DOI: 10.1111/eufm.12408

ORIGINAL ARTICLE

EUROPEAN FINANCIAL MANAGEMENT WILEY

Machine learning methods in finance: Recent applications and prospects

Daniel Hoang | **Kevin Wiegratz**

Institute for Finance, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Correspondence
Daniel Hoang, Institute for Finance, Karlsruhe Institute of Technology, Kaiserstr. 12, 76131 Karlsruhe, Germany.
Email: daniel.hoang@kit.edu

Abstract

We study how researchers can apply machine learning (ML) methods in finance. We first establish that the two major categories of ML (supervised and unsupervised learning) address fundamentally different problems than traditional econometric approaches. Then, we review the current state of research on ML in finance and identify three archetypes of applications: (i) the construction of superior and novel measures, (ii) the reduction of prediction error, and (iii) the extension of the standard econometric toolset. With this taxonomy, we give an outlook on potential future directions for both researchers and practitioners. Our results suggest many benefits of ML methods compared to traditional approaches and indicate that ML holds great potential for future research in finance.

KEYWORDS

artificial intelligence, big data, machine learning

JEL CLASSIFICATION

C45, G00

We appreciate helpful comments and suggestions made by John A. Doukas (the editor), two anonymous referees, Renée Adams, Andreas Benz, Francesco D'Acunto, Martin Ruckes, Fabian Silbereis, Michael Weber, and participants at the 2022 European Conference of the Financial Management Association (Lyon).

1 | INTRODUCTION

Artificial intelligence is increasingly entering our day-to-day life with impressive applications: face detection enables safe and efficient airport travel, voice recognition allows for seamless communication with personal assistants on smartphones and smart home devices, and ever more firms are using chatbots for quick customer support. Almost everyone interacts with modern artificial intelligence many times per day.

The main technology behind artificial intelligence is machine learning (ML). ML methods enable machines to conduct such complex tasks as detecting faces, understanding speech, or answering messages. Given the power of ML technology, it is natural to ask whether ML methods can also be applied elsewhere. This paper addresses the use of ML to solve problems in finance research.

Several overview papers indicate the potential of ML in finance. Varian (2014) describes ML as an appropriate tool in the economic analysis of big data and presents some ML methods with examples in economics. He further hints at potential ML applications in econometrics. Mullainathan and Spiess (2017) identify prediction problems as the main use case of ML in economics and present different categories of existing and potential future applications. Athey and Imbens (2019) illustrate the most relevant ML methods from an econometric perspective. They also provide an overview of ML's potential beyond pure prediction, especially for causality in economic questions.

While the usage of ML in finance research is still in its infancy, the number of applications that exploit the potential of ML has grown tremendously over the last few years. In 2018, the number of ML publications more than tripled compared to the yearly average of the years 2010 to 2017. In 2019, the increase was already more than fivefold. In 2020, the increase was almost sevenfold, and in 2021, there were almost 11 times as many publications using ML than before. Even though the universe of ML applications in finance has greatly expanded recently, it is still mostly unclear where and how to apply ML to solve research problems in finance.

The contribution of this paper is threefold. First, we present a high-level primer on ML for financial economists. We illuminate the different types of ML, their purposes and functionalities, and the available methods for each type. Given our focus on finance, we place special emphasis on the difference between traditional econometric methods and ML. We also demonstrate the benefits of ML over traditional linear methods (particularly for prediction problems) by applying ML to a high-dimensional asset pricing problem in finance. Our introduction allows researchers in the field to quickly grasp the essentials of ML that are relevant for applications in finance without assuming any prior knowledge of ML.

Second, we construct a taxonomy of current and future ML applications in finance. Given the increasing number of recent studies, earlier classifications do not capture existing applications well. We review the up-to-date literature in the field and divide it into three distinct archetypes. Our taxonomy allows researchers to better understand the current state of the literature and how different contributions relate to each other. Furthermore, it serves as guidance for future ML applications in finance.

Third, we study future prospects of ML applications in finance. We systematically analyse ML applications in finance and how their publication success differs by research field (asset pricing, corporate finance, financial intermediation, household finance) and application type. Our results not only suggest a high potential for ML applications in general but also provide researchers with indications of the most promising future directions.

Traditional econometrics aims to provide causal explanations for economic phenomena by analysing relationships between economic variables. ML, in contrast, allows researchers to obtain unique insights from high-dimensional data. There are two major types of high-dimensional data for which ML offers benefits over traditional methods such as linear regression. First, ML can deal with *high-dimensional, numerical data*, that is, data consisting of a high number of variables relative to the number of observations. Such high-dimensional data arises if there is a plethora of economically relevant variables or if nonlinearities and interaction effects play an important role. ML methods leverage the informational content of such data for predictions with small out-of-sample prediction errors. Second, in contrast to traditional methods, ML allows the exploitation of *unconventional data* (such as text, images, or videos), which are inherently high-dimensional. ML methods can extract economically relevant information from such data, which then serves as a starting point for further economic analyses.

ML is strongly related to the concept of big data. Big data consists of a high number of observations, a high number of variables, or both (Stock & Watson, 2020, p. 515). In general, data with a high number of observations improve the accuracy of ML predictions (in a similar way to how they improve the precision of parameter estimates of ordinary least squares [OLS] regressions). If the data exhibit a high number of variables (relative to the number of observations), ML outperforms simpler, traditional methods such as linear regression. Applying ML to data with high numbers of observations *and* variables combines both benefits as it can yield high prediction accuracy as well as outperformance over traditional methods.

Based on our review of the finance literature, we classify ML applications into three distinct archetypes: (1) construction of superior and novel measures, (2) reduction of prediction error in economic prediction problems, and (3) extension of the existing econometric toolset.

First, researchers can use ML to construct *superior* and *novel* measures. For instance, when applied to exploit unconventional data, the extracted information can serve as a superior or novel measure of an economic variable. *Superior* ML measures may exhibit lower measurement error and, therefore, can enable more precise estimates of economic relationships than traditional measures can. *Novel* ML measures enable analyses with previously unmeasurable economic variables.

Second, researchers can use ML to reduce prediction error in economic prediction problems. For instance, the fundamental problem of pricing financial or real assets is the prediction of adequate market prices. Given that a main functionality of ML is prediction, ML methods can provide better results than traditional approaches in solving such economic prediction problems.

Third, researchers can use ML to extend the existing econometric toolset. Econometric tools often contain a prediction component. For instance, the first stage of an instrumental variable design is effectively a prediction problem. ML methods can enhance such existing econometric tools by improving the performance of their prediction component. Furthermore, some ML methods themselves directly serve as new econometric tools. For instance, ML-based clustering methods extend the set of existing clustering methods from econometrics.

To demonstrate the benefits of ML over traditional methods at a typical prediction problem, we apply ML to real estate asset pricing, which is particularly relevant in the areas of household finance and real estate economics.[1] Real estate asset pricing is an inherent high-dimensional problem due to the large number of property characteristics, nonlinearities, and interaction

[1]Our exemplary application cannot yield generalisable results about the performance of ML compared to traditional methods, but illustrates how to apply ML to a typical problem in finance with high-dimensional data.

effects (for instance, a kitchen's marginal value likely interacts with house type, e.g., luxury apartment vs. standard single-family house.) We predict real estate asset prices in the German residential housing market using various ML methods (which exploit the large number of individual property characteristics in our data set) and compare their accuracy with estimates from traditional hedonic pricing (linear regression with the OLS estimator). Figure 1 illustrates our key results. The two charts compare the actual property prices with the OLS estimates (chart on the left) and with the price predictions of our best-performing ML method (chart on the right, boosted regression trees). On average, the price predictions from the ML approach are much closer to the actual prices than the OLS estimates. The difference in pricing accuracy is especially pronounced at the upper end of the price range: while the OLS estimates show large deviations from the actual prices, the ML-based price predictions are much closer.

In the final part of our paper, we conduct a bibliometric analysis and examine the publication success of articles published in major finance journals during the 2010–2021 period. Specifically, we address the following questions: (1) How important is ML as a novel methodology for research in finance? (2) What is the methodological purpose of ML (beyond prediction) in its applications for research in finance? (3) How do these findings differ across the various subfields in finance?

We find that although ML is a relatively new method in finance research, it has already found broad acceptance in the scientific community. The share of ML papers has grown in recent years and accounts for approximately 3%–4% of the publications in the top three finance journals (*The Journal of Finance, Journal of Financial Economics, The Review of Financial Studies*) in 2021. This share is similar for somewhat lower-ranked journals. Furthermore, our analysis reveals that the two main areas of finance—financial markets/asset pricing and banking/corporate finance—leverage the potential of ML in fundamentally different ways. While the literature in the field of financial markets/asset pricing tends to apply ML to

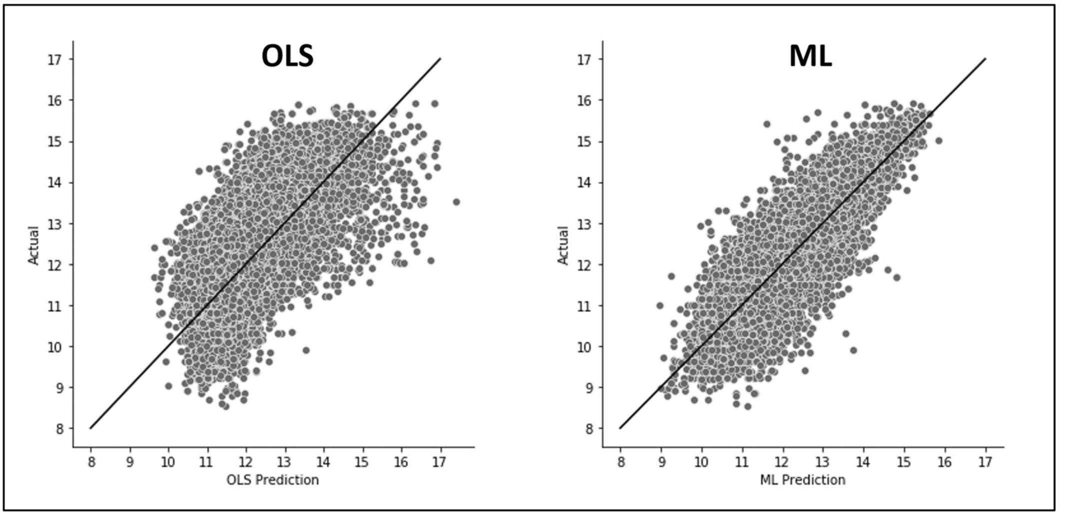

FIGURE 1 Comparison of the accuracy of hedonic pricing (OLS) and ML in predicting real estate asset prices. This figure depicts the accuracy of traditional hedonic pricing (OLS) and ML in predicting real estate asset prices in the German residential housing market. On average, the ML-based price estimates are much closer to the actual prices than the OLS estimates are. The benefit of ML is most pronounced at the upper end of the price range, where OLS performs especially poorly.

economic prediction problems, most publications in the fields of banking and corporate finance use ML to construct superior and novel measures. Interestingly, publications in the highest-ranked journals use ML disproportionally often to construct superior and novel measures. This effect is especially large within the fields of banking and corporate finance. Our results indicate a particularly large potential of applying ML to unconventional data to construct superior and novel measures for topics related to financial institutions and corporate finance.

