: 10.1287/mnsc.2022.4449.

Bali, T. G., Beckmeyer, H., Moerke, M., & Weigert, F. (2023) Option return predictability with machine learning and big data. *The Review of Financial Studies* 36(9):3548–3602. doi: 10.1093/rfs/hhad017.

Banerjee, M., Cole, B. M., & Ingram, P. (2023) "Distinctive from what? And for whom?" Deep learning-based product distinctiveness, social structure, and third-party certifications. *Academy of Management Journal* 66(4):1016–1041. doi: 10.5465/amj.2021.0175.

Bawack, R. E., Wamba, S. F., Carillo, K. D. A., & Akter, S. (2022) Artificial intelligence in e-commerce: A bibliometric study and literature review. *Electronic Markets* 32(1):297–338. doi: 10.1007/s12525-022-00537-z.

Beam, A. L., Manrai, A. K., & Ghassemi, M. (2020) Challenges to the reproducibility of machine learning models in health care. *JAMA* 323(4):305–306. doi: 10.1001/jama.2019.20866.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March) On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. doi: 10.1145/3442188.3445922.

Bergh, D. D., Sharp, B. M., Aguinis, H., & Li, M. (2017) Is there a credibility crisis in strategic management research? Evidence on the reproducibility of study findings. *Strategic Organization* 15(3):423–436. doi: 10.1177/1476127017701076.

Bishop, C. M. (2006) *Pattern Recognition and Machine Learning* (Springer, New York).

Boghrati, R., Berger, J., & Packard, G. (2023) Style, content, and the success of ideas. *Journal of Consumer Psychology* 33(4):688–700. doi: 10.1002/jcpy.1346.

Bosquet, C., & Combes, P. P. (2013) Are academics who publish more also more cited? Individual determinants of publication and citation records. *Scientometrics* 97:831–857. doi: 10.1007/s11192-013-0996-6.

Brodeur, A., Cook, N., & Heyes, A. (2020) Methods matter: P-hacking and publication bias in causal analysis in economics. *American Economic Review* 110(11):3634–3660. doi: 10.1257/aer.20190687.

Brynjolfsson, E., & Li, D. (2024) The economics of generative AI. *The Reporter* 1:16–18. url: nber.org/reporter/2024number1/economics-generative-ai.

Bzdok, D., Altman, N., & Krzywinski, M. (2018) Statistics versus machine learning. *Nature Methods* 15(4):233–234. doi: 10.1038/nmeth.4642.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... & Wu, H. (2016) Evaluating replicability of laboratory experiments in economics. *Science* 351(6280):1433–1436. doi: 10.1126/science.aaf0918.

Cai, E., Juan, D. C., Stamoulis, D., & Marculescu, D. (2017) NeuralPower: Predict and deploy energy-efficient convolutional neural networks. *Proceedings of Machine Learning Research (Asian Conference on Machine Learning 2017)* 77, 622–637. https://proceedings.mlr.press/v77/cai17a.html.

Cao, S., Jiang, W., Yang, B., & Zhang, A. L. (2023) How to talk when a machine is listening: Corporate disclosure in the age of AI. *The Review of Financial Studies* 36(9):3603–3642. doi: 10.1093/rfs/hhad021.

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010) Detecting management fraud in public companies. *Management Science* 56(7):1146–1160. doi: 10.1287/mnsc.1100.1174.

Chen, M., Hanauer, M.X., & Kalsbach, T. (2024) Design choices, machine learning, and the cross-section of stock returns. *SSRN working paper*. doi: 10.2139/ssrn.5031755.

Chen, X., Cho, Y. H., Dou, Y., & Lev, B. (2022) Predicting future earnings changes using machine learning and detailed financial data. *Journal of Accounting Research* 60(2):467–515. doi: 10.1111/1475-679X.12429.

Chernozhukov, V., Hansen, C., & Spindler, M. (2015) Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review* 105(5):486–490. doi: 10.1257/aer.p20151022.

Chou, Y. C., Chuang, H. H. C., Chou, P., & Oliva, R. (2023) Supervised machine learning for theory building and testing: Opportunities in operations management. *Journal of Operations Management* 69(4):643–675. doi: 10.1002/joom.1228.

Choudhury, P., Allen, R. T., & Endres, M. G. (2020) Machine learning for pattern discovery in management research. *Strategic Management Journal* 42(1):30–57. doi: 10.1002/smj.3215.

Clarivate (2024) Journal Citation Reports. url: https://jcr.clarivate.com.

Christensen, G., & Miguel, E. (2018) Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature* 56(3):920–980. doi: 10.1257/jel.20171350.

Cui, R., Gallino, S., Moreno, A., & Zhang, D. J. (2018) The operational value of social media information. *Production and Operations Management* 27(10):1749–1769. doi: 10.1111/poms.12707.

Deng, L. (2018) Artificial intelligence in the rising wave of deep learning: The historical path and future outlook. *IEEE Signal Processing Magazine* 35(1):180–177. doi: 10.1109/MSP.2017.2762725.

Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019) Show your work: Improved reporting of experimental results. *arXiv preprint*. doi: 10.18653/v1/D19-1224.

Dodge, J., Prewitt, T., Tachet des Combes, R., Odmark, E., Schwartz, R., Strubell, E., Luccioni, A.S., Smith, N.A., DeCario, N., & Buchanan, W. (2022) Measuring the carbon intensity of AI in cloud instances. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1877–1894. doi: 10.1145/3531146.3533234.

Ebert, K., Alder, N., Herbrich, R., & Hacker, P. (2024) AI, climate, and regulation: From data centers to the AI Act. *arXiv preprint*. doi: 10.48550/arXiv.2410.06681.

Fišar, M., Greiner, B., Huber, C., Katok, E., Ozkes, A. I. (2024) Reproducibility in management ccience. *Management Science* 70(3):1343–1356. doi: 10.1287/mnsc.2023.03556.

Franceschet, M., & Costantini, A. (2010) The effect of scholar collaboration on impact and quality of academic papers. *Journal of Informetrics* 4(4):540–553. doi: 10.1016/j.joi.2010.06.003.

García-Martín, E., Rodrigues, C. F., Riley, G., & Grahn, H. (2019) Estimation of energy consumption in machine learning. *Journal of Parallel and Distributed Computing* 134:75–88. doi: 10.1016/j.jpdc.2019.07.007.

Gentzkow, M., Kelly, B., & Taddy, M. (2019) Text as data. *Journal of Economic Literature* 57(3):535–574. doi: 10.1257/jel.20181020.

Gundersen, O. E., & Kjensmo, S. (2018) State of the art: Reproducibility in artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence* 32(1):1644–1651. doi: https://doi.org/10.1609/aaai.v32i1.11503.

Goodfellow, I., Bengio, Y., & Courville, A. (2016) *Deep Learning* (MIT Press, Cambridge).

Goldstein, I., Spatt, C. S., & Ye, M. (2021) Big data in finance. *The Review of Financial Studies* 34(7):3213–3225. doi: 10.1093/rfs/hhab038.

Gu, S., Kelly, B., & Xiu, D. (2020) Empirical asset pricing via machine learning. *The Review of Financial Studies* 33(5):2223–2273. doi: 10.1093/rfs/hhaa009.

Heinrich, K., & Keshavarzi, A. (2024) Are our predictions healthy? A comparative meta-analysis of machine learning studies in predictive healthcare. *ECIS 2024 Proceedings*.

Hoetker, G. (2007) The use of logit and probit models in strategic management research: Critical issues. *Strategic Management Journal* 28(4):331–343. doi: 10.1002/smj.582.

Hofman, J. M., Sharma, A., & Watts, D. J. (2017) Prediction and explanation in social systems. *Science* 355(6324):486–488. doi: 10.1126/science.aal3856.

Huang, D., Li, J., & Wang, L. (2021) Are disagreements agreeable? Evidence from information aggregation. *Journal of Financial Economics* 141(1):83–101. doi: 10.1016/j.jfineco.2021.02.006.

Hünermund, P., Kaminski, J., & Schmitt, C. (2022) Causal machine learning and business decision making. *SSRN working paper*. doi: 10.2139/ssrn.3867326.

Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A., & Narayanan, A. (2022) The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* 335–348. doi: 10.1145/3514094.3534196.

Janiesch, C., Zschech, P., & Heinrich, K. (2021) Machine learning and deep learning. *Electronic Markets*, 31(3):685–695. doi: 10.1007/s12525-021-00475-2.

Kapoor, S., & Narayanan, A. (2023) Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4(9):100804. doi: 10.1016/j.patter.2023.100804.

Kapoor, S., Cantrell, E. M., Peng, K., Pham, T. H., Bail, C. A., Gundersen, O. E., ... & Narayanan, A. (2024) REFORMS: Consensus-based recommendations for machine-learning-based science. *Science Advances* 10(18):1–17. doi: 10.1126/sciadv.adk3452.

Ketzenberg, M. E., Abbey, J. D., Heim, G. R., & Kumar, S. (2020) Assessing customer return behaviors through data analytics. *Journal of Operations Management* 66(6):622–645. doi: 10.1002/joom.1086.

Kelly, B., & Xiu, D. (2023) Financial machine learning. *Foundations and Trends® in Finance* 13(3–4):205–363. doi: 10.1561/0500000064.

Kraus, M., Feuerriegel, S., & Oztekin, A. (2020) Deep learning in business analytics and operations research: Models, applications and managerial implications. *European Journal of Operational Research* 281(3):628–641. doi: 10.1016/j.ejor.2019.09.018.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012) Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* 25.

Kruschel, S., Hambauer, N., Weinzierl, S., Zilker, S., Kraus, M., & Zschech, P. (2024) Challenging the performance-interpretability trade-off: An evaluation of interpretable machine learning models. *arXiv preprint*. doi: 10.48550/arXiv.2409.14429.

Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019) Quantifying the carbon emissions of machine learning. *arXiv preprint*. doi: 10.48550/arXiv.1910.09700.

Larivière, V., Gingras, Y., Sugimoto, C. R., & Tsou, A. (2015) Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology* 66(7):1323–1332. doi: 10.1002/asi.23266.

Leavitt, K., Schabram, K., Hariharan, P., & Barnes, C. M. (2021) Ghost in the machine: On organizational theory in the age of machine learning. *Academy of Management Review* 46(4):750–777. doi: 10.5465/amr.2019.0247.

LeCun, Y., Bengio, Y., & Hinton, G. (2015) Deep learning. *Nature* 521(7553):436–444. doi: 10.1038/nature14539.

Leech, G., Vazquez, J. J., Kupper, N., Yagudin, M., & Aitchison, L. (2024) Questionable practices in machine learning. *arXiv preprint*. doi: 10.48550/arXiv.2407.12220.

Lévesque, M., Obschonka, M., & Nambisan, S. (2020) Pursuing impactful entrepreneurship research using artificial intelligence. *Entrepreneurship Theory and Practice* 46(4):803–832. doi: 10.1177/1042258720927369.

Lin, J. (2018) The neural hype and comparisons against weak baselines. *ACM SIGIR Forum* 52(2):40–51. doi: 10.1145/3308774.3308781.

Luccioni, S., Jernite, Y., & Strubell, E. (2024) Power hungry processing: Watts driving the cost of AI deployment? *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 85–99. doi: 10.1145/3630106.3658542.

Matz, S. C., Segalin, C., Stillwell, D., Müller, S. R., & Bos, M. W. (2019) Predicting the personal appeal of marketing images using computational methods. *Journal of Consumer Psychology* 29(3):370–390. doi: 10.1002/jcpy.1092.

Maula, M., & Stam, W. (2020) Enhancing rigor in quantitative entrepreneurship research. *Entrepreneurship Theory and Practice* 44(6):1059–1090. doi: 10.1177/1042258719891388.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006) A proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine* 27(4):12. doi: 10.1609/aimag.v27i4.1904.