Overall, our results suggest a promising future for ML applications in finance. The many benefits of ML over traditional econometric methods, the strong and consistent increase in the number of ML publications in the last few years, and the widespread usage of ML by studies published in the highest-ranked journals of the profession leave little reason to expect otherwise.[2]

Our paper is related to a growing literature focused on ML applications in finance. For instance, there is a small number of finance textbooks that either survey specific areas of finance in which ML techniques have recently emerged (e.g., Nagel, 2021, for asset pricing; De Prado, 2018, for asset management) or provide mathematical foundations for ML in quantitative finance (e.g., Dixon et al., 2020). The aim of these important contributions is to show how to carefully adapt ML techniques and how to deal with the specific characteristics of certain subfields in finance—with a particular focus on financial markets. Our perspective on ML is clearly different from the ones used in these important contributions as our interest lies in detecting promising ML applications beyond (prediction problems in) financial markets. We also add to a small number of survey papers that review the applications of ML in finance. These studies differ from ours in their use of classification techniques, scope, and focus. One group of surveys uses (mostly) automated techniques, such as textual analysis (Aziz et al., 2022) or citation-based approaches (Goodell et al., 2021), to classify ML applications across all finance subfields into *application areas* (such as risk forecasting or financial fraud). Another group of surveys adopts a more selective perspective and manually reviews either ML applications in certain subfields of finance, such as risk management (Aziz & Dowling, 2019), or applications of specific ML methods, such as deep learning (Ozbayoglu et al., 2020). Our study differs from these studies, which focus on *application areas* (i.e., *where* ML is applied), in that we classify the literature based on the *methodological purpose* of ML in finance (i.e., *how* ML is applied). This somewhat different angle—based on our novel taxonomy—allows us to uncover a frequently overlooked (but promising) group of ML applications in finance: While many of the existing surveys (tend to) focus on ML for prediction purposes, we show that two other types of ML applications are gaining importance: the construction of superior and novel measures and the extension of the existing econometric toolset for finance research. Furthermore, we also manually review all these ML papers instead of relying on automated techniques that might miss important context. Additionally, to the best of our knowledge, none of the existing reviews examines ML applications in finance with a bibliometric performance analysis based on the publication success of existing work by *research field* and *methodological purpose*.

The remainder of this paper is organised as follows. Section 2 gives a high-level introduction to ML together with an illustrative application of ML to a typical problem in finance. In Section 3, we present the three archetypes of ML applications and review the corresponding

[2]ML has received considerable attention not only from finance academia but also from practitioners. Table A1 in the appendix presents a selection of public announcements of large institutions (such as banks, insurance companies, and asset management firms) that make use of ML in their day-to-day business operations (e.g., HSBC and Deutsche Bank apply ML to predict and detect fraudulent transactions). These practice use cases mostly centre around prediction problems (the second archetype in our taxonomy).

literature. Section 4 outlines the most promising future directions for applying ML in finance. Section 5 concludes the paper.

2 | FUNDAMENTALS OF ML

In this section, we provide a primer of ML to lay the groundwork for subsequent chapters. Our focus is on the mechanics of the different types of ML, the problems for which ML has proven to be well suited for solving, and the methods with widespread use in the finance literature. We also emphasise the differences between ML and traditional econometric methods.

Most studies in empirical finance aim at analysing economic relationships between economic variables. A typical example is an analysis of how certain factors affect the capital structure or how regulatory changes affect the expectations of economic agents. Traditional econometric methods provide estimates $\hat{\beta}$ for the direction and strength of these factors.

ML, in contrast, serves different purposes. Instead of providing direct insights into the relationships between economic variables, ML tends to serve as a method for prediction or for data structure inference. Methods for prediction take the given observations to infer estimates for the dependent variable $\hat{y}$ of new observations based on their covariates X. For instance, the observed prices and property characteristics in the real estate market could be used to predict the prices of previously unobserved properties based on their characteristics. The first major type of ML, *supervised learning*, encompasses methods to make such predictions (see Section 2.1).

Methods for data structure inference derive structural information from given data X. A typical example is the identification of clusters in the data to learn how different observations relate to each other. The second major type of ML, *unsupervised learning*, comprises such methods to arrive at structural information from data (see Section 2.2).

Table 1 gives an overview of the differences between traditional econometrics and these two major types of ML, supervised and unsupervised learning. Most importantly, the three approaches serve different purposes. As explained above, traditional econometrics aims at extracting economic relationships (Samuelson & Nordhaus, 2009, p. 5) and thus solves so-called $\hat{\beta}$-problems (Mullainathan & Spiess, 2017). Supervised learning provides predictions; thus, it is mainly intended to solve so-called $\hat{y}$-problems (Mullainathan & Spiess, 2017). Unsupervised learning infers the data structure from given data without a special y-variable; thus, it solves X-problems.

The three approaches also differ with regard to their general methodology. Every approach makes use of data. In traditional econometrics, there is a dependent variable y and multiple independent variables X. In ML jargon, such data are called 'labelled data', as there is a *special label* y for each observation (which is the dependent variable y in regression jargon). The dominant method in traditional econometrics is linear regression, mainly due to its flexibility and interpretability. Linear regression with the OLS estimator provides an explanatory model in the form of a regression line and different metrics of statistical significance, such as *t*-values and *p*-values. Finally, these results can indicate causal relationships between economic variables.

Supervised learning also relies on labelled data. The special label y represents the target variable to be predicted based on the predictor variables X. Applying a supervised ML method on the given data yields a prediction model as well as estimates for its expected prediction performance. The prediction model can then be used to make out-of-sample predictions, that is, predictions of the value of the target variable of previously unobserved examples based on their characteristics.

TABLE 1 Differences between traditional econometrics and the two major types of ML: supervised and unsupervised learning

This table reports an overview of how traditional econometrics and the two major types of ML, supervised and unsupervised learning, differ with regard to the used data, method, results, usage, and purpose. Traditional econometrics enables explanations of economic phenomena, while supervised learning provides predictions and unsupervised learning infers data structure.

Approach	Data	Method	Results	Usage	Purpose
Traditional econometrics	Labelled data $(X_i, Y_i)_i$	Linear regression (OLS)	Explanatory model and statistical significance	(Causal) relationship	Explanation '$\hat{\beta}$'
Supervised learning	Labelled data $(X_i, Y_i)_i$	Supervised ML method	Prediction model and prediction performance	Out-of-sample predictions	Prediction '$\hat{y}$'
Unsupervised learning	Unlabelled data $(X_i)_i$	Unsupervised ML method	Data structure model and data structure characteristics	Structural information from data	Data structure inference 'X'

Unsupervised learning relies on unlabelled data, which is the defining distinction between unsupervised and supervised learning in the literature (Hastie et al., 2009, pp. 485–486). Unlabelled data means that there is no label y (i.e., no dependent variable y in regression jargon); all variables are considered 'equal'. Applying an unsupervised ML method to the given data yields a data structure model and data structure characteristics. Finally, both results can be used to infer structural information from the data.[3]

In the following sections, we describe the two major categories of ML—supervised and unsupervised learning—in more detail and give an overview (whose coverage is naturally selective) of the relevant methods for each category. Then, we provide an illustrative application of ML to a typical problem from the field of household finance: the prediction of real estate prices. Finally, we discuss limitations, caveats, and drawbacks of ML.

2.1 | Supervised learning

Supervised learning aims at making out-of-sample predictions with high prediction performance. To accurately assess the expected prediction performance on previously unseen observations, the given data are divided into *training data* and *test data*. Then, a supervised ML method is applied to the training data to build a prediction model. Finally, applying the prediction model to the test data yields an estimate of the expected out-of-sample prediction performance.

To build a prediction model, various supervised ML methods of differing complexity have been developed. In general, more complex methods tend to enable higher prediction performance but reduce interpretability. Figure 2 gives an overview of common methods of supervised ML arranged by typical prediction performance and interpretability.

The simplest method is linear regression with the *OLS* estimator. *OLS* provides excellent interpretability. However, its out-of-sample prediction performance has turned out to be generally weak. One way to improve the prediction performance of the linear *OLS* model would be to add nonlinear transformations and interactions of the original predictor variables to the model specification. In many cases, however, it is ex ante unclear which nonlinearities and interactions are actually relevant. Including all possible combinations is generally difficult since it results in an exorbitant number of variables that can quickly exceed the number of observations. In many cases, the sheer size of the resulting data sets would also lead to computational problems.

Since *OLS* (under certain conditions) is the best linear *unbiased* estimator (BLUE), one way that has been proposed to improve the prediction performance is to allow for bias. In contrast to explanation problems, prediction problems aim to achieve maximal prediction performance; thus, they do not require unbiasedness of variable coefficients. Regularised linear methods offer a way to systematically introduce bias to improve *OLS* prediction performance (Hastie et al., 2009, pp. 61–79). More specifically, regularisation means that such methods shrink the coefficients of the predictor variables to increase prediction performance.[4] The most common method for regularised linear regression is the *least absolute shrinkage and selection operator*

[3]While supervised and unsupervised learning are arguably the most important categories of ML, there also exist other categories of ML that are less common but relevant for specific applications: reinforcement learning for sequential decision problems (Sutton & Barto, 2018), semisupervised learning for problems with mostly unlabelled training data (Zhu, 2005), and active learning for problems with costly training data (Settles, 2009).