McDermott, M. B., Wang, S., Marinsek, N., Ranganath, R., Foschini, L., & Ghassemi, M. (2021) Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine* 13(586). doi: 10.1126/scitranslmed.abb1655.

Menkveld, A. J., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., Neusüß, S., Razen, M., Weitzel, U. et al. (2024) Nonstandard errors. *The Journal of Finance* 79(3):2339–2390. doi: 10.1111/jofi.13337.

Messeri, L., & Crockett, M. J. (2024) Artificial intelligence and illusions of understanding in scientific research. *Nature* 627, 49–58. doi: 10.1038/s41586-024-07146-0.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., ... & Van der Laan, M. (2014) Promoting transparency in social science research. *Science* 343(6166):30–31. doi: 10.1126/science.1245317.

Miguel, E. (2021) Evidence on research transparency in economics. *Journal of Economic Perspectives* 35(3):193–214. doi: 10.1257/jep.35.3.193.

Mullainathan, S., & Spiess, J. (2017) Machine learning: An applied econometric approach. *Journal of Economic Perspectives* 31(2):87–106. doi: 10.1257/jep.31.2.87.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Yarkoni, T. (2015) Promoting an open research culture. *Science* 348(6242):1422–1425. doi: 10.1126/science.aab2374.

Papies, D., Ebbes, P., & McDonnel Freit, E. (2023) Endogeneity and causal inference in marketing. In R.S. Winer & S.A. Neslin (eds.), *The History of Marketing Science*, *World Scientific-Now Publisher Series in Business, 18* (World Scientific Publishing Co. Pte. Ltd.), 253–300. doi: 10.2139/ssrn.4091717.

Pérez-Pons, M. E., Parra-Dominguez, J., Omatu, S., Herrera-Viedma, E., & Corchado, J. M. (2022) Machine learning and traditional econometric models: a systematic mapping study. *Journal of Artificial Intelligence and Soft Computing Research* 12(2):79–100. doi: 10.2478/jaiscr-2022-0006.

Pérignon, C., Akmansoy, O., Hurlin, C., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., Menkveld, A.J., & Weitzel, U. (2024) Computational reproducibility in finance: Evidence from 1,000 tests. *The Review of Financial Studies*. doi: 10.1093/rfs/hhae029.

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., ... & Larochelle, H. (2021) Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research* 22:1–20.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., ... & Schönlieb, C. B. (2021) Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence* 3(3):199–217. doi: 10.1038/s42256-021-00307-0.

Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S., & Scheinost, D. (2024) Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications* 15(1):1829. doi: 10.1038/s41467-024-46150-w.

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020) Green AI. *Communications of the ACM* 63(12):54–63. doi: 10.1145/3381831.

Sculley, D., Snoek, J., Wiltschko, A., & Rahimi, A. (2018) Winner's curse? On pace, progress, and empirical rigor. *ICLR 2018 Workshop Track*.

Serra-Garcia, M., & Gneezy, U. (2021) Nonreplicable publications are cited more than replicable ones. *Science Advances* 7(21), eabd1705. doi: 10.1126/sciadv.abd1705.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* 22(11):1359–1366. doi: 10.1177/0956797611417632.

Starr, M. A. (2014) Qualitative and mixed-methods research in economics: Surprising growth, promising future. *Journal of Economic Surveys* 28(2):238–264. doi: 10.1111/joes.12004.

Stone, M., & Rasp, J. (1991) Tradeoffs in the choice between logit and OLS for accounting choice studies. *The Accounting Review* 66(1):170–187. url: jstor.org/stable/247712.

Strubell, E., Ganesh, A., & McCallum, A. (2019) Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650. doi: 10.48550/arXiv.1906.02243.

Strubell, E., Ganesh, A., & McCallum, A. (2020) Energy and policy considerations for modern deep learning research. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(09):13693–13696. doi: 10.1609/aaai.v34i09.7123.

Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2021) Deep learning's diminishing returns: The cost of improvement is becoming unsustainable. *IEEE Spectrum* 58(10):50–55. doi: 10.1109/MSPEC.2021.9563954.

Vandewiele, G., Dehaene, I., Kovács, G., Sterckx, L., Janssens, O., Ongenae, F., ... & Demeester, T. (2021) Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling. *Artificial Intelligence in Medicine* 111, 101987. doi: 10.1016/j.artmed.2020.101987.

Varian, H. R. (2014) Big data: New tricks for econometrics. *Journal of Economic Perspectives* 28(2):3–28. doi: 10.1257/jep.28.2.3.

Varoquaux, G., & Cheplygina, V. (2022) Machine learning for medical imaging: Methodological failures and recommendations for the future. *npj Digital Medicine*, 5. doi: 10.1038/s41746-022-00592-y.

Wu, L., Wang, D., & Evans, J. A. (2019) Large teams develop and small teams disrupt science and technology. *Nature* 566(7744):378–382. doi: 10.1038/s41586-019-0941-9.

Xu, X., Xiong, F., & An, Z. (2023) Using machine learning to predict corporate fraud: Evidence based on the gone framework. *Journal of Business Ethics* 186(1):137–158. doi: 10.1007/s10551-022-05120-2.

Figures

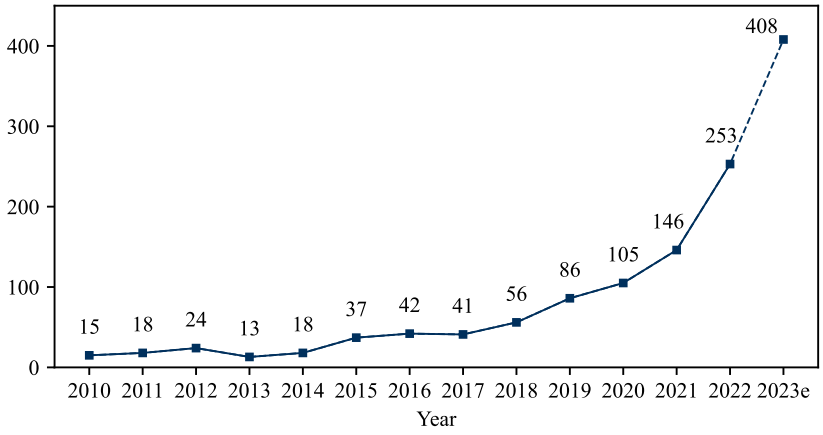

Figure 1: Number of publications in journals from the Financial Times Research Rank (FT50), in which the authors apply ML models for their empirical analysis. We collected studies published between 2010 and June 2023. The expected number of publications for 2023 (2023e) is based on 204 studies we identified that were published in print or online issues between January and June 2023.

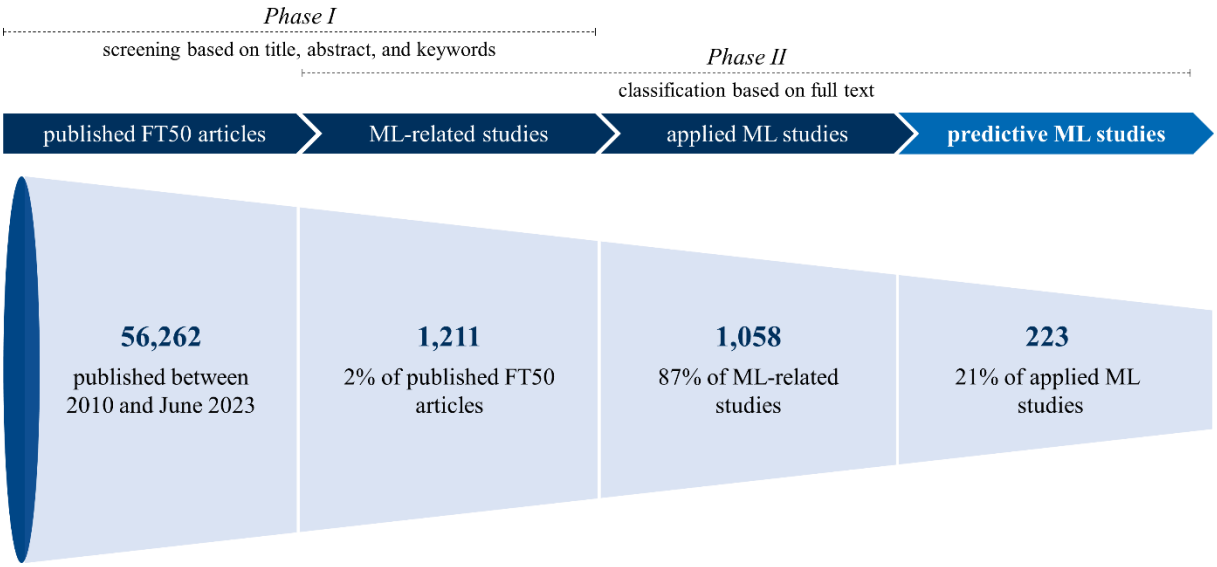

Figure 2: Identification of relevant studies and identified number of studies that involve ML in general (ML-related studies), studies in which the authors apply ML models (applied ML studies), and studies in which the authors predict the variable of interest with at least one ML model in order to answer a research question that is of central importance to their study (predictive ML studies).

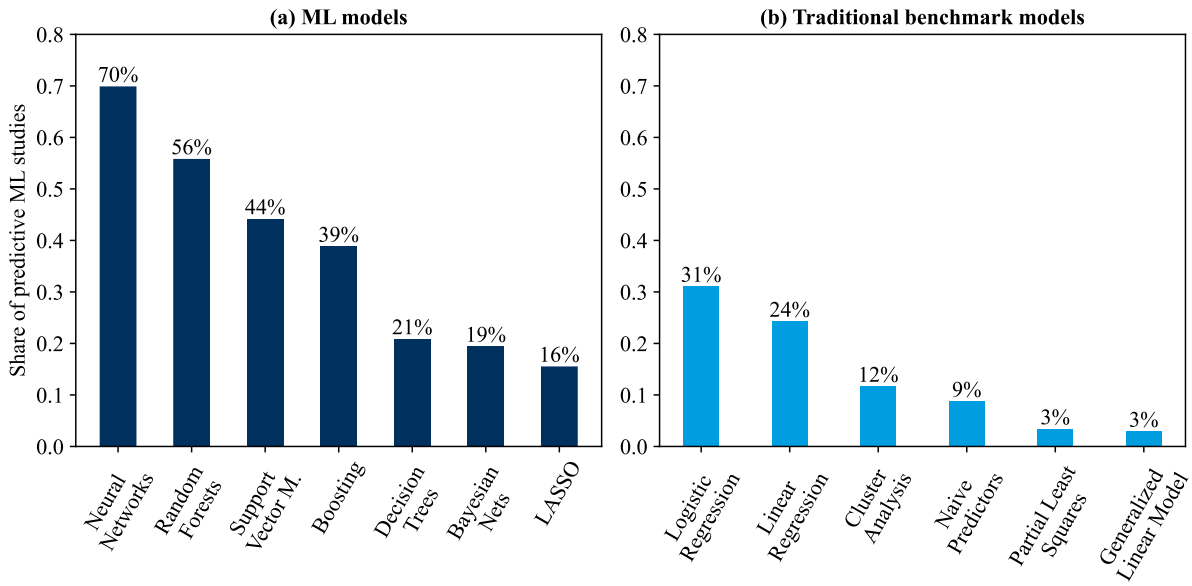

Figure 3: Share of predictive ML studies that cover the most reported models in our sample grouped by (a) ML models and (b) traditional statistical models considered as benchmarks in our paper.