[4]The introduction of bias can increase prediction performance because of the bias-variance tradeoff. See, for instance, Hastie et al. (2009, pp. 37–38, 219–228) for technical details.

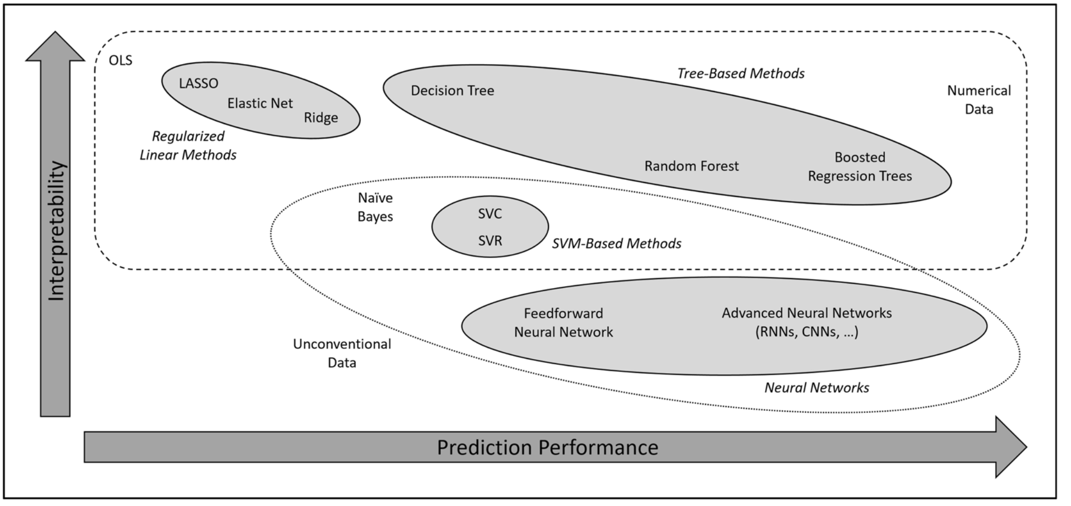

FIGURE 2 Overview of common methods in supervised ML arranged by typical prediction performance and interpretability. This figure gives an overview of the most common methods in supervised ML. The methods differ by complexity: more complex methods typically achieve higher prediction performance but are less interpretable. For numerical data, less complex methods tend to work well, while unconventional data (such as text, images, or videos) often require more complex methods.

(LASSO). *LASSO* works similarly to *OLS* but introduces bias by adding a penalty term in its optimisation function to penalise large variable coefficients with little informational content. The specific functional form of the penalty term drives irrelevant coefficients to zero. Hence, *LASSO* is often used for variable selection in addition to pure prediction and also provides relatively good interpretability.

In addition to *LASSO*, there are other regularised linear methods that differ with regard to the functional form of the penalty term. *Ridge regression* uses a penalty term that does not drive coefficients to exactly zero and is therefore less interpretable. However, *ridge regression* often provides superior prediction performance compared to *LASSO*. *Elastic net regression* combines the two methods (Zou & Hastie, 2005). Its penalty term is a linear combination of the penalty terms of *LASSO* and *ridge regression* to incorporate their respective strengths.

In contrast to the linear methods just discussed, more complex ML methods automatically consider relevant nonlinearities and interaction effects. For numerical data, tree-based ML methods are widespread (Hastie et al., 2009, pp. 305–334). The simplest tree-based method is the *decision tree*, which also acts as the building block of all other tree-based methods. Panel A in Figure 3 depicts a simplified *decision tree* trained for house price prediction. It consists of nodes at which the tree splits depending on the value of a certain predictor variable. *Decision trees* typically contain multiple layers of nodes, so they implicitly consider interactions between multiple variables. When the tree reaches a leaf node, that is, a node after which there is no further split, the tree returns a prediction value. Given that the relevant predictor variables and thresholds are directly observable in the splits, *decision trees* are characterised by relatively high interpretability.[5]

[5]For more details on decision trees, see, for example, Loh (2011).

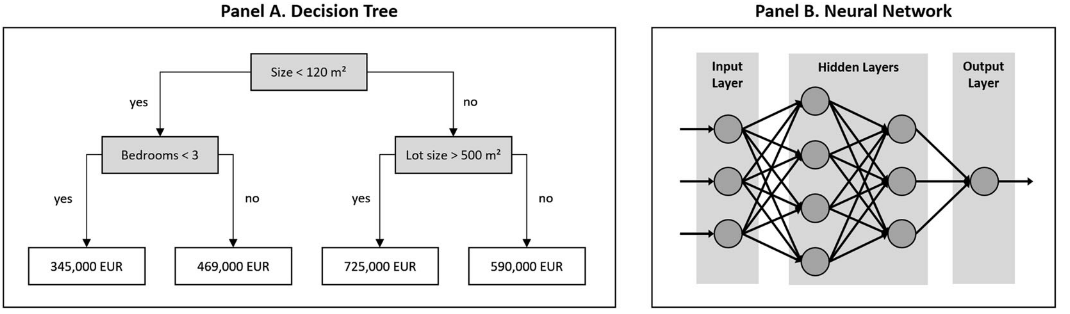

FIGURE 3 Illustrations of a decision tree and a neural network. This figure depicts a *decision tree* (Panel A) and a *neural network* (Panel B). The *decision tree* was trained for house price prediction. It reaches its prediction decision by evaluating the value of certain predictor variables at each split. *Neural networks* consist of multiple layers of neurons through which the given data are processed. The shown *neural network* uses a simple feed-forward architecture, which means that data only flow from left to right.

Random forests combine multiple *decision trees* (Breiman, 2001). More specifically, the *random forest* method repeatedly draws bootstrap samples from the given data and builds a separate *decision tree* from each sample. The prediction of a *random forest* is then the average prediction value of the different trees. *Random forests* typically achieve much higher prediction performance than single *decision trees* but are inherently less interpretable.

Boosted regression trees extend the concept of *random forests* to further improve their prediction performance (Hastie et al., 2009, pp. 353–358). Instead of combining many independent *decision trees*, the *boosted regression tree* method builds the trees iteratively and considers which observations the previous trees could not predict well. *Boosted regression trees* typically not only outperform *random forests* but are often among the winning algorithms in data science competitions, which highlights their state-of-the-art prediction performance level.

While *tree-based ML methods* and, in particular, *boosted regression trees* achieve state-of-the-art prediction performance with numerical data, *neural networks* often excel with unconventional data such as text, images, or videos. Panel B in Figure 3 depicts a small *neural network*. A *neural network* consists of two components: neurons (arranged in so-called layers) and links between neurons (Hastie et al., 2009, pp. 389–415). The links describe the flow of data between the neurons. First, a *neural network's* input layer receives the predictor variables, for instance, pixel-level image data. Then, the hidden layers iteratively process the data and deliver them to the output layer, which returns the final prediction value. In its most basic version, a neuron first calculates a weighted sum of the data that arrive from the neurons of the previous layer (the weights are determined endogenously during the training process). Then, it applies a nonlinear function (e.g., a logistic function) to this weighted sum. Finally, the neuron sends the result of this calculation to all neurons of the next layer to which it is connected. The number of layers, the number of neurons in each layer, the links between neurons, and the functional forms of the nonlinear functions are (exogenously) specified by the designer of the neural network and depend on the given problem. *Neural networks* used in real applications can be very large with many hidden layers and thousands of neurons and links. Furthermore, they do not have to be fully connected, so not every neuron of a layer necessarily needs to forward its output to every neuron of the next layer. Various architectures have been proposed to build *neural networks*. One of the simplest architectures is the feed-forward network: neurons come in their most basic variant, and no backlinks exist so that data simply flow from left to right.[6] Due to their high complexity, *neural networks* are inherently difficult to interpret. In general, very little information can be inferred from the hidden layers, which represent the learned knowledge of a *neural network*. Improving the interpretability of *neural networks* is subject to ongoing research in computer science.

In addition to the methods just discussed, there are older ML methods that (compared to newer methods) typically achieve worse prediction performance and/or provide lower interpretability, such as the *naïve Bayes* method (Rish, 2001), which uses Bayes' theorem to classify observations into categories, or *support vector machine (SVM)* methods (Hastie et al., 2009, pp. 417–455). We refer the interested reader to the mentioned literature for more details on these methods.

2.2 | Unsupervised learning

The purpose of unsupervised learning is data structure inference. Since the data structure subsumes many different types of information, we divide the methods of unsupervised learning into different subcategories. The two most common subcategories in unsupervised learning are *clustering* and *dimensionality reduction*.

In *clustering*, observations are grouped in a way that results in high within-group similarity and low cross-group similarity. Various kinds of clustering methods have been proposed. First, *centroid-based methods* form clusters by arranging the observations around multiple central points (so-called centroids). After the initial positioning of the centroids, iterative updates of their position yields increasingly suitable clusters. A common example of a very early but still heavily used centroid-based method is K-means (MacQueen, 1967). Second, *density-based methods* build clusters depending on the differing density in the space of observations. In other words, they group observations with many similar observations nearby into clusters. An example of a *density-based clustering method* is DBSCAN from Ester et al. (1996), which is also one of the most widely applied clustering methods. Third, *distribution-based methods* assign observations to clusters based on whether they likely belong to the same statistical distribution. Hence, these methods require knowledge of the distribution of the underlying data process in advance. For normally distributed data, Gaussian mixture models are widespread (Rasmussen, 1999). Finally, *hierarchical methods* construct clusters that consider the hierarchical relationship in the data. They start with initial clusters, where each cluster consists of a single observation. Then, they iteratively combine smaller clusters into larger clusters to build a hierarchy. A common method for hierarchical clustering is BIRCH (Zhang et al., 1996).