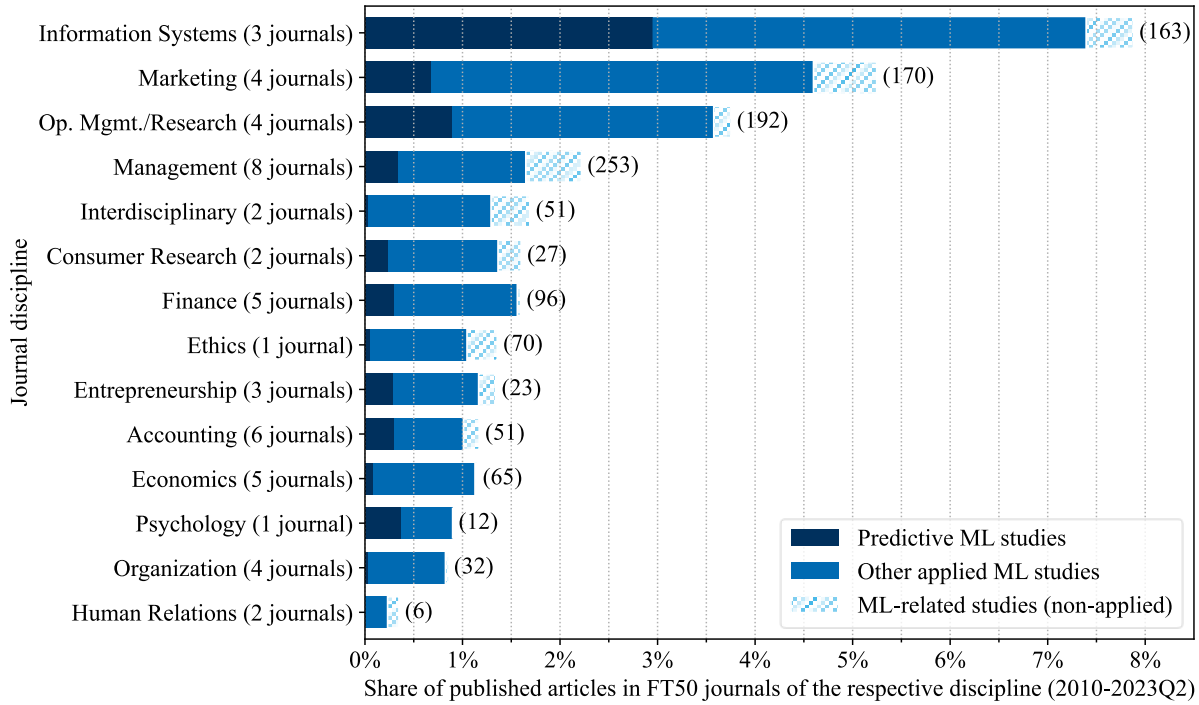

Figure 4: Share (number) of ML-related studies grouped by the research discipline of the journal and categorized into articles that apply at least one ML model to answer a predictive research question that is of central interest to the study (predictive ML studies), articles that apply ML models for other purposes, and articles that are about ML without applying ML models themselves. The categorization of FT50 journals by research discipline is described in Table A.2.

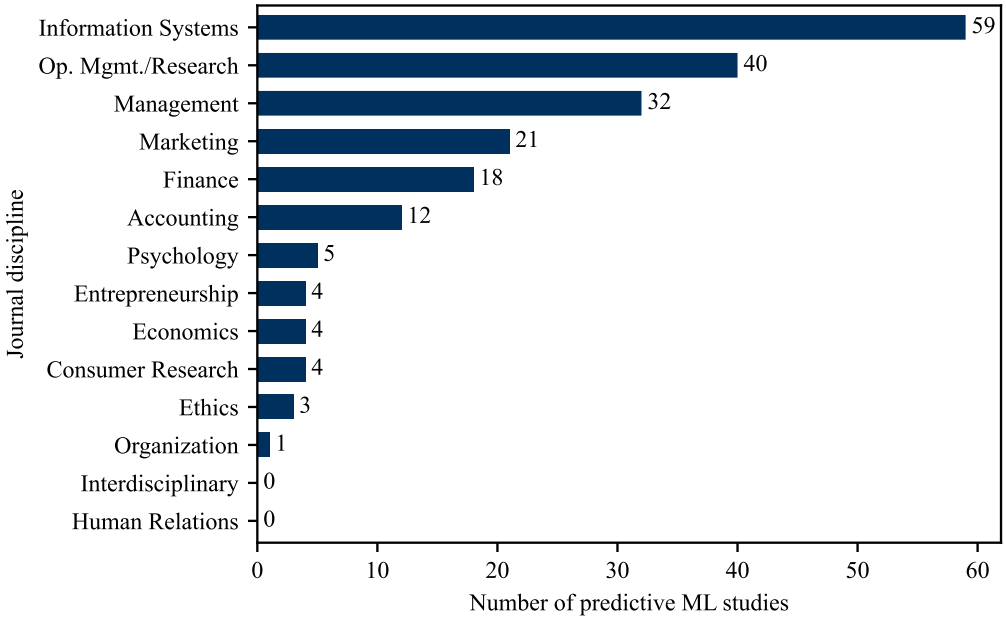

Figure 5: Number of predictive ML studies in our sample grouped by the research discipline of the journals in which they are published (n=203).

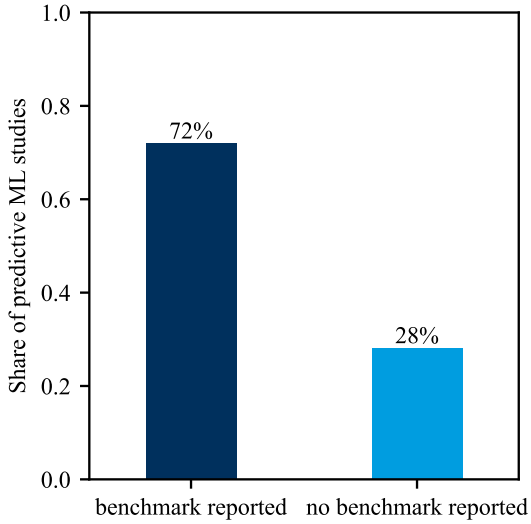

Figure 6: Share of predictive ML studies that report the predictive performance of at least one traditional benchmark model and those that do not report the predictive performance of traditional benchmark models (n=203).

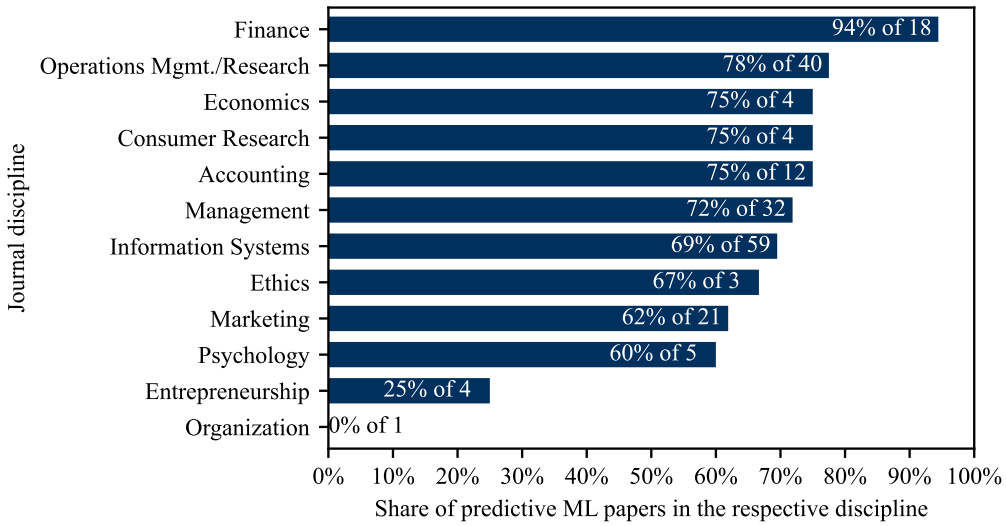

Figure 7: Share of predictive ML studies that report the predictive performance for at least one traditional benchmark model grouped by the research discipline of the journals in which they are published (n=203).

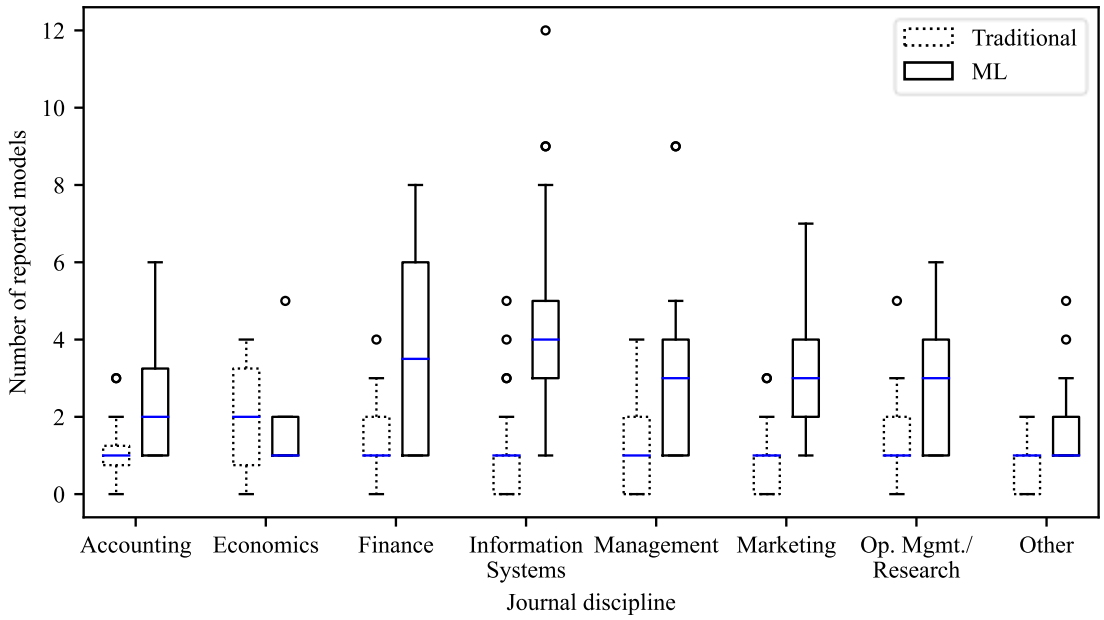

Figure 8: Number of traditional benchmark and ML models for which predictive ML studies report predictive performance grouped by the research discipline of the journals in which they are published. The group "Other" includes the research disciplines: Entrepreneurship, Ethics, Organization, Psychology, and Consumer Research.

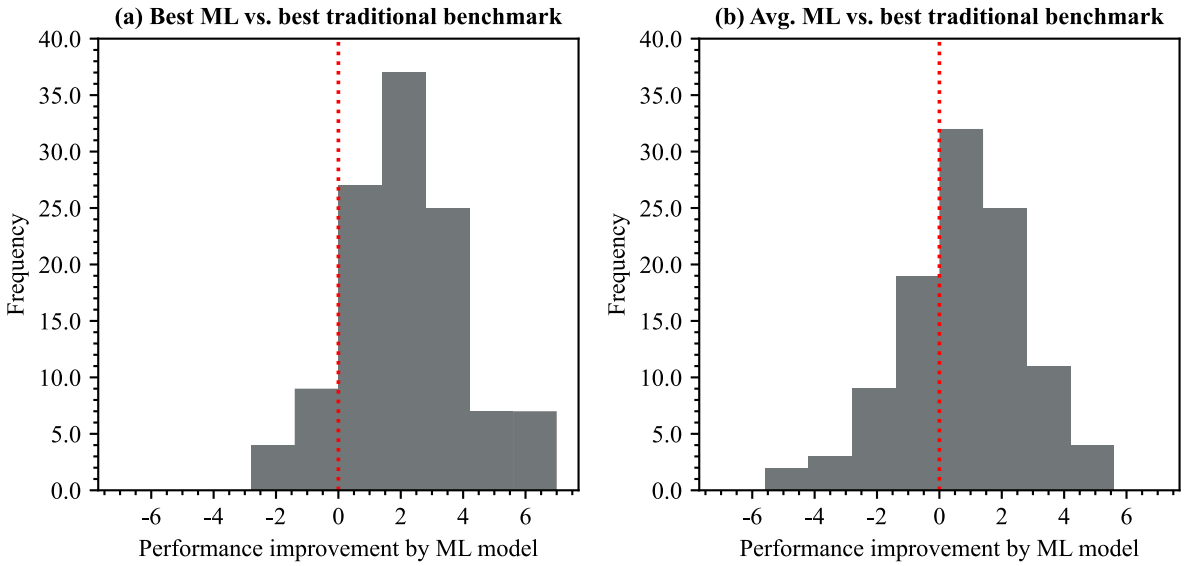

Figure 9: Absolute performance difference of the (a) best and (b) average ML model over the best-performing traditional benchmark model relative to the mean incremental performance increase as described in Eq. (2). The histogram in panel (a) includes the prediction results of 116 studies, while the histogram in panel (b) includes results of 105 studies since those that only report one ML model were removed. We only take studies into account that report the results for more than two models in order to calculate the mean incremental performance increase.