Dimensionality reduction aims at increasing the information density of the given data by decreasing their dimensionality while retaining most of the inherent information. There are various methods for dimensionality reduction, of which we cover only the two most common ones. First, methods based on *principal component analysis* (*PCA*) derive linear combinations of

[6]Advanced *neural networks* employ more complex neurons and architectures. *Recurrent neural networks (RNNs)* are designed for sequential data such as text (Medsker & Jain, 2001). The special architecture of *RNNs* allows hidden-layer neurons to accumulate information over multiple related observations (for instance, words in a sentence). There are different possibilities for designing this information storage mechanism. Widespread design examples are gated recurrent units (GRU) and long short-term memory (LSTM). *Convolutional neural networks (CNNs)* are another type of advanced neural networks whose general architecture fits well with visual data such as images and videos (Albawi et al., 2017). Simply put, their hidden layers represent trainable filters that iteratively detect increasingly complex structures. The architecture of *CNNs* is typically highly customised toward a specific application. Adequately designed *CNNs* show outstanding performance for tasks such as face detection or general image recognition.

the original variables ('principal components') that cover as much of the data's variance as possible. While the basic variant of *PCA* is inherently linear, nonlinear generalisations also exist. For more details on the different *PCA*-based methods, see, for instance, Hastie et al. (2009, pp. 534–552). Second, *methods based on neural networks* reduce dimensionality with special architectures. A widely used method is the autoencoder neural network (Goodfellow et al., 2016, pp. 499–523). An autoencoder consists of an encoder network that creates a condensed representation of the input data and a subsequent decoder network that reconstructs the original data from the condensed representation. A special bottleneck layer connects the encoder and decoder networks to train them on given data. If the autoencoder is able to reconstruct the original data well, then the condensed data representation in the bottleneck layer has successfully retained most of the information in the data while reducing its dimensionality.

In addition to *clustering* and *dimensionality reduction*, further subcategories of unsupervised learning exist but are (to date) used somewhat less often for applications in finance. *Association rule mining* tries to identify relations between variables (Agrawal et al., 1993). For instance, it can learn from customer purchase data which products are often bought together. *Outlier detection* tries to find observations that substantially differ from the remaining data. While many traditional methods for outlier detection exist, ML-based methods often provide superior performance, especially in high-dimensional settings (Domingues et al., 2018). Methods in *synthetic data generation* try to generate new data that satisfy certain requirements. Generative adversarial networks, for instance, use neural networks to create new, synthetic data that closely mimic the given training data (Goodfellow et al., 2020). Their neural network architecture makes them especially useful for unconventional data, for example, to create artificial images that are similar to existing images.

2.3 | Application: Real estate price prediction

To illustrate the differences between ML methods and more traditional approaches, we now apply ML to the problem of real estate price prediction. The prediction of real estate prices is a particularly good example to illustrate the benefits of ML to solve problems in finance for three reasons. First, real estate is one of the most important asset classes in the economy. In the United States, the total value of real estate assets is comparable to the size of the equities and fixed income markets combined. For most households, real estate is the greatest source of wealth. The Global Financial Crisis in 2007/2008 exemplified how spillover effects from the real estate sector can destabilise economies around the world. Consequently, the reduction of prediction errors in the area of real estate pricing is of particular economic importance. Second, real estate assets show a high level of heterogeneity (each property is unique), which makes real estate pricing challenging. Third, the high number of property characteristics variables as well as potentially relevant nonlinearities and interaction effects makes real estate pricing an inherently high-dimensional problem, where ML provides unique benefits over traditional methods. The traditional approach to derive price estimates for individual properties is hedonic pricing. Hedonic pricing first regresses the property characteristics on the observed property prices with OLS to obtain a linear pricing model. Then, this model can produce price estimates for new, previously unobserved properties. It is also possible to interpret the regression coefficients as the characteristics' shadow prices. However, hedonic pricing relies on an inherently linear model and therefore does not directly consider nonlinearities and interaction

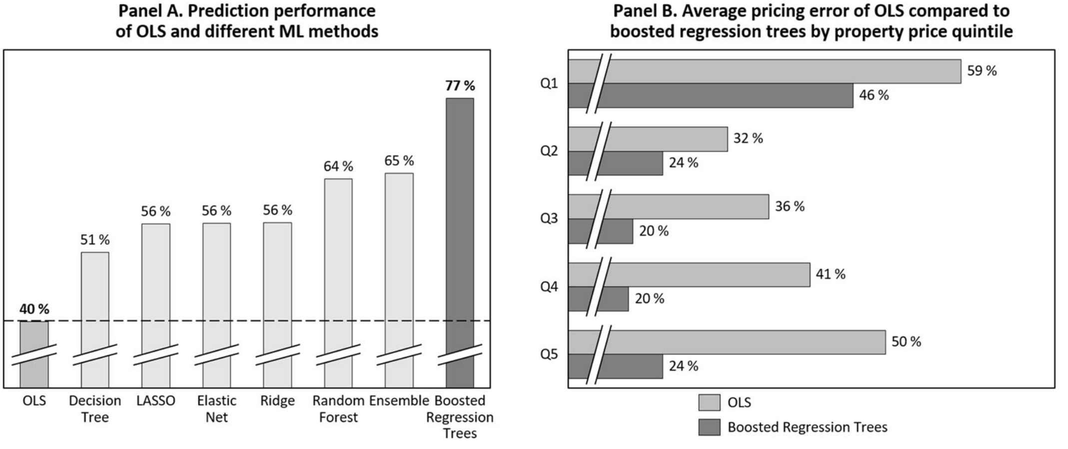

FIGURE 4 Prediction performance and *average* pricing errors of hedonic pricing (OLS) and ML methods. Panel A depicts the prediction performance (R^2) of traditional hedonic pricing (OLS) *compared to* different ML methods. While most ML methods outperform OLS, the boosted regression trees method performs best by far and almost doubles the OLS performance. Panel B shows the average pricing error (measured by mean absolute error [MAE]) for the best-performing ML method, boosted regression trees, and for the OLS baseline in the five price quintiles. In all quintiles, the boosted regression trees method significantly outperforms OLS. The reduction in pricing error from ML is most pronounced in the highest price quintile, where OLS performs relatively poorly.

effects. For instance, we can assume relevant interactions between lot size and location: an additional m^2 in lot size for a property in a city centre is likely worth more than in a suburb. While we could manually add such specific effects to the linear model, there may exist a plethora of unknown nonlinear and interaction effects. By ignoring these effects, the linear model of hedonic pricing potentially leaves important information contained in the data unexploited. ML methods, in contrast, automatically consider nonlinearities and interactions. Therefore, supervised ML can potentially generate price predictions that exhibit lower pricing error than the linear model from hedonic pricing. In the following, we study whether and how ML provides superior price estimates for individual real estate assets.

We exploit a comprehensive collection of more than four million residential real estate listings in Germany between January 2000 and September 2020 from the five major real estate online platforms and major newspapers.[7] The data set contains offer prices and all relevant individual property characteristics (floor area, number of rooms, construction year, location, lot size, etc.). We use these data to train different ML models for the prediction of individual property prices and compare these models with the linear OLS model from hedonic pricing. Panel A in Figure 4 shows the key result of our analysis.[8] ML methods strongly improve the accuracy of price predictions over the OLS baseline. Our best-performing ML method, boosted regression trees, dramatically increases out-of-sample R^2 to 77%, compared to 40% for OLS; thus, it almost doubles the amount of explained price variation. On average, the predictions from boosted regression trees deviate from the actual prices by approximately 27%, compared to 44% for OLS. In monetary terms, the superior prediction performance of boosted regression

[7]According to the data provider, the data set covers more than 95% of the public listings during the given period.
[8]See the online appendix for more details on the sample and our methodology.

trees corresponds to an average pricing error of approximately 94,000 EUR, compared to 176,000 EUR for OLS. Since the mean property price in our sample is 393,000 EUR, the improvements in pricing accuracy from ML are not only statistically significant but also economically large.

While the improvements in pricing accuracy induced by ML are already impressive on average, their benefits become even more pronounced at the upper end of the price range. Panel B in Figure 4 depicts the prediction performance of the best-performing ML method, boosted regression trees, compared to that of OLS in the five property price quintiles. The boosted regression trees method outperforms OLS in all quintiles. While OLS performs worst at the extremes of the price range, ML is especially useful in reducing the pricing error for the most expensive properties. In the highest price quintile, the boosted regression trees method lowers the average pricing error to 24%, compared to 50% for OLS. In monetary units, the superior prediction performance of boosted regression trees relative to that of OLS corresponds to a reduction in the average pricing error by more than 240,000 EUR in the highest price quintile. Given that the average property price in the top quintile is approximately 884,000 EUR, the improvements in pricing power from ML are dramatic. Our results indicate that nonlinearities and interaction effects are relevant in real estate pricing and especially important for the most expensive properties.

Our results demonstrate the benefits of using ML to reduce the prediction error in economic prediction problems. ML can yield a statistically and economically significant reduction in prediction error compared to traditional linear regression with OLS in addressing the problem of real estate price prediction. The already large benefits of ML on average further increase for assets at specific price ranges. Hence, ML methods not only improve prediction accuracy in general but also especially for observations where traditional approaches struggle.[9]

2.4 | Limitations, caveats, and drawbacks of ML

While the results from our illustrative application of ML to real estate asset pricing show the benefits of ML over traditional methods for problems with high-dimensional data, there also exist limitations, caveats, and drawbacks of using ML. In the following, we discuss three important aspects in detail.

First, ML methods tend to exhibit *low interpretability*. While ML models can produce predictions with low prediction error, it is often not directly observable how the algorithm has generated its results. Hence, ML is generally not suited for problems that require a deep understanding of the economic determinants of the prediction target. Nevertheless, the quickly advancing field of interpretable ML tries to offer solutions to the model interpretability problem with several kinds of approaches (see, for instance, Burkart & Huber, 2021, for an overview of the available methods).

[9]Our real estate asset pricing example is primarily meant to illustrate the advantages of ML over traditional methods for a problem with high-dimensional data. Nevertheless, it represents (to the best of our knowledge) the first application of ML to real estate pricing for an entire major economy, spanning a comprehensive data set of all real estate listings—both, online and offline—for a sample period of more than 20 years. Our data set contains more than four million observations, which far exceeds the scale of prior work. Most existing studies in the real estate asset pricing literature apply ML to predict individual house prices in narrow regions within different countries, such as the United States (Mullainathan & Spiess, 2017; Park & Bae, 2015; Pérez-Rave et al., 2019), France (Tchuente & Nyawa, 2022), Spain (Rico-Juan & Taltavull de La Paz, 2021), the Netherlands (Guliker et al., 2022), Turkey (Erkek et al., 2020), Hong-Kong (Ho et al., 2021), and Colombia (Pérez-Rave et al., 2019). In addition to predicting individual real estate prices, a small group of studies uses ML to predict the general price level in the real estate market (Milunovich, 2020; Yu et al., 2021).

Second, ML generally requires *large data sets*. Data sets can be large in two dimensions: the number of relevant variables and the number of observations. ML offers benefits over traditional methods for prediction tasks if the number of relevant variables is large relative to the number of observations. At the same time, ML usually provides good prediction performance only if there is a high number of observations on which an ML model can be trained. Unfortunately, large-scale data are not always available for many research questions in finance. In some cases, using ML models that have already been pretrained with large amounts of comparable data can solve this problem. Such pretrained models exist for many common ML tasks, such as textual analysis or face recognition, so researchers can directly apply them to the problem at hand independent of the amount of available data. In addition, the general trend toward increasing data collection in all aspects of life should more and more alleviate the data problem.

Finally, using ML often has *high computational costs*. Compared to traditional methods such as linear regression, training ML models requires significantly more time and computing power. The problem typically becomes worse with more sophisticated ML methods. In particular, neural networks with complex architectures typically have the highest computational costs. As a result, using cloud computing services often becomes necessary to deal with this problem.

3 | TAXONOMY OF ML APPLICATIONS IN FINANCE

An increasing number of finance papers that use ML in at least some part of their study go on to be published. However, many researchers are still unaware of how and where to apply ML in the field of finance. In this section, we present a taxonomy of existing ML applications, which serves multiple purposes. First, it outlines where ML can add value in finance research. Second, it provides a systematic overview of existing ML applications in the field of finance. Third, it enables a better understanding of new contributions and how they relate to the existing literature. Finally, it may guide researchers in discovering possible applications and thus may facilitate new ML studies in finance.

As explained above, ML solves different problems compared to traditional econometric methods. The workhorse model of finance research, linear regression with OLS, has one major objective: identification of causal relationships between economic variables to explain economic phenomena. In contrast, ML provides predictions that minimise prediction error or infers structural information from given data.

To survey the ML literature in finance, we first identify ML-related papers in major journals in finance, the NBER working paper series, and the Financial Economics Network of the SSRN preprint repository; then, we search for ML method names and their variations (e.g., LASSO, random forest, etc., see Section 2). We study these papers and categorise the ML research strategies in these papers into the following three distinct archetypes:

(1) Construction of superior and novel measures: $y = \beta \boldsymbol{X} + \varepsilon$.
(2) Reduction of prediction error in economic prediction problems: $\hat{\boldsymbol{y}} = f(X)$.
(3) Extension of the existing econometric toolset: $y = \beta X + \varepsilon$ *and* $\boldsymbol{ML}$.

Studies of the first archetype use ML to construct a superior or novel measure for one of the independent variables X. The main analyses of these papers still largely rely on a traditional (linear) model, which is estimated, for example, with OLS. Studies of the second archetype use ML to reduce the prediction error of predictions $\hat{y}$ in economic prediction problems. Supervised ML methods achieve superior prediction performance by using flexible functional forms $f(*)$ in the prediction model. Studies of the third archetype use ML to extend the existing econometric toolset. ML methods either serve as new econometric methods themselves or optimise some part of a traditional econometric method. In the following subsections, we review the literature related to each of the three archetypes of ML applications in finance in detail.[10]

3.1 | Construction of superior and novel measures

The first archetype of ML applications in finance is the construction of superior and novel measures. Studies of this archetype use ML to extract information from high-dimensional, unconventional data such as text, images, or videos and construct a numerical measure of an economic variable. For textual data, traditional approaches use word counts based on domain-specific dictionaries.[11] For image and video data, only human assessments have been available for a long time. ML-based approaches provide easier and, at the same time, more powerful access to the information contained in unconventional data. All types of ML methods are applicable: predictions from supervised learning, data structure information from unsupervised learning, and results from other types of ML can be used to construct measures of economic variables.

The superior or novel measure finally serves as an independent variable in the main analysis of an economic relation. Using superior measures (i.e., with lower measurement error than existing measures) reduces attenuation bias, which leads to more precise estimates of the parameters describing an economic relationship. Novel measures enable new analyses with previously unmeasurable economic aspects. In the main analysis, most studies that construct ML-based measures apply traditional econometric methods such as linear regression with OLS.

Table 2 presents a selection of studies that use ML to construct superior or novel measures. In the following, we present them in three categories: (1) measures of sentiment, (2) measures of corporate executives' characteristics, and (3) measures of firm characteristics.

3.1.1 | Measures of sentiment

Measures of sentiment describe beliefs of people, usually on a positive–negative scale. Most studies in this subcategory construct measures of sentiment from textual data. There are multiple approaches to construct a one-dimensional (positive vs. negative) measure of sentiment from textual data. Loughran and McDonald (2011) present a dictionary approach to derive sentiment from financial texts. More specifically, they count negative words based on a finance-specific word list. Dictionary approaches, however, miss the context of words within a sentence (Loughran & McDonald, 2016). In contrast, flexible ML-based approaches can

[10]Given the quickly evolving nature of the field, our review is necessarily selective regarding some ML applications. For instance, we may not consider important papers outside of the 'standard' finance domain, such as genuine computer science papers that apply ML to specific finance problems. Finally, our manual review is to a certain degree subjective, especially compared to automated review techniques (such as textual analysis [Aziz et al., 2022] or citation-based approaches [Goodell et al., 2021]).

[11]See Loughran and McDonald (2016) for an overview of mostly traditional text analytics methods in accounting and finance.

TABLE 2 Overview of studies that use ML to construct superior and novel measures

This table reports an overview of the relevant studies in finance that apply ML to construct superior and novel measures. There are three main categories: measures of sentiment, measures of corporate executives' characteristics, and measures of firm characteristics.

Category	Subcategory	Measures
Measures of Sentiment	Stocks	– Investor sentiment in social media – Sentiment in news – Sentiment in analyst reports – Sentiment in annual reports
	Sovereign Debt	– Sentiment in news
	Products	– Consumer sentiment in social media – Expert sentiment in product-technology articles
Measures of Corporate Executives' Characteristics	Personality Traits	– Big Five scores – Risk tolerance
	Beliefs	– Confidence in expressing opinions
	Emotions	– Facial emotions (e.g., happiness, sadness, anger, fear, disgust) – Verbal emotions (e.g., positive, negative, warmth, ability) – Vocal emotions (e.g., valence, arousal, happiness, sadness)
	Actions and Working Patterns	– Answer avoidance in conference calls – Working style (high- vs. low-level activities) – Communication style
	Quality	– Expected shareholder support
	Looks	– (Facial) Attractiveness – (Facial) Trustworthiness – (Facial) Dominance – (Facial) Masculinity
Measures of Firm Characteristics	Financial Characteristics and Risk Exposures	– Financial constraints – Risk exposures (e.g., COVID-19, cybersecurity)
	Corporate Culture	– Cultural values (e.g., innovation, integrity, teamwork) – Gender culture – Board responsibilities
	Connectedness	– Political connectedness – Venture capital communities – Mutual fund voting behaviour

consider not only the context of words within a sentence but also how different sentences interrelate with each other. For an extensive review of sentiment with traditional econometric and ML-based approaches, see Algaba et al. (2020).

Sentiment exists for many topics and is derived from many sources. In finance, our interest mainly lies in the aggregate sentiment of markets such as the *stock market*, which is the most common target of ML-based measures of sentiment. The majority of the relevant studies use measures of sentiment for stocks to study their effect on future stock returns and various financial reporting numbers.

There are multiple studies that construct a measure of investor sentiment from social media. Antweiler and Frank (2004) use the ML methods naïve Bayes and SVM to classify user posts on the Yahoo Finance message board as positive or negative. Then, they aggregate their classifications to construct a measure of stock market sentiment. Renault (2017) similarly classifies user posts on the finance-focused social network StockTwits to construct a measure of investor sentiment. Vamossy (2021) also relies on StockTwits but measures investor emotions by extracting different emotional states from user posts with textual analysis based on deep learning. The studies by Sprenger et al. (2014), Bartov et al. (2018), Giannini et al. (2018), and Gu and Kurov (2020) derive investor sentiment from user posts on Twitter. Liew and Wang (2016) also apply ML to extract sentiment information from Twitter but for pre-IPO sentiment.

In addition to social media, news articles are another source of sentiment for stocks. Barbon et al. (2019) enhance the naïve Bayes method to build a sentiment variable based on firm-specific news. Ke et al. (2019) implement a customised ML-based approach that specialises in extracting information relevant for stock returns. Their method then allows them to extract a measure of sentiment for stocks from Dow Jones Newswire articles. Similarly, Boudoukh et al. (2019) also analyse Dow Jones Newswire articles but focus on the saliency of firm-specific news. Manela and Moreira (2017) deviate from the traditional measures of sentiment that use a positive–negative scale. Instead, they construct a measure of stock market uncertainty from *Wall Street Journal* front-page articles. von Beschwitz et al. (2020) study how ML-based news analytics (i.e., computer algorithms that investors use to interpret financial news) affect stock prices, trading volumes, and liquidity. Calomiris and Mamaysky (2019) use ML to measure sentiment from country-level news articles and study how it affects returns and volatilities. In addition to the analysis of text, Obaid and Pukthuanthong (2022) apply ML to news photos to derive a measure of sentiment for stocks and find that it can act as a substitute of text-based measures.