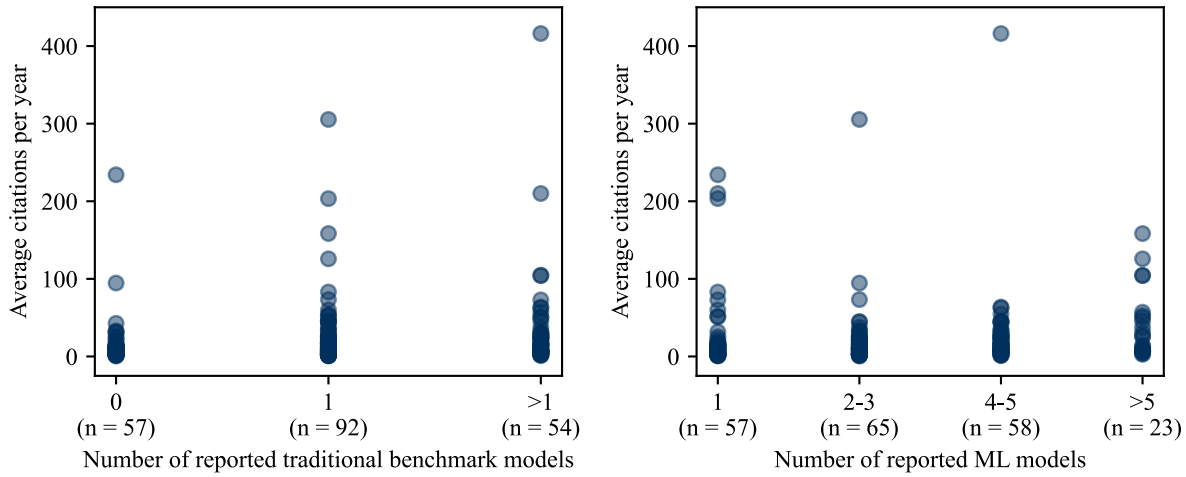

Figure 10: Average number of citations per year of predictive ML studies depending on the number of reported traditional benchmark (left panel) or respectively ML (right panel) models. The number of studies per group is denoted by n.

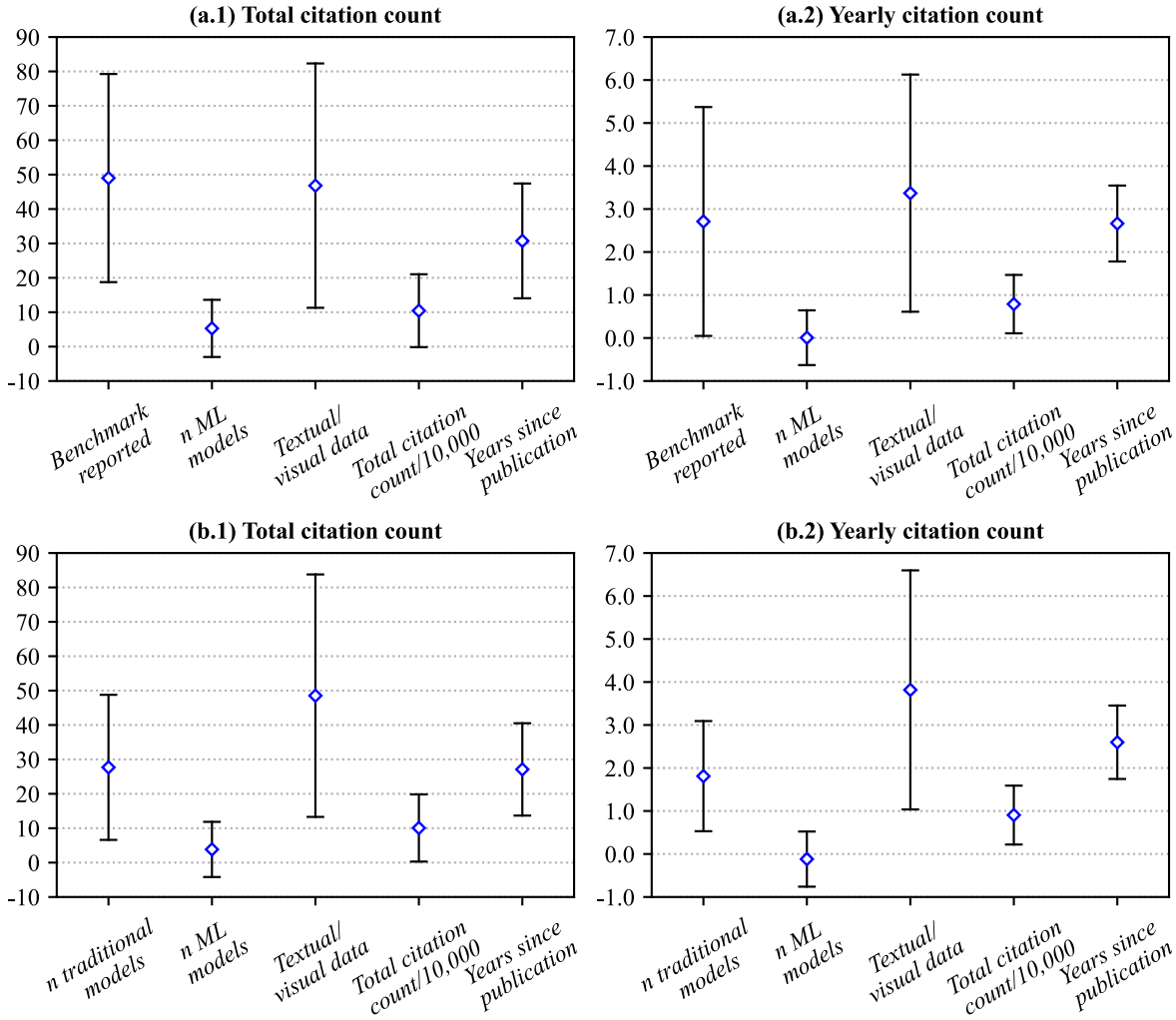

Figure 11: Confidence intervals of estimated average marginal effects (95%) for the total citation count (left plots) and the yearly citation count (right plots) based on the negative binomial regressions presented in Table 4. The upper plots (panel a) present the results if we measure the transparency of the reporting of traditional benchmark results with a binary variable (*Benchmark reported*). The lower plots (panel b) show the results if we use the number of reported traditional benchmark models (*n traditional models*) as the explanatory variable to estimate the relation of transparency and citation count. We omit the remaining, insignificant control variables for graphical illustration.

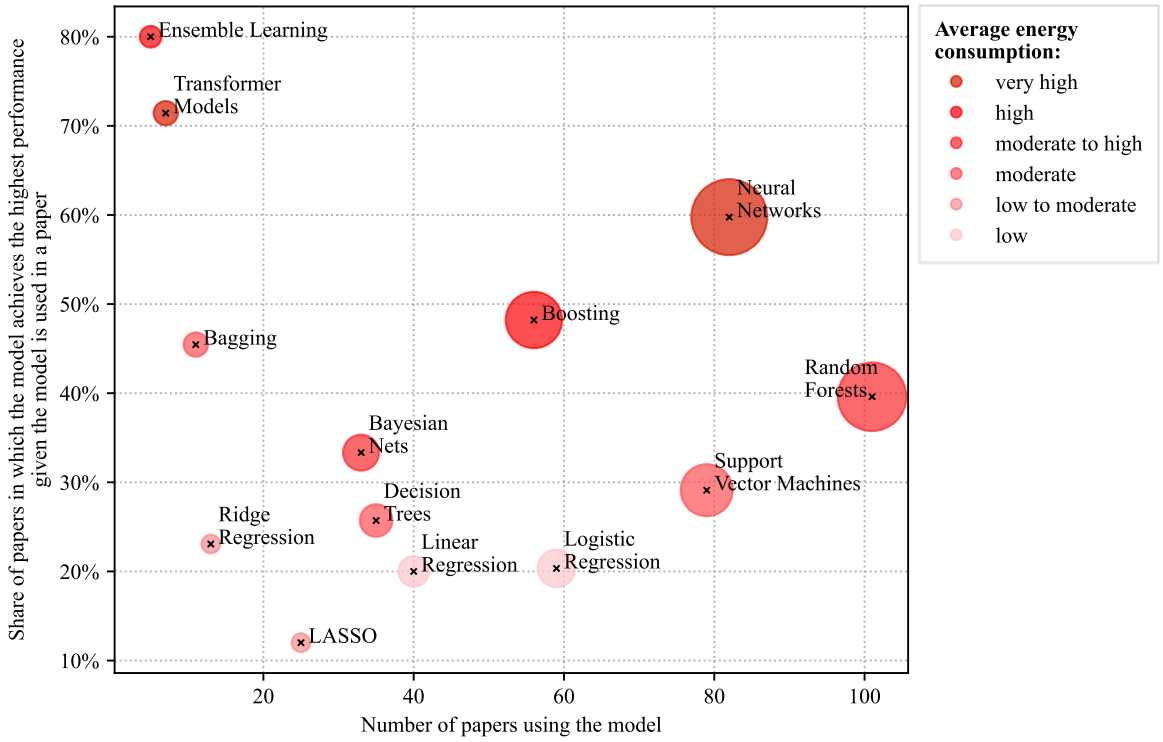

Figure 12: Models with the highest predictive performance per paper and assumed average energy consumption. This chart shows the share of papers in which a certain model type achieves the highest predictive performance (for at least one evaluation measure) of all models reported in a paper relative to all papers that use this model type. The size of the circles is determined by the absolute number of papers in which the model type achieves the highest predictive performance according to at least one evaluation measure. We define the average energy consumption per model type according to their general computational intensity from low (= simple models with minimal computational effort) to very high (= models with complex architectures, extensive parameters, requiring high-dimensional and large-scale data sets). This figure only includes the most used model types in our sample and papers that report the results for at least two models.

Tables

Table 1: Summary statistics of the identified predictive ML studies.

	Obs.	Mean	S.D.	Median	Min	Max
Dependent variables						
Transparency about the model performance						
Traditional benchmark	203	0.719	0.450	1.000	0.000	1.000
n traditional models	203	1.113	1.011	1.000	0.000	5.000
Performance difference						
Best ML vs. best benchmark	116	2.097	2.000	2.000	−2.769	7.000
Avg ML vs best benchmark	105	0.696	1.992	0.697	−5.487	4.933
Citation analysis						
Total citation count	203	105.409	208.129	37.000	1.000	1942.000
Yearly citation count	1,371	16.689	36.616	5.000	0.000	573.000
Paper-specific variables						
n ML models	203	3.182	2.032	3.000	1.000	12.000
Textual/visual data	203	0.424	0.495	0.000	0.000	1.000
Years since publication	1,371	1.207	3.018	1.000	−7.000	14.000
Publication year						
= 2010	2					
= 2011	1					
= 2012	1					
= 2013	1					
= 2014	2					
= 2015	4					
= 2016	10					
= 2017	3					
= 2018	7					
= 2019	16					
= 2020	29					
= 2021	28					
= 2022	57					
= 2023	42					
Author-specific variables						
Average citation count / 1000	203	4.394	11.162	1.840	0.002	146.706
Total citation count / 1000	203	11.887	19.680	4.659	0.002	146.706
Average FT50 publications	203	10.006	11.640	5.500	0.000	93.000
Average years since PhD	201	12.670	6.935	11.500	1.000	49.500
Team size	203	3.360	1.183	3.000	1.000	8.000
Journal-specific variables						
Journal impact factor	203	5.545	2.076	5.000	1.683	11.775

Notes: The dependent variables *traditional benchmark*, *n traditional models*, *best ML vs. best benchmark*, and *avg ML vs. best benchmark* are also used as explanatory variables in other regression models. Definitions of the variables are provided in Table A3.