Other studies use analyst reports or annual reports for measures of sentiment. Huang et al. (2014) apply the naïve Bayes method to analyst reports to construct a measure of stock sentiment. Azimi and Agrawal (2021) apply deep learning methods to 10-Ks to measure sentiment and study its effect on abnormal returns and trading volumes.

While most studies that construct ML-based measures of sentiment consider sentiment for stocks, Cathcart et al. (2020) study sentiment for *sovereign debt markets*. More specifically, they leverage news sentiment information from Thomas Reuters News Analytics to investigate the impact of media content on sovereign credit risk.

Beyond sentiment for financial markets, two studies examine sentiment for *products*. Tang (2018) uses a commercial service to create a measure of consumer sentiment based on Twitter posts. The subsequent main analysis studies the effect of consumer sentiment on firm sales. Nauhaus et al. (2021) construct a measure of expert sentiment from articles concerning specific technology domains and then study how it affects firms' capital allocation among the business units engaged in these domains.

3.1.2 | Measures of corporate executives' characteristics

The prominent role of a firm's leadership and its large implications has led to a vast amount of finance literature that studies various aspects of corporate executives. Related to this stream of the literature, ML enables the construction of superior and novel measures of executives' characteristics. While most measures in this category rely on textual data, there are also some studies that construct measures from analysing images and videos.

Multiple studies construct ML-based measures of *executives' personality traits*. Gow et al. (2016) use ML to extract CEOs' Big Five personality scores (agreeableness, conscientiousness, extraversion, neuroticism, and openness to experience) from the Q&A part of conference call transcripts. Then, the authors use the extracted scores to analyse the effect of personality on financing choices, investment choices, and operating performance. Similarly, Hrazdil et al. (2020) determine the Big Five personality scores of CEOs and CFOs by using the commercial service IBM Watson Personality Insights. From these scores, they construct a novel measure of executives' risk tolerance to analyse its effect on audit fees.

Other studies construct measures of *executives' own beliefs*. For instance, Du et al. (2019) apply ML to mutual fund managers' letters to shareholders to construct a measure of managers' level of confidence in expressing opinions. Their main analysis then studies the effect of confidence on future performance.

Recent advances in ML also enable studies that construct measures of *executives' emotions*. Akansu et al. (2017) apply ML-based face-reading software to videos of CEOs during press interviews to extract facial emotions and quantify CEO mood. They measure emotions such as anger, disgust, fear, happiness, sadness, or surprise and study their effect on firm performance. Hu and Ma (2021) use ML to construct measures of startup founders' emotions during investor pitch videos. More specifically, they measure three dimensions of emotions: facial emotions, verbal emotions, and vocal emotions. Finally, they analyse the effect of the three dimensions on the probability of obtaining a venture capital investment. Breaban and Noussair (2018) use ML-based face-reading software to extract the emotional state of traders in an experimental setting.

Another stream of the literature addresses *executives' actions and working patterns*. Barth et al. (2020) propose an ML-based measure of the degree to which executives obstruct the flow of information during earnings conference calls by giving so-called nonanswers to investors' and analysts' questions. Bandiera et al. (2020) apply ML to CEO survey data to construct a measure of CEO working style. More specifically, their measure captures whether a given CEO performs more low-level or more high-level activities. Then, this novel measure enables the authors to study firm-CEO assignment frictions. Choudhury et al. (2019) construct a measure of executives' communication style by applying ML to transcripts and videos from interviews of emerging market CEOs. Dávila and Guasch (2022) construct a measure of entrepreneurs' nonverbal communication style during pitch presentations with ML-based computer vision software and analyse its relation to firm valuations and funding success rates.

The study by Erel et al. (2021) uses ML to measure director quality. They predict the (excess) level of directors' shareholder support over the first 3 years of tenure using various ML methods. By interpreting these predictions as a measure of director quality, the authors study firms' decision-making process in the selection of corporate directors.

Finally, the large amount of image data freely available on the internet allows many studies to systematically exploit the information that the *looks of corporate executives*—in particular, their facial traits—may contain. Hsieh et al. (2020) extract a measure of trustworthiness from executives' business headshot images. More specifically, they detect and use certain facial features (such as eyebrow angle or face roundness) to predict perceived trustworthiness. Their main analysis studies the effect of executives' trustworthiness on audit fees. Peng et al. (2022) leverage the social network LinkedIn and apply ML to profile photos of sell-side analysts to construct measures of trustworthiness, dominance, attractiveness, etc. Kamiya et al. (2019) use ML to first measure the width-to-height ratio of CEOs' faces from portrait photos and then infer a measure of facial masculinity to study its effect on firms' riskiness.

3.1.3 | Measures of firm characteristics

Studies in the third category construct measures of firm characteristics with ML methods. The first subcategory consists of measures of *firms' financial characteristics and risk exposures*. Buehlmaier and Whited (2018) apply ML to annual reports to construct a measure of financial constraints. Their ML-based measure achieves superior performance compared to the existing measures. Hanley and Hoberg (2019) construct a measure of aggregate risk exposure in the financial sector from individual banks' annual reports by using a commercial ML-based service. They use their measure to study the effect of financial sector risk on banks' stock returns and volatility as well as bank failure. Li et al. (2021a) apply ML-based textual analysis methods to construct measures of firms' exposure and response to COVID-19 based on the information from earnings calls. Alan et al. (2021) measure firm-level cybersecurity risk with ML-based methods from computational linguistics. More generally, Lima and Keegan (2020) provide an overview on how ML-based textual analysis can be applied to social media to assess cybersecurity risk.

ML can also help to study *corporate culture*. Li et al. (2021b) extract aspects of corporate culture from conference call transcripts with ML and build measures of five different corporate culture values. Using these measures allows them to analyse the effect of corporate culture on firm policies such as executive compensation and risk-taking. Furthermore, they study the effect on firm performance metrics such as operational efficiency and firm value. Adams, Akyol, et al. (2021) apply ML to firms' reports to a gender-equality agency to construct multiple measures of corporate gender culture. Their novel measures allow them to systematically study how firms treat female employees. Adams, Ragunathan, et al. (2021) apply ML-based textual analysis to extract boards' and board committees' responsibilities and meeting frequencies.

Finally, the capabilities of ML enable the construction of novel measures of firms' *connectedness*. Mazrekaj et al. (2021) apply ML to construct a measure of firms' political connections, which helps identify potential conflicts of interest. Bubna et al. (2020) study venture capital syndications and create a measure of venture capital relatedness. More specifically, they cluster venture capital firms using ML to identify syndication groups and study their effect on startup maturation and innovation. Bubb and Catan (2021) apply clustering methods from unsupervised learning to mutual funds' proxy votes to determine to which voting parties they belong.

3.2 | Reduction of prediction error in economic prediction problems

Studies of the second archetype of ML applications in finance apply ML to reduce prediction error in economic prediction problems. While many problems in economics require the identification of causal relationships between economic variables, some problems directly require prediction. ML can reduce the prediction error in such problems, that is, generate more accurate predictions than simpler approaches such as fitted values from linear regression with OLS.

Predictions can be generated from numerical data as well as unconventional data such as text, images, or videos. Since the purpose of ML in this category is to minimise prediction error in economic prediction problems, by definition, only supervised ML is directly applicable here. Given the large number of available ML methods, most studies use a multitude of different methods to assess which method works best on the given data. Applying supervised ML methods finally results in predictions of an economic variable, which directly helps in solving an economic prediction problem.[12]

Table 3 gives an overview of the relevant studies that use ML in economic prediction problems to reduce prediction error. In the following, we present these studies in the three categories of (1) prediction of asset prices and trading mechanisms, (2) prediction of credit risk, and (3) prediction of firm outcomes and financial policy.

3.2.1 | Prediction of asset prices and trading mechanisms

The prediction of asset prices and trading mechanisms is of central importance in studying capital markets. ML can reduce the prediction error in various types of prediction problems. We distinguish among predictions in the following seven different subcategories: equities, bonds, foreign exchange, derivatives, general market prices, investors, and market microstructure.

The most common ML-based prediction in the subcategory of *equities* is the prediction of future stock returns, which is closely related to the field of cross-sectional asset pricing. Rasekhschaffe and Jones (2019) provide an overview of the use of ML for predicting the cross-section of stock returns and the selection of individual stocks. Martin and Nagel (2022) emphasise the challenges of cross-sectional asset pricing with high-dimensional data. Gu et al. (2020) directly predict future stock returns based on firm characteristics, historical returns, and macroeconomic indicators. They use ML methods with varying complexity ranging from regularised linear models to neural networks. Furthermore, they analyse which predictor variables are the most informative in predicting the cross-section of stock returns. Rossi (2018) predicts future stock returns and future stock volatility based on established predictor variables from Welch and Goyal (2008). The studies by Moritz and Zimmermann (2016), Kelly et al. (2019), Gu et al. (2021), and Freyberger et al. (2020) all predict future stock returns based on firm characteristics and historical returns. However, they differ with respect to the specific ML methods applied. Grammig et al. (2020) construct a hybrid approach that combines traditional methods based on financial theory with ML to predict future excess stock returns. Chinco et al. (2019) apply LASSO to predict ultra-short-term future stock returns based on the cross-section of ultrashort-term historical returns. Akyildirim et al. (2021) use various ML methods to predict intraday excess returns based on high-frequency order and trade information. Amel-Zadeh et al. (2020) predict abnormal stock returns around earnings announcements based on financial statement variables. They use LASSO, random forests, and neural networks and analyse which financial statement variables are the most informative. Chinco et al. (2021) use ridge regression

[12]Most studies only focus on the predictions themselves. However, there are also some studies that try to analyse how the predictor variables affect the predictions. While most ML models do not allow for direct observation of how the algorithm generates its predictions, methods from the field of interpretable ML try to 'open the black box' (see, e.g., Murdoch et al., 2019).

TABLE 3 Overview of studies that use ML in economic prediction problems

This table reports an overview of relevant studies in finance that apply ML in economic prediction problems to reduce prediction error. There are three main categories of economic prediction problems for which ML is relevant: prediction of asset prices and trading mechanisms, prediction of credit risk, and prediction of firm outcomes and financial policy.

Category	Subcategory	Prediction targets
Prediction of asset prices and trading mechanisms	Equities	– Stock returns – Stock volatility – Stock covariance – Equity risk premium
	Bonds	– Future excess returns of US treasury bonds
	Foreign exchange	– Direction of changes in exchange rates
	Derivatives	– Prices of options on index futures – Prices of general derivatives
	General financial claims	– Stochastic discount factor – Financial crises
	Investors	– Mutual fund performance – Retail investors' portfolio allocations and performance
	Market microstructure	– Lifespan of trading orders – General microstructure variables
Prediction of credit risk	Consumer credit risk	– General consumer default – Credit card delinquency and default – Bill payment in developing countries – Credit card repayment patterns
	Real estate credit risk	– Mortgage loan risk – Commercial real estate default
	Corporate credit risk	– Firms' credit rating changes – Corporate bankruptcy – Fintech loan default – Recovery rates of corporate bonds
Prediction of firm outcomes and financial policy	Financial outcomes	– Capital structure – Earnings
	Corporate misconduct	– Accounting fraud – Regulatory violations
	Startups' success	– Startup acquisitions – Startup valuations and success probabilities

to determine the probability of encountering stock return anomalies. Feng et al. (2020) propose an ML-based method to evaluate the contribution of the plethora of potential risk factors in explaining stock returns. Two studies focus on financial market volatility: Kogan et al. (2009) predict future stock volatility based on annual reports; Osterrieder et al. (2020) predict the intraday volatility index VIX from option prices. Rossi and Timmermann (2015) use ML to study how stock returns and economic activity are related. They apply boosted regression trees to predict covariances between stock returns and a daily economic activity index.

In addition to predictions of individual stock returns, ML can reduce the prediction error in predicting aggregate stock market behaviour, particularly the equity risk premium. Jacobsen et al. (2019) predict the equity risk premium based on established stock market predictor variables from Welch and Goyal (2008) with an ensemble of multiple ML models. Routledge (2019) predicts the equity risk premium from macroeconomic indicators and FOMC texts. Adämmer and Schüssler (2020) extract topics discussed in general news articles with ML to predict the equity risk premium.

Some studies predict certain aspects of *bonds*. For instance, Bianchi et al. (2021) apply various ML methods to predict future excess returns of US treasury bonds from general yield data and macroeconomic indicators.

In the subcategory of *foreign exchange*, the study by Colombo et al. (2019) applies SVM to predict the direction of changes in exchange rates based on indicators of market uncertainty.

Other studies use ML to price *derivatives*, which is also an early application of ML in finance. Hutchinson et al. (1994) price options on the S&P 500 future based on the Black-Scholes variables with an early variant of neural networks. Similarly, Yao et al. (2000) price options on the Nikkei 225 future. In more recent work, De Spiegeleer et al. (2018) find that ML methods can price derivatives much faster than advanced mathematical models while achieving only slightly worse accuracy.

Instead of focusing on certain asset classes, there are also studies concerning *general financial claims*. Two studies directly predict the stochastic discount factor. Chen et al. (2019) use generative adversarial networks based on deep neural networks with different predictors, such as firm characteristics, historical returns, and macroeconomic indicators. Kozak et al. (2020) develop a custom ML method based on Bayesian priors to predict the stochastic discount factor from firm characteristics and historical returns. The study by Oh et al. (2006) applies ML to detect and predict financial crises from financial market volatility. Similarly, Coffinet and Kien (2019) develop an ML toolkit to detect banking crises.

In addition to asset prices and returns, prediction problems also arise in studies concerning retail and professional *investors'* trading decisions and performance. Li and Rossi (2020) apply boosted regression trees to predict mutual funds' performance, which then allows for fund selection. Rossi and Utkus (2020) study which type of retail investors benefit (the most) from robo-advising. More specifically, they apply boosted regression trees to predict changes in investors' portfolio allocations and performance.

Finally, some studies focus on predicting certain aspects of the *market microstructure* with ML. McInish et al. (2019) apply random forests to predict the lifespan of orders based on order characteristics and market data. Easley et al. (2021) predict a variety of variables relevant for market participants, such as bid-ask spreads, changes in volatility, and sequential return correlations from established microstructure measures with random forests.

3.2.2 | Prediction of credit risk

Credit risk is a typical economic prediction problem: its ultimate goal is to know which prospective borrowers will eventually default. As such, ML can lower prediction errors and improve decision making, such as in loan origination. We divide the current literature concerning ML-based predictions of credit risk into the following three subcategories: consumer credit risk, real estate credit risk, and corporate credit risk.

Studies on *consumer credit risk* apply ML to make default predictions for any type of consumer credit. Albanesi and Vamossy (2019) study general consumer credit default. They use advanced ML methods such as boosted regression trees and deep neural networks to derive more accurate predictions from credit bureau data compared to standard credit scoring models. Furthermore, they analyse which predictors are the most relevant and how the different predictors affect the predictions. Similarly, Tantri (2021) predicts consumer credit default with boosted regression trees based on borrower and loan characteristics data and finds that using ML-based default predictions can improve lending efficiency. Khandani et al. (2010) predict consumer credit card default based on transaction data and traditional credit bureau data. Similarly, Butaru et al. (2016) predict credit card default but consider more general account data and macroeconomic indicators. They both use tree-based ML methods that automatically consider nonlinearities and interactions between predictor variables. Butaru et al. (2016) also attempts to identify which predictor variables drive default predictions. Björkegren and Grissen (2018, 2020) focus on bill payment and apply random forests to mobile phone metadata to predict the payment of consumer bills in developing countries. The ability to make credit risk predictions based on easily obtainable data from mobile phones can help unbanked people in developing countries without a credit score obtain access to loans. Slightly different from the studies above, Gathergood et al. (2019) use credit card transaction data to predict credit card repayment patterns. They predict not whether customers pay their credit card bills but how customers split repayment on multiple cards with different interest rates. They also apply various ML methods and analyse which predictors are most informative.

Whenever algorithm-based decisions affect people, algorithmic bias is a potential issue. Since ML-based predictions of consumer credit risk directly affect credit approval decisions, it is necessary that the algorithm does not discriminate against people based on attributes such as gender or race. The literature does not paint a uniform picture of whether ML reduces or increases bias in consumer credit decisions. Rambachan, Kleinberg, Ludwig, et al. (2020) and Rambachan, Kleinberg, Mullainathan, et al. (2020) argue that discrimination by algorithms crucially depends on the given data. Since algorithms base their decisions on the data on which they have been trained, they might propagate biases present in the data. Fuster et al. (2022) apply ML to a concrete data set to create an ML model for credit decisions. They find that ML increases the disparity between and within different groups relative to simpler methods. In particular, it disadvantages Hispanic and Black borrowers compared to traditional approaches. Hence, awareness of the potential discrimination by ML-based algorithms is required if their predictions influence decisions that directly affect people, such as lending.

On the other hand, there are also studies showing that ML use can decrease bias in consumer credit decisions. Based on a theoretical model, Philippon (2019) shows how algorithms can reduce discrimination in credit markets. Dobbie et al. (2021) train an ML model to maximise expected profit from credit applications and find that the resulting lending decisions eliminate bias. Kleinberg et al. (2018) show that including problematic variables, such as gender and race, in ML models can actually reduce discrimination. To conclude the discussion concerning algorithmic bias in consumer credit risk, to date, there is no uniform picture in the literature. Some studies find that using ML to determine consumer credit risk increases bias, while other studies find that it decreases bias.

The second subcategory of ML-based credit risk predictions, *real estate credit risk*, involves the risk of mortgages and commercial real estate loans. Sadhwani et al. (2021) use deep neural networks to predict mortgage loan risk from mortgage origination and performance data and macroeconomic indicators. They also analyse which predictor variables are the most important and how they affect the predictions. Cowden et al. (2019) use various ML methods to predict commercial real estate default based on property characteristics.

Corporate credit risk is another area in which ML can provide superior credit risk predictions. Jones et al. (2015) predict firms' credit rating changes based on firm fundamentals, analyst forecasts, and macroeconomic indicators. Tian et al. (2015) and Sermpinis et al. (2022) directly predict corporate bankruptcy from firms' financial statements and market data. Lahmiri and Bekiros (2019) similarly predict bankruptcy from firm fundamentals but additionally include general risk indicators. They use more sophisticated neural networks. Croux et al. (2020) apply LASSO to predict fintech loan default from loan and borrower characteristics as well as macroeconomic indicators. In contrast to the above studies, Nazemi and Fabozzi (2018) focus on the time after credit default and predict the recovery rates of corporate bonds based on bond and industry characteristics and macroeconomic indicators with various ML methods.

3.2.3 | Prediction of firm outcomes and financial policy

The analysis of the determinants of specific firm outcomes (e.g., capital structure), as an important subject of study in the field of corporate finance, can also be the target of ML-based predictions. We divide the current literature in this category into the following three subcategories based on the specific target of the prediction: financial outcomes, corporate misconduct, and startups' success.

Two studies use ML to predict different *financial outcomes*. Amini et al. (2021) study firms' capital structure as a typical problem in corporate finance. They predict corporate leverage based on the standard capital structure determinants in the literature (Frank & Goyal, 2009) with various ML methods. Furthermore, they analyse which determinants are actually informative for capital structure and how they influence the predictions in detail. The study by van Binsbergen et al. (2020) applies random forests to predict firms' future earnings based on their accounting data, macroeconomic data, and analyst forecasts.

Corporate misconduct represents another typical prediction problem in the category of firm outcomes and financial policy. The most common type of corporate misconduct studied in the literature is accounting fraud. While traditional approaches can be used to predict accounting fraud (such as the Beneish, 1999 model of earnings manipulation), some studies argue that ML can provide superior prediction accuracy. Bao et al. (2020) apply boosted regression trees to raw financial statement variables to predict accounting fraud. They find that ML-based predictions outperform simpler existing fraud models. Brown et al. (2020) also predict accounting fraud by applying ML-based textual analysis to firms' annual reports. They further analyse which topics are the most informative and how they affect fraud predictions. Bertomeu et al. (2021) use boosted regression trees to predict material misstatements based on a large set of potential predictor variables. In addition to accounting fraud, Campbell and Shang (2022) apply textual analysis and ML to predict general violations of regulatory rules from firms' employee reviews on websites such as Glassdoor.