Table 2: **Transparency about the predictive performance of traditional benchmark models (RQ2b).** This table shows coefficient estimates (1) and average marginal effects (2)–(4) for regressing a variable indicating whether a traditional benchmark was reported or not (1)–(2), the number of reported traditional benchmark models (3), and the number of reported ML models (4) on paper-, author-, and journal-specific variables.

	y: benchmark reported (1/0)		y: n traditional (I) / ML (II) models	
	(1) LPM	(2) Probit	(3) Poisson I	(4) Poisson II
Paper-specific variables				
Textual/visual data	−0.133*	−0.135**	−0.436***	0.715***
	(0.073)	(0.063)	(0.150)	(0.275)
n ML models	0.025	0.029	0.096***	
	(0.019)	(0.018)	(0.035)	
n traditional models				0.399***
				(0.140)
Author-specific variables				
Seniority				
Average citation count / 1000	−0.002	−0.002	−0.015	−0.000
	(0.004)	(0.003)	(0.020)	(0.014)
Average FT50 publications	0.005	0.005	0.012	−0.014
	(0.004)	(0.004)	(0.009)	(0.014)
Team size	−0.006	−0.010	−0.075	0.261**
	(0.032)	(0.029)	(0.057)	(0.115)
Journal-specific variables				
Journal impact factor	−0.009	−0.010	−0.008	−0.063
	(0.028)	(0.027)	(0.057)	(0.112)
Intercept	0.976***	Yes	Yes	Yes
	(0.239)			
Fixed effects				
Journal discipline	Yes	Yes	Yes	Yes
Publication Year	Yes	Yes	Yes	Yes
Observations	203	196	203	203
Adj. R^2	0.0002			
Pseudo R^2		0.1140		
AIC	273.84	257.78	558.92	807.77
LL	−109.92	−103.89	−250.46	−376.89

Notes: Fixed effects include the research discipline of the journal in which a study was published and the year the study was published. Asterisks indicate the significance of the estimated parameters at the ***1%, **5%, and *10% level. Robust standard errors are shown in parentheses.

Table 3: **Reported performance difference of ML and traditional benchmark models (RQ3b).** This table shows OLS coefficient estimates for regressing the performance difference of the best-performing ML (1)–(2) or, respectively, the average ML model (3)–(4) to the best-performing traditional benchmark model on paper-, author-, and journal-specific variables. The performance difference is measured relative to the mean incremental performance increase of all reported models in a study as described in Eq. (2).

	y: Best ML vs. best benchmark		y: Avg. ML vs. best benchmark	
	(1)	(2)	(3)	(4)
Paper-specific variables				
Textual/visual data	0.391	0.308	0.644	0.627
	(0.524)	(0.547)	(0.583)	(0.609)
n ML models	0.574***	0.548***	0.139	0.122
	(0.113)	(0.114)	(0.135)	(0.141)
n traditional models	−0.248	−0.320	−0.714***	−0.716***
	(0.252)	(0.237)	(0.242)	(0.267)
Author-specific variables				
Seniority				
Average citation count / 1000	0.166***	0.138***	0.178***	0.158***
	(0.057)	(0.047)	(0.065)	(0.054)
Average FT50 publications	−0.054*		−0.056*	
	(0.028)		(0.031)	
Average years since PhD		−0.074**		−0.069**
		(0.030)		(0.034)
Team size	0.407*	0.419*	0.443*	0.438*
	(0.215)	(0.223)	(0.240)	(0.249)
Journal-specific variables				
Journal impact factor	0.078	−0.017	0.171	0.069
	(0.149)	(0.156)	(0.173)	(0.195)
Intercept	0.372	4.134	−2.203	0.292
	(3.747)	(2.997)	(2.192)	(2.128)
Fixed effects				
Journal discipline	Yes	Yes	Yes	Yes
Prediction context	Yes	Yes	Yes	Yes
Performance measure	Yes	Yes	Yes	Yes
Publication Year	Yes	Yes	Yes	Yes
Observations	114	113	103	102
Adj. R^2	0.3970	0.3880	0.3890	0.3780
AIC	441.57	435.83	401.34	397.15
LL	−184.79	−182.91	−165.67	−164.58

Notes: Fixed effects include the research discipline of the journal in which a study was published, the thematic context of the variable that is predicted (see Table A.1), the type of measure that was used in the study to evaluate the predictive performance, and the year the study was published. Asterisks indicate the significance of the estimated parameters at the ***1%, **5%, and *10% level. Robust standard errors are shown in parentheses.

Table 4: **Citation analysis (RQ4a)**. This table shows average marginal effects for regressing the total (left panel) and yearly (right panel) citation count of predictive ML studies on the transparency about the performance of traditional benchmark models.

	y: Total citation count				*y*: Yearly citation count			
	(1) Poisson I	(2) NB I	(3) Poisson II	(4) NB II	(1) Poisson I	(2) NB I	(3) Poisson II	(4) NB II
Transparency about the model performance								
Benchmark reported	7.709	49.000***			6.783***	2.710**		
	(18.210)	(15.430)			(2.176)	(1.358)		
n traditional models			29.610*	27.690**			3.960**	1.811***
			(16.380)	(10.760)			(1.531)	(0.654)
Paper-specific controls								
n ML models	4.471	5.295	1.229	3.830	0.835	0.007	0.627	−0.119
	(3.537)	(4.239)	(3.628)	(4.091)	(0.593)	(0.325)	(0.574)	(0.327)
Textual/visual data	28.040	46.800**	44.090***	48.530***	6.811***	3.369**	7.092***	3.817***
	(19.430)	(18.120)	(16.050)	(17.980)	(2.521)	(1.407)	(2.506)	(1.418)
Author-specific controls								
Total citation count / 1000	0.941**	1.044*	1.255***	1.007**	0.146**	0.079**	0.142**	0.090**
	(0.385)	(0.541)	(0.469)	(0.498)	(0.073)	(0.035)	(0.068)	(0.035)
Team size	12.110*	7.511	12.030	9.480	1.180	0.311	1.417	0.338
	(7.182)	(7.883)	(7.306)	(7.707)	(1.165)	(0.620)	(1.133)	(0.610)
Journal-specific controls								
Journal impact factor	−1.302	−11.800	0.786	−10.040	−1.706	−0.025	−1.450	−0.048
	(5.681)	(7.451)	(5.954)	(7.162)	(1.045)	(0.427)	(1.004)	(0.422)
Other controls								
Years since publication	23.900***	30.730***	23.170***	27.100***	3.914***	2.663***	3.438***	2.599***
	(4.382)	(8.513)	(4.335)	(6.848)	(1.101)	(0.450)	(0.882)	(0.436)
Fixed effects								
Journal discipline	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Publication Year	No	No	No	No	Yes	Yes	Yes	Yes
Observations	203	203	203	203	1335	1335	1335	1335
AIC	19026.28	2127.82	17478.54	2125.41	11876.60	8095.19	11872.61	8090.68
LL	−9495.14	−1045.91	−8721.27	−1044.70	−5904.30	−4012.59	−5902.30	−4010.34

Notes: Fixed effects include the research discipline of the journal in which a study was published and the year the study was published. All models include an intercept. Panel regressions (y: yearly citation count) employ a random effects estimator. Asterisks indicate the significance of the estimated parameters at the ***1%, **5%, and *10% level. We report robust standard errors except for the negative binomial panel regressions. Standard errors are shown in parentheses.

Table 5: **Citation analysis (RQ4b)**. This table shows average marginal effects for regressing the total (left panel) and yearly (right panel) citation count of predictive ML studies on the reported performance difference between ML and traditional benchmark models.

	y: Total citation count				*y*: Yearly citation count			
	(1) Poisson I	(2) NB I	(3) Poisson II	(4) NB II	(1) Poisson I	(2) NB I	(3) Poisson II	(4) NB II
Performance improvement by ML models								
Best ML vs. best traditional	15.460**	9.732			1.430	0.239		
	(6.089)	(6.802)			(0.934)	(0.723)		
Avg ML vs. best traditional			10.650	7.800			1.165	0.402
			(6.675)	(7.431)			(1.031)	(0.767)
Paper-specific controls								
n traditional models	47.460***	11.820	45.880**	1.278	1.847	3.236**	0.355	2.541
	(17.200)	(13.740)	(22.090)	(16.650)	(1.964)	(1.526)	(2.390)	(1.789)
n ML models	0.039	−3.432	8.640	5.983	−0.378	−0.795	0.951	−0.337
	(6.470)	(7.031)	(7.482)	(6.616)	(0.976)	(0.754)	(0.930)	(0.778)
Textual/visual data	22.540	45.620	39.160	52.790	6.368	6.681*	7.422	5.810
	(32.820)	(32.870)	(36.770)	(35.620)	(4.542)	(3.552)	(5.015)	(3.561)
Author-specific controls								
Total citation count / 1000	−1.195	0.687	−1.101	0.944	0.094	0.123	0.128	0.118
	(0.817)	(0.960)	(0.965)	(0.993)	(0.133)	(0.110)	(0.137)	(0.114)
Team size	42.900***	26.090*	43.560***	20.970	3.651*	3.386**	3.019	3.079**
	(13.030)	(14.900)	(14.770)	(14.870)	(2.047)	(1.498)	(2.058)	(1.545)
Journal-specific controls								
Journal impact factor	−12.750	−5.978	−12.480	−5.474	−0.929	−0.240	−0.865	−0.068
	(9.395)	(6.978)	(8.720)	(7.144)	(0.992)	(0.822)	(1.030)	(0.849)
Other controls								
Years since publication	19.050***	33.000***	18.410***	32.460***	4.120***	2.862***	4.083***	2.733***
	(6.738)	(8.759)	(6.377)	(8.585)	(1.131)	(0.837)	(1.136)	(0.829)
Fixed effects								
Journal discipline	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Prediction context	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Performance measure	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Publication Year	No	No	No	No	Yes	Yes	Yes	Yes
Observations	114	114	103	103	756	756	681	681
AIC	7617.29	1256.58	6619.15	1135.32	7353.86	4830.41	6594.66	4346.44
LL	−3777.65	−597.29	−3280.57	−538.66	−3627.93	−2365.20	−3251.33	−2126.22

Notes: The performance improvement by ML models is calculated as the performance difference of the best-performing ML or, respectively, the average ML model to the best-performing traditional benchmark model relative to the mean incremental performance increase of all reported models in a study as described in Eq. (2). Fixed effects include the research discipline of the journal in which a study was published, the thematic context of the variable that is predicted (see Table A.1), the type of measure that was used in the study to evaluate the predictive performance, and the year the study was published. All models are estimated with an intercept. Panel regressions (y: yearly citation count) employ a random effects estimator. Asterisks indicate the significance of the estimated parameters at the ***1%, **5%, and *10% level. We use robust standard errors except for the negative binomial panel regressions. Standard errors are shown in parentheses.

Online Appendix

A. Supplementary Tables

Table A.1: Categories of the thematic context of the variable that is predicted in a study to group studies with research questions from similar topics, and the number (share) of predictive ML studies assigned to the categories.

Category	Description and examples	Number
Corporate Finance & Lending	Audit and corporate finance-related topics as well as crowdfunding like loan defaults, start-up valuations, survival rates, funding	24 (11.8%)
Customer Behavior	Predict the behavior or characteristics of customers such as purchase decisions, customer churn	29 (14.3%)
Cybersecurity	Detect cyber threats, e.g., phishing, malware, spam detection	10 (4.9%)
Employee Behavior	Predict the behavior or characteristics of employees such as turnover, performance, job search status, value of education	11 (5.4%)
Financial Markets	Predict financial market variables, e.g., asset (stocks, bonds, options) returns, volatility, financial risk	18 (8.9%)
Fraud	Detect fraud such as accounting fraud, financial fraud, fraudulent behavior	10 (4.9%)
Healthcare	Predict healthcare-related variables such as new infections of a disease, hospital length of stay, cancer, patient behavior	23 (11.3%)
Human Behavior	Predict the behavior of humans (if not assignable to customer or employee behavior) like outcome of bargaining, social preferences, emotions	8 (3.9%)
Sales, Price, & Demand	Predict sales and earnings of companies, prices (except for financial assets) or demand for goods and related topics such as the forecasting of macroeconomic variables	27 (13.3%)
Text & News Sentiment	Predict the sentiment of text, news, and reviews, and other related topics such as the detection of fake news	25 (12.3%)
Transportation & Logistics	Predicting variables that are related to transportation and logistics, e.g., travel-time, driving crashes, inventory	7 (3.4%)

Notes: This table describes the categories used to group studies with similar predictive research questions. We categorized 203 predictive ML studies of which 11 studies (5.4%) are assigned to the residual category *Other*. Note that we use the most specific category that can be assigned to a given research problem if it potentially would fit to multiple categories. For example, if a study is about predicting fraudulent behavior on crowdfunding platforms, it could be potentially assigned to the categories *Corporate Finance & Lending*, *Human Behavior*, and *Fraud.* The category that would describe the research problem most specifically would be *Fraud* in this case.

Table A.2: Journals listed in the Financial Times Research Rank (FT50) classified by their research discipline. The assigned research disciplines of the journals are used as fixed effects in the regression models.

Journal discipline	Assigned FT50 journals
Accounting	Accounting, Organizations and Society; Contemporary Accounting Research; Journal of Accounting and Economics; Journal of Accounting Research; Review of Accounting Studies; The Accounting Review
Consumer Research	Journal of Consumer Psychology; Journal of Consumer Research
Economics	American Economic Review; Econometrica; Journal of Political Economy; Quarterly Journal of Economics; Review of Economic Studies
Entrepreneurship	Entrepreneurship Theory and Practice; Journal of Business Venturing; Strategic Entrepreneurship Journal
Ethics	Journal of Business Ethics
Finance	Journal of Finance; Journal of Financial and Quantitative Analysis; Journal of Financial Economics; Review of Finance; Review of Financial Studies
Human Relations	Human Relations; Human Resource Management
Information Systems	Information Systems Research; Journal of Management Information Systems; MIS Quarterly
Management	Academy of Management Journal; Academy of Management Review; Harvard Business Review; Journal of Management; Journal of Management Studies; Management Science; Sloan Management Review; Strategic Management Journal
Marketing	Journal of Marketing; Journal of Marketing Research; Journal of the Academy of Marketing Science; Marketing Science
Interdisciplinary	Journal of International Business Studies; Research Policy
Operations Mgmt./Research	Journal of Operations Management; Manufacturing and Service Operations Management; Operations Research; Production and Operations Management
Organization Studies	Administrative Science Quarterly; Organization Science; Organization Studies; Organizational Behavior and Human Decision Processes
Psychology	Journal of Applied Psychology

Table A.3: Variable definitions.

Variable	Definition
Dependent variables	
Traditional benchmark	Binary variable indicating whether the predictive performance for at least one traditional statistical model is reported (=1) or not (=0)
n traditional models	The number of traditional statistical models for which the predictive performance is reported
Best ML vs. best benchmark	The absolute performance difference between the best ML and the best traditional statistical model relative to the mean incremental increase of all models reported in a paper as described in Eq. (2)
Avg ML vs. best benchmark	The absolute performance difference between the mean predictive performance of all ML models and the best traditional statistical model relative to the mean incremental increase of all models reported in a paper as described in Eq. (2)
Total citation count	The total number of citations on Google Scholar that a predictive ML study gathered as of August 26, 2024
Yearly citation count	The number of citations on Google Scholar per year that a predictive ML study gathered as of August 26, 2024
Explanatory and control variables	
Journal impact factor	The journal impact factor in the publication year according to Clarivate (2024)
n ML models	The number of ML models for which the predictive performance is reported
Seniority: average citation count	The total number of citations on Google Scholar averaged over the authors prior to the publication year without pre-publication citations of the predictive ML study
Seniority: average FT50 publications	The total number of publications in journals listed in the Financial Times Research Rank averaged over the authors prior to the publication year
Seniority: average years since PhD	The average difference in years between the publication year and the year in which the authors obtained their PhD
Team size	The number of authors
Textual/visual data	Binary variable indicating whether the prediction models use textual or visual (image, video) data as input (=1) or not (=0)
Total citation count	The sum of total citations on Google Scholar of all authors prior to the publication year without pre-publication citations of the predictive ML study
Years since publication	The number of years since the publication year. For cross-sectional data, the years since publication are measured by the number of years between 2024 and the publication year
Fixed effects	
Journal discipline	Overall research discipline of the journal as listed in Table A.2
Prediction context	The thematic context of the variable that is predicted as listed in Table A.1
Publication year	The year in which the study was published
Performance measure	The type of measure that was used to evaluate the models' predictive performance (confusion matrix, area under the curve, loss functions, out-of-sample R-squared, profit, or other)

Notes: The dependent variables *traditional benchmark*, *n traditional models*, *best ML vs. best benchmark*, and *avg ML vs. best benchmark* are also used as explanatory variables in other regression models.

Table A.4: Pearson correlations of all dependent, control, and explanatory variables used in the regression models presented in section 4.

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
(1)	*Traditional benchmark*	1														
(2)	*n traditional models*	.69	1													
(3)	*Best ML vs. best benchmark*		−.15	1												
(4)	*Avg ML vs best benchmark*		−.29	.84	1											
(5)	*Total citation count*	.06	.21	.09	.01	1										
(6)	*Yearly citation count*	−.01	.11	.11	.06	.68	1									
(7)	*n ML models*	.08	.13	.48	.05	.02	.00	1								
(8)	*Textual/visual data*	−.15	−.21	.11	.07	−.02	.01	.12	1							
(9)	*Average citation count / 1000*	.02	−.04	.21	.25	.07	.06	−.03	.04	1						
(10)	*Total citation count / 1000*	−.07	−.12	.18	.18	.20	.16	.03	.02	.78	1					
(11)	*Average FT50 publications*	.08	.04	.04	.12	.05	.04	−.05	.00	.66	.55	1				
(12)	*Average years since PhD*	.05	−.04	−.02	.01	−.04	−.06	.02	−.12	.27	.40	.45	1			
(13)	*Team size*	.00	−.04	.18	.16	.03	−.02	.14	.04	.00	.25	−.02	.11	1		
(14)	*Journal impact factor*	−.07	−.09	.06	.11	−.05	.01	.08	.01	.18	.27	−.06	.04	.16	1	
(15)	*Years since publication*	−.02	.04	−.12	−.19	.17	.24	−.10	.00	−.04	−.06	−.02	−.00	−.05	−.32	1

Notes: A grey cell indicates that these variables do not occur together as covariates in any of our regression models.

B. Supplementary Figures

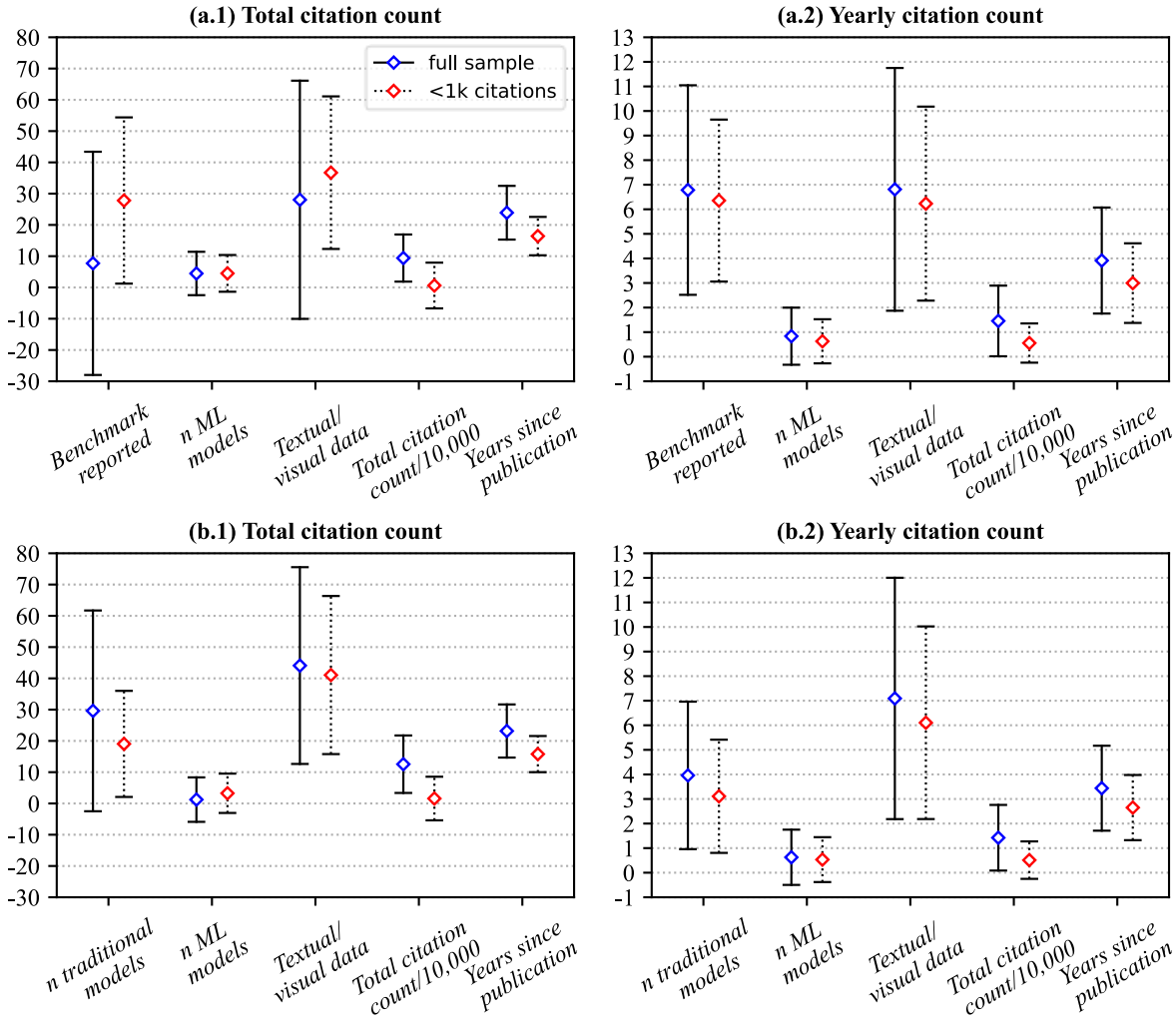

Figure B.1: Confidence Intervals of estimated average marginal effects (95%) for the total citation count (left plots) and the yearly citation count (right plots) based on the Poisson regressions presented in Table 4. The upper plots (panel a) present the results if we measure the transparency of the reporting of traditional benchmark results with a binary variable (*Benchmark reported*). The lower plots (panel b) show the results if we use the number of reported traditional benchmark models (*n traditional models*) as the explanatory variable to estimate the relation of transparency and citation count. Due to robustness reasons, the average marginal effects are presented for using the full sample of studies (solid line) and only for studies with less than 1,000 citations (dotted line). We omit the remaining, insignificant control variables for graphical illustration.

C. Robustness to a different classification of models into machine learning and traditional statistical models

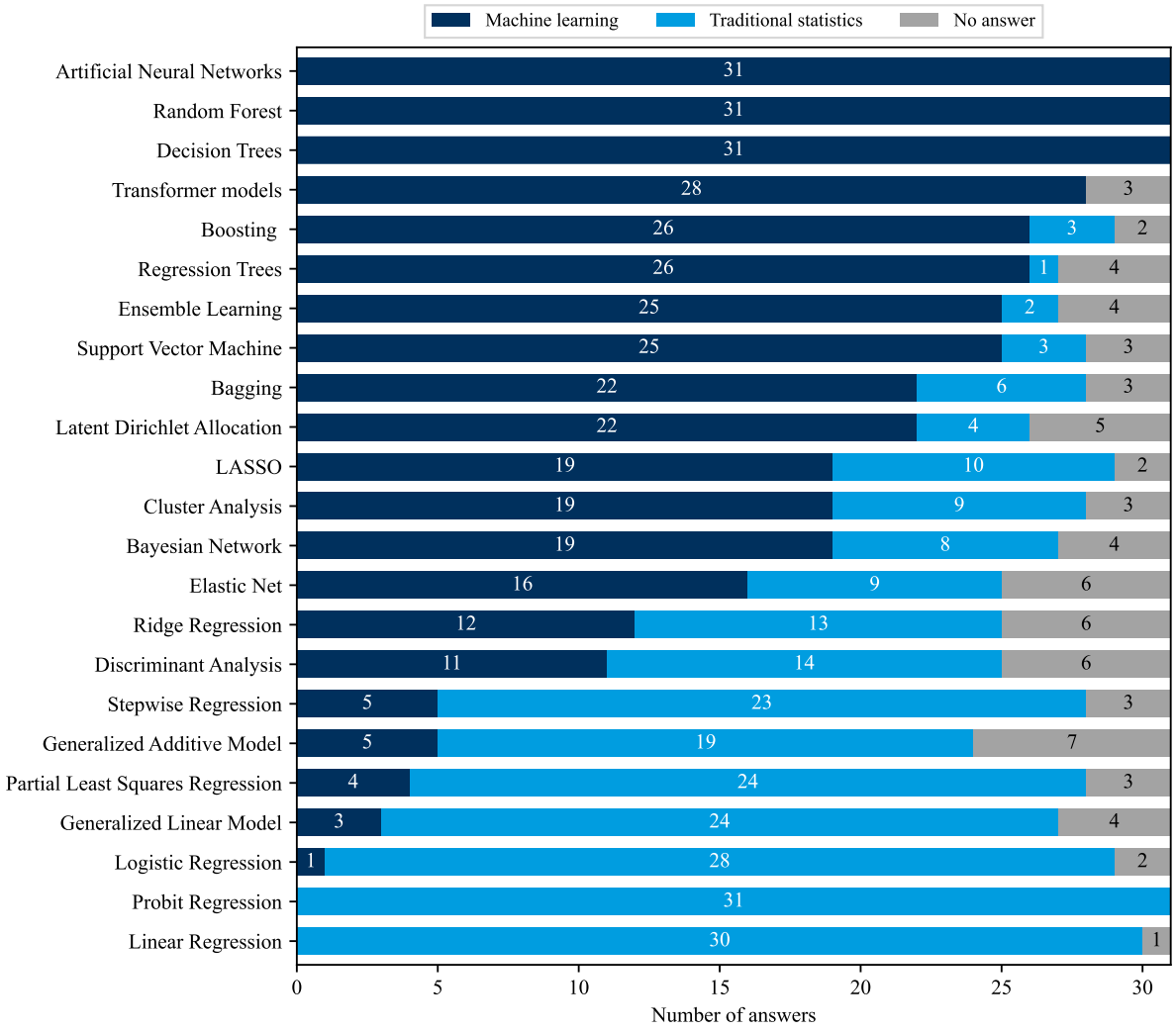

Figure C.1: Classification of the most used models in our sample into ML and traditional statistical models by survey respondents. In total, 31 authors of the studies in our sample or of studies that discuss the use of ML in business and economic research participated in our survey.

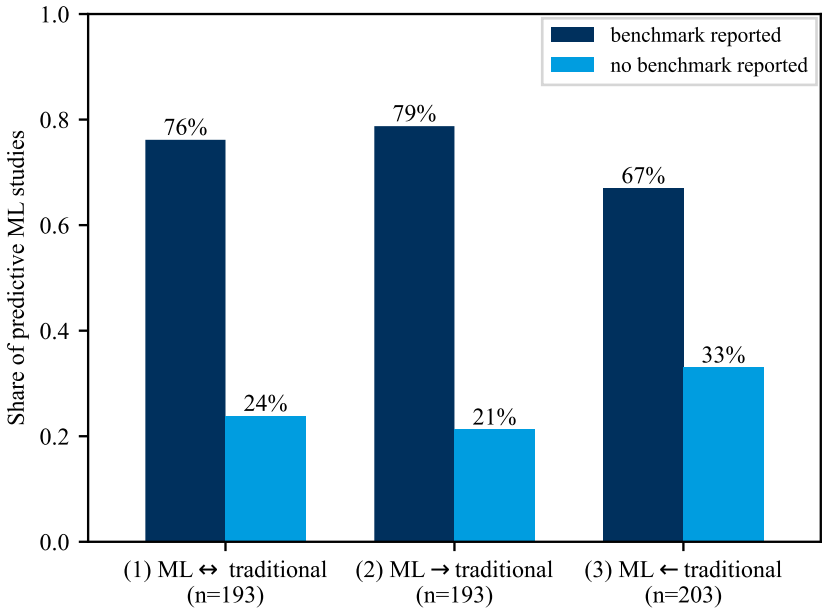

Figure C.2: Share of predictive ML studies that report the predictive performance of at least one traditional benchmark model and those that do not report the predictive performance of traditional benchmark models if (1) we classify all ambiguous models in the opposite direction to our initial classification, (2) we classify all ambiguous models as traditional statistical models, and (3) we classify all ambiguous models as ML. Ambiguous models according to the results of our survey (Figure C.1) include Bayesian networks, elastic nets, general additive models, cluster analysis, discriminant analysis, LASSO, and ridge regression.

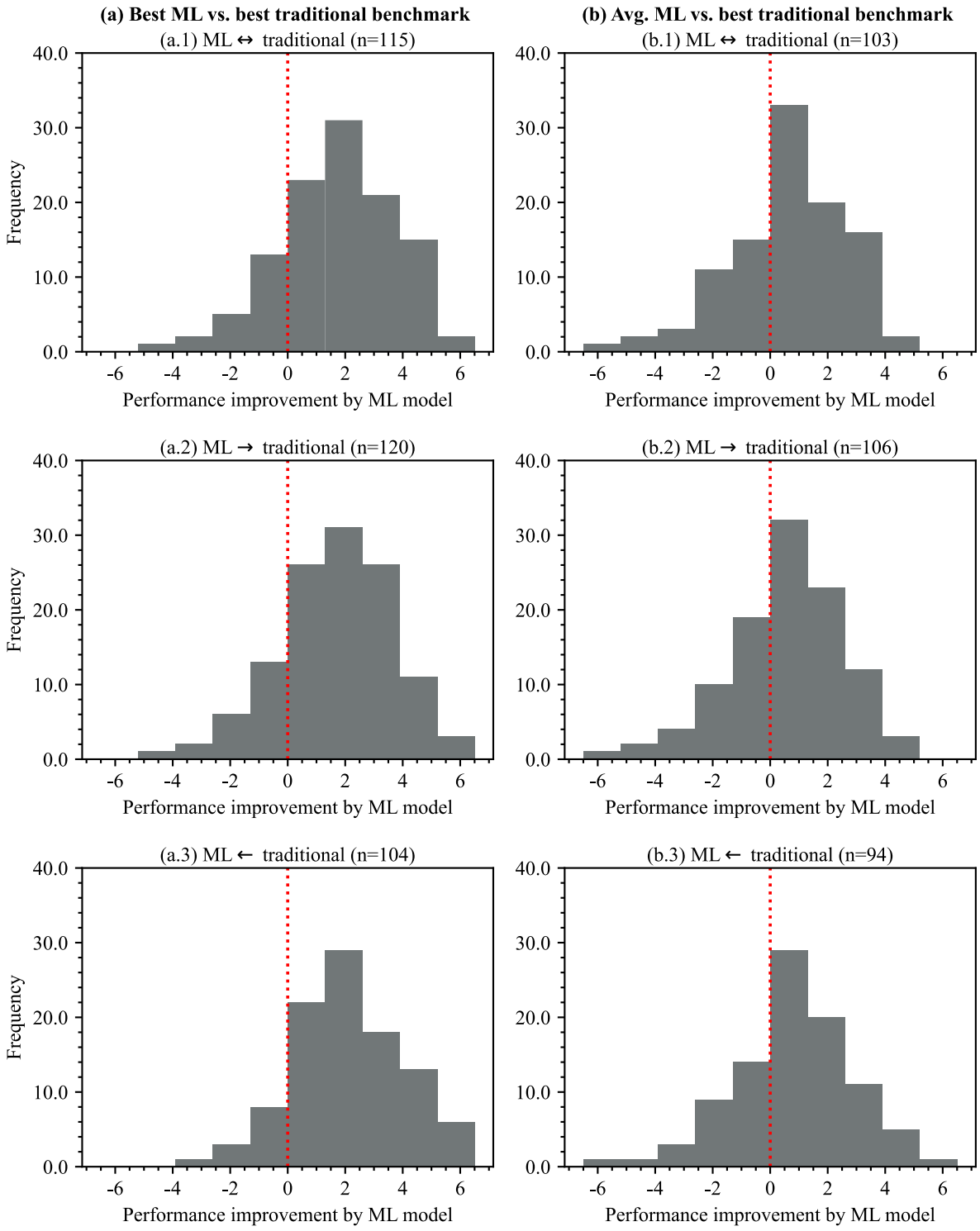

Figure C.3: Absolute performance difference of the (a) best and (b) average ML model over the best-performing traditional benchmark model relative to the mean incremental performance increase as described in Eq. (2) if (1) we classify all ambiguous models in the opposite direction to our initial classification, (2) we classify all ambiguous models as traditional statistical models, and (3) we classify all ambiguous models as ML. Ambiguous models according to the results of our survey (Figure C.1) include Bayesian networks, elastic nets, general additive models, cluster analysis, discriminant analysis, LASSO, and ridge regression. We only take studies into account that report the results for more than two models in order to calculate the mean incremental performance increase.

Table C.1: Robustness of the results regarding the transparency about the predictive performance of traditional benchmark models (RQ2b, Table 2) to alternative classifications of prediction models. This table shows coefficient estimates (1) and average marginal effects (2)–(4) for regressing a variable indicating whether a traditional benchmark was reported or not (1)–(2), the number of reported traditional benchmark models (3), and the number of reported ML models (4) on paper-, author-, and journal-specific variables.

	y: benchmark reported (1/0)						y: n traditional (I) / ML (II) models					
	(a) ML ↔ traditional		**(b) ML → traditional**		**(c) ML ← traditional**		**(a) ML ↔ traditional**		**(b) ML → traditional**		**(c) ML ← traditional**	
	(1) LPM	(2) Probit	(1) LPM	(2) Probit	(1) LPM	(2) Probit	(1) Poisson I	(2) Poisson II	(1) Poisson I	(2) Poisson II	(1) Poisson I	(2) Poisson II
Textual/visual data	−0.093	−0.104	−0.120*	−0.127**	−0.120	−0.122*	−0.490**	0.668**	−0.499**	0.683***	−0.440***	0.661**
	(0.073)	(0.064)	(0.071)	(0.060)	(0.078)	(0.068)	(0.194)	(0.303)	(0.202)	(0.256)	(0.147)	(0.328)
n ML models							0.072		0.151***		0.027	
							(0.045)		(0.053)		(0.029)	
n traditional models								0.205*		0.326***		0.165
								(0.116)		(0.103)		(0.150)
Team size								0.242**		0.243**		0.249*
								(0.119)		(0.104)		(0.130)
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Intercept	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	193	184	193	184	203	196	193	193	193	193	203	203
Adj. R^2	−0.0113		0.0018		−0.0038							
Pseudo R^2		0.1144		0.1361		0.0920						
AIC	242.96	229.29	224.70	214.45	293.07	276.15	596.78	756.24	617.03	724.08	529.93	840.51
LL	−94.48	−90.65	−85.35	−83.22	−119.53	−113.07	−268.39	−351.12	−278.52	−335.04	−235.97	−393.25

Notes: This table shows coefficient estimates for (a) classifying all ambiguous models in the opposite direction to our initial classification, (b) classifying all ambiguous models as traditional statistical models, and (c) classifying all ambiguous models as ML. Ambiguous models according to the results of our survey (Figure C.1) include Bayesian networks, elastic nets, general additive models, cluster analysis, discriminant analysis, LASSO, and ridge regression. This table only displays the coefficient estimates and corresponding robust standard errors in parentheses for variables with *p*-values below .1 and for variables with *p*-values below .1 in the baseline regression (Table 2). All other variables are included in the regression model. Other controls include: *average citation count, average FT50 publications, journal impact factor*. Fixed effects include the research discipline of the journal in which a study was published and the year the study was published. Asterisks indicate the significance of the estimated parameters at the ***1%, **5%, and *10% level.

Table C.2: Robustness of the results regarding the reported performance difference of ML and traditional benchmark models (RQ3b, Table 3) to alternative classifications of prediction models. This table shows OLS coefficient estimates for regressing the performance difference of the best-performing ML (1)–(2) or, respectively, the average ML model (3)–(4) to the best-performing traditional benchmark model on paper-, author-, and journal-specific variables. The performance difference is measured relative to the mean incremental performance increase of all reported models in a study as described in Eq. (2).

	y: Best ML vs. best benchmark						y: Avg. ML vs. best benchmark					
	(a) ML ↔ traditional		**(b) ML → traditional**		**(c) ML ← traditional**		**(a) ML ↔ traditional**		**(b) ML → traditional**		**(c) ML ← traditional**	
	(1)	(2)	(1)	(2)	(1)	(2)	(3)	(4)	(3)	(4)	(3)	(4)
n ML models	0.614***	0.620***	0.518***	0.531***	0.585***	0.570***						
	(0.128)	(0.131)	(0.118)	(0.120)	(0.133)	(0.329)						
n traditional models							−0.407	−0.428	−0.182	−0.220	−0.906**	−0.823**
							(0.320)	(0.332)	(0.229)	(0.235)	(0.343)	(0.357)
Avg citation count /1000	0.016	0.015	0.033	0.022	0.210***	0.198***	0.027	0.019	0.029	0.013	0.247***	0.227***
	(0.047)	(0.047)	(0.044)	(0.042)	(0.065)	(0.058)	(0.053)	(0.052)	(0.049)	(0.047)	(0.068)	(0.055)
Avg FT50 publications	−0.010		−0.014		−0.471		−0.018		−0.021		−0.068**	
	(0.022)		(0.022)		(0.308)		(0.025)		(0.027)		(0.031)	
Avg years since PhD		−0.019		−0.015		−0.079**		−0.022		−0.016		−0.092***
		(0.030)		(0.029)		(0.036)		(0.033)		(0.033)		(0.033)
Team size	0.516**	0.508**	0.537**	0.529**	0.379	0.379	0.462*	0.444*	0.479**	0.466**	0.415	0.403
	(0.229)	(0.234)	(0.207)	(0.211)	(0.256)	(0.258)	(0.237)	(0.242)	(0.226)	(0.231)	(0.262)	(0.266)
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Intercept	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	115	114	120	119	99	98	103	102	106	105	93	92
Adj. R^2	0.4138	0.4025	0.4175	0.4040	0.4457	0.4453	0.2447	0.2288	0.263	0.244	0.3937	0.3933
AIC	446.74	442.19	454.15	450.29	390.10	382.81	412.80	408.71	425.27	421.85	366.36	360.22
LL	−189.37	−188.10	−192.08	−191.15	−161.05	−158.41	−172.40	−171.35	−177.63	−176.92	−149.18	−147.11

Notes: This table shows coefficient estimates for (a) classifying all ambiguous models in the opposite direction to our initial classification, (b) classifying all ambiguous models as traditional statistical models, and (c) classifying all ambiguous models as ML. Ambiguous models according to the results of our survey (Figure C.1) include Bayesian networks, elastic nets, general additive models, cluster analysis, discriminant analysis, LASSO, and ridge regression. This table only displays the coefficient estimates and corresponding robust standard errors in parentheses for variables with *p*-values below .1 and for variables with *p*-values below .1 in the baseline regression (Table 3). All other variables are included in the regression model. Other controls include: *textual/visual data*, *journal impact factor*. Fixed effects include the research discipline of the journal in which a study was published, the thematic context of the variable that is predicted (see Table A.1), the type of measure that was used in the study to evaluate the predictive performance, and the year the study was published. Asterisks indicate the significance of the estimated parameters at the ***1%, **5%, and *10% level.

Table C.3: Robustness of the results regarding the citation analysis (RQ4a, Table 4) to alternative classifications of prediction models. This table shows average marginal effects for regressing the total (left panel) and yearly (right panel) citation count of predictive ML studies on the transparency about the performance of traditional benchmark models.

	y: Total citation count				*y*: Yearly citation count			
	(1) Poisson I	(2) NB I	(3) Poisson II	(4) NB II	(1) Poisson I	(2) NB I	(3) Poisson II	(4) NB II
Panel A: ML ↔ traditional								
Benchmark reported	10.306	40.534**			5.566**	1.974		
	(20.732)	(16.694)			(2.350)	(1.468)		
n traditional models			17.806	24.926**			3.524**	0.942*
			(12.065)	(9.873)			(1.397)	(0.519)
Controls and fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	193	193	193	193	1,273	1,273	1,273	1,273
AIC	18222.66	2033.56	17346.13	2027.99	11477.83	7776.56	11469.87	7774.90
LL	−9093.33	−998.78	−8655.06	−995.99	−5704.92	−3853.28	−5700.94	−3852.45
Panel B: ML → traditional								
Benchmark reported	11.597	51.537***			7.080***	3.270**		
	(21.613)	(16.092)			(2.271)	(1.472)		
n traditional models			19.247	25.474***			3.594**	1.153**
			(11.738)	(9.354)			(1.322)	(0.514)
Controls and fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	193	193	193	193	1,273	1,273	1,273	1,273
AIC	18228.93	2031.80	17267.53	2027.37	11475.67	7773.93	11469.09	7773.25
LL	−9096.47	−997.90	−8615.76	−995.69	−5703.83	−3851.96	−5700.54	−3851.63
Panel C: ML ← traditional								
Benchmark reported	3.810	41.017***			5.642**	2.184*		
	(15.870)	(15.891)			(2.236)	(1.307)		
n traditional models			28.571	27.750**			3.962**	1.465**
			(17.628)	(11.505)			(1.639)	(0.653)
Controls and fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	203	203	203	203	1,335	1,335	1,335	1,355
AIC	19023.01	2129.52	17578.43	2125.65	11878.78	8096.04	11872.98	8093.52
LL	−9493.50	−1046.76	−8771.22	−1044.83	−5905.39	−4013.02	−5902.49	−4011.76

Notes: This table shows coefficient estimates for (Panel A) classifying all ambiguous models in the opposite direction to our initial classification, (Panel B) classifying all ambiguous models as traditional statistical models, and (Panel C) classifying all ambiguous models as ML. Ambiguous models according to the results of our survey (Figure C.1) include Bayesian networks, elastic nets, general additive models, cluster analysis, discriminant analysis, LASSO, and ridge regression. Fixed effects include the research discipline of the journal in which a study was published (all regression models) and the year the study was published (only panel regressions; y: yearly citation count). All models include an intercept. Controls include: *n ML models, textual/visual data, total citation count, team size, journal impact factor, years since publication*. Panel regressions (y: yearly citation count) employ a random effects estimator. Asterisks indicate the significance of the estimated parameters at the ***1%, **5%, and *10% level. We report robust standard errors except for the negative binomial panel regressions. Standard errors are shown in parentheses.

Table C.4: Robustness of the results regarding the citation analysis (RQ4b, Table 5) to alternative classifications of prediction models. This table shows average marginal effects for regressing the total (left panel) and yearly (right panel) citation count of predictive ML studies on the reported performance difference between ML and traditional benchmark models.

	y: Total citation count				*y*: Yearly citation count			
	(1) Poisson I	(2) NB I	(3) Poisson II	(4) NB II	(1) Poisson I	(2) NB I	(3) Poisson II	(4) NB II
Panel A: ML ↔ traditional								
Best ML vs. best traditional	12.742*	11.212			1.558	0.967		
	(7.011)	(7.179)			(0.977)	(0.684)		
Avg ML vs. best traditional			1.120	7.056			0.969	0.458
			(4.983)	(6.175)			(0.829)	(0.598)
Controls and fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	115	115	103	103	766	766	685	685
AIC	6730.70	1245.51	3895.96	1115.10	7467.13	4878.59	6764.34	4358.28
LL	−3335.35	−594.76	−1919.98	−529.55	−3683.56	−2388.29	−3335.17	−2131.14
Panel B: ML → traditional								
Best ML vs. best traditional	18.290**	14.139*			2.035**	1.332*		
	(7.454)	(7.413)			(1.019)	(0.727)		
Avg ML vs. best traditional			6.082	11.190*			1.599*	0.915
			(5.266)	(6.345)			(0.871)	(0.626)
Controls and fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	120	120	106	106	795	795	695	695
AIC	6906.12	1300.35	4211.43	1151.34	7703.88	5063.68	6625.75	4406.68
LL	−3422.06	−621.17	−2076.72	−546.67	−3801.94	−2480.84	−3265.88	−2155.34
Panel C: ML ← traditional								
Best ML vs. best traditional	0.222	3.036			0.412	−0.060		
	(7.363)	(7.530)			(1.041)	(0.741)		
Avg ML vs. best traditional			0.521	3.655			0.502	0.173
			(6.047)	(7.959)			(1.107)	(0.727)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	99	99	93	93	646	646	605	605
AIC	5626.77	1078.83	4934.11	1012.40	6195.71	4100.25	5708.86	3806.05
LL	−2784.38	−511.41	−2439.06	−478.20	−3048.85	−2000.13	−2808.43	−1856.02

Notes: This table shows coefficient estimates for (Panel A) classifying all ambiguous models in the opposite direction to our initial classification, (Panel B) classifying all ambiguous models as traditional statistical models, and (Panel C) classifying all ambiguous models as ML. Ambiguous models according to the results of our survey (Figure C.1) include Bayesian networks, elastic nets, general additive models, cluster analysis, discriminant analysis, LASSO, and ridge regression. The performance improvement by ML models is calculated as the performance difference of the best-performing ML or, respectively, the average ML model to the best-performing traditional benchmark model relative to the mean incremental performance increase of all reported models in a study according to Eq. (2). Fixed effects include the research discipline of the journal in which a study was published, the thematic context of the variable that is predicted, the type of measure that was used to evaluate predictive performance (all regression models), and the year the study was published (only panel regressions; y: yearly citation count). All models include an intercept. Controls include: *n traditional models*, *n ML models*, *textual/visual data*, *total citation count*, *team size*, *journal impact factor*, *years since publication*. Panel regressions (y: yearly citation count) employ a random effects estimator. Asterisks indicate the significance of the estimated parameters at the ***1%, **5%, and *10% level. We report robust standard errors except for the negative binomial panel regressions. Standard errors are shown in parentheses.