Finally, studies in the field of entrepreneurial finance use ML to predict *startups' success*. Xiang et al. (2012) apply ML-based textual analysis to predict startup acquisitions based on firms' fundamental data and firm-specific news. Similarly, Ang et al. (2022) predict startups' valuations and their probabilities of success with ML-based textual analysis and boosted regression trees.

3.3 | Extension of the existing econometric toolset

Studies of the third archetype of ML applications extend the existing econometric toolset. Many commonly used econometric methods contain a prediction component. For instance, the first stage of instrumental variable regression with 2SLS is effectively a prediction problem, as only the fitted (predicted) value of the instrumented variable enters the second stage. ML methods can provide superior predictions and hence improve the capabilities of such econometric methods. On the other hand, some ML methods already serve similar purposes as existing econometric methods. For instance, clustering is a known problem in econometrics and in ML. ML-based methods often provide superior performance, so they can directly extend the econometric toolset. Table 4 gives an overview of the literature on ML-based econometric methods. We distinguish between causal ML that uses ML for the estimation of treatment effects and other isolated applications of ML in econometrics. Within the category of causal ML, we further divide the literature into ML-enhanced methods for instrumental variable regression, novel methods of causal trees and causal forests, and other approaches related to causal ML. In the following, we briefly review the corresponding literature.

3.3.1 | Causal ML

While traditional econometric methods aim for causality, ML methods are designed for prediction or for data structure inference. The field of causal ML tries to combine the

TABLE 4 Overview of ML-based methods that extend the existing econometric toolset

This table reports the different categories of ML-based methods that extend the existing econometric toolset. The largest category is causal ML for the estimation of treatment effects. ML enhances existing methods, such as instrumental variable regression, or introduces new methods, such as causal trees and causal forests. ML also provides other methods relevant for the estimation of treatment effects, such as verifying the balance between treatment and control groups. The second category includes special applications of ML in econometric approaches in addition to treatment effects, such as the generation of simulated data.

Category	Subcategory	Approaches
Causal ML	Instrumental variable regression	– 2SLS first stage with LASSO, ridge regression, or neural networks
	Causal-tree based methods and applications	– Causal trees – Causal forests – Applications of causal forests
	Other causal ML	– Direct prediction of treatment effects – ML-based propensity score – Balance verification between treatment and control groups – Counterfactual prediction
Special Applications		– Predictive power of economic theories – Completeness of economic theories – Handling of imbalanced data – Generation of artificial data – ML-augmented preanalysis plans

advantages of both to create superior econometric methods suitable for causality and especially for the estimation of treatment effects. The most developed methods within causal ML are ML-enhanced instrumental variable regression and the novel methods of causal trees and forests.

As noted before, ML can directly improve the first stage of *instrumental variable regression*. By providing better predictions for the instrumented variable, the coefficient of determination R^2 of the first stage improves, resulting in more precise estimates in the second stage. Concrete implementations of this idea already exist for different ML methods, including LASSO (Belloni et al., 2012), ridge regression (Carrasco, 2012; Hansen & Kozbur, 2014), and neural networks (Hartford et al., 2017). However, Angrist and Frandsen (2022) argue that ML-enhanced instrumental variable methods might not be superior to existing specialised approaches in selecting instrumental variables.

For the estimation of treatment effects with ML, *causal trees and causal forests* are other well-developed methods. The seminal work by Athey and Imbens (2016) introduced the causal tree approach, which uses tree-based ML methods to partition data into subpopulations with different magnitudes of treatment effects. Causal forests proposed by Athey and Wager (2019) extend this concept by using an entire ensemble of causal trees. Some studies apply causal forests to concrete problems in finance. Gulen et al. (2020) apply causal forests to estimate heterogeneous treatment effects of debt covenant violations on firms' investment levels. O'Malley (2021) estimates the treatment heterogeneity of a legislative change in home repossession risk on mortgage default with causal forests.

In addition to causal trees and causal forests, *other approaches* use ML to improve the estimation of treatment effects. Lee et al. (2010) estimate the propensity score with ML. Mullainathan and Spiess (2017) suggest the use of ML to verify the balance between treatment and control groups. They argue that if it is possible to predict the treatment assignment with ML, then the split into treatment and control groups cannot be balanced. However, this idea works in only one direction: it is possible to infer imbalance but not balance by applying ML to predict the treatment assignment (since the chosen ML methods may not be powerful enough to predict the treatment assignment of imbalanced data). Chernozhukov et al. (2017, 2018) directly calculate treatment effects from ML-based predictions of treatment assignment and outcome. Finally, Athey et al. (2019) predict the counterfactual with ensemble methods to estimate treatment effects from panel data.

3.3.2 | Special applications of ML in econometrics

While causal ML for the estimation of treatment effects is currently the most developed application of ML in econometrics, there are various special applications of ML in econometrics that also extend the existing econometric toolset.

Above, we presented how ML can create measures of economic variables. By generalising this concept, ML can also construct a predictability measure of entire economic theories. Peysakhovich and Naecker (2017) introduce the notion that ML can be used to derive an upper bound of the predictive power of theories: the explainable variation in the dependent variable in a given data set with ML methods. Fudenberg et al. (2019) extend this idea to construct a completeness measure for economic theories. They calculate completeness by comparing two prediction errors: the error achieved from using the model and variables hypothesised by economic theory and the error achievable with ML. In general, different data sets contain different levels of information, so they allow different levels of predictability. By comparing prediction errors to those achievable with ML methods, it is possible to create a fairer and more informative measure for a comparison of different economic theories.

A different problem relevant in econometrics as well as in ML is imbalanced data. For instance, in loan performance data, actual defaults are much rarer than uneventful repayments. Sigrist and Hirnschall (2019) combine ML with traditional econometric methods to address such problem types. More specifically, they use boosted regression trees to enhance the traditional Tobit model. They also illustrate the advantages of their method in a concrete problem by applying it to loan defaults in Switzerland.

In the field of simulation, Athey et al. (2021) use generative adversarial networks instead of traditional Monte Carlo methods to simulate data that more closely mimic real data. They illustrate their method by using simulated data for performance comparisons across different econometric estimators. Adams, Kräussl, et al. (2021) use deep neural networks to generate artificial paintings to study gender discrimination in art prices.

Finally, Ludwig et al. (2019) introduce ML-augmented preanalysis plans to avoid p-hacking. They augment standard linear regression with new regressors from ML. The new regressors aggregate many potentially relevant variables into a single index. Hence, their method avoids the otherwise necessary prespecification of concrete analysis choices in standard preanalysis plans.

4 | FUTURE PROSPECTS OF ML IN FINANCE

The benefits of ML over traditional methods as illustrated above together with the existing but still limited number of ML applications in finance suggest a still mostly untapped potential for future research. However, it is unclear whether the usage of ML methods will actually gain broad popularity in the finance community. Furthermore, prospective users of ML need to know whether ML applications can also reach the most prestigious journals of the profession or if they tend to be published only in specialty journals. Finally, the different application categories of ML described by our taxonomy and the wide variety of research fields in finance make it difficult to pinpoint exactly where the most promising applications of ML in finance research lie. In this section, we give indicative answers to these questions by systematically analysing the existing finance literature that already uses ML methods. In particular, we investigate the publication success of such papers and how it differs by research field and application type. Our results may not only indicate the future prospects of ML in finance but also show where and how researchers can apply ML to maximise its future potential.

4.1 | Sample of finance research papers that apply ML

For a systematic analysis of the existing finance research that applies ML, we begin by constructing a sample of relevant publications. We build our sample by focusing on research papers that have been published in major finance journals. As our starting point, we choose the 45 most highly ranked finance journals (categories A+, A and B) of the journal ranking of the German Academic Association of Business Research (VHB-JOURQUAL3).[13] Then,

[13]In an alternative approach, we choose the 37 journals that are ranked as 4*, 4 or 3 within the finance category of the AJG 2018 ranking of the Chartered Association of Business Schools. Those ranks are largely comparable to the A+, A and B ranks of the VHB-JOURQUAL3 ranking. Our results remain qualitatively unchanged when using this alternative set of journals.

we visit each journal website and download all papers that have been published in the years 2010 to 2021 and that contain any of the following keywords either in the title, abstract, or full text:

- *General ML-related terms*: 'machine learning', 'big data', 'artificial intelligence'
- *ML method categories*: 'supervised learning', 'unsupervised learning', 'reinforcement learning', 'semisupervised learning'
- *Specific ML methods*: 'lasso', 'ridge', 'elastic net', 'decision tree', 'random forest', 'boosted regression trees', 'gradient boosting', 'support vector machine', 'support vector classification', 'support vector regression', 'neural network', 'naïve bayes'

We read each paper in this initial sample and manually exclude papers that do not use machine learning in any part of their analysis (for instance, if they mention the keyword(s) above only while describing the work of others). Finally, we arrive at a sample that consists of 346 papers.

To investigate possible differences in publication success by research field and application type, we classify each paper in both dimensions. For the classification by research field, we make use of JEL codes.[14] In the few cases where EconLit provides no JEL codes or if none of the provided codes fall into the financial economics code range (G), we instead use author-provided JEL codes obtained directly from the papers. We then classify each paper in our sample into exactly one of the five JEL subfields within financial economics (G1–G5 code range).[15] Since some papers carry multiple JEL codes, we manually classify 68 papers in our sample for which the subfield assignment is ambiguous. In 29 cases, we can resolve the ambiguity by choosing the subfield according to the majority of a paper's JEL codes. In the remaining 39 cases, we manually assign the most appropriate subfield.

Regarding the classification by application type, we inspect each paper's methodology in detail and then classify it into one of the three archetypes of our taxonomy described in Section 3: (i) superior and novel measures, (ii) economic prediction problems, and (iii) new econometric tools.

4.2 | How promising are ML applications in finance?

To provide indications of the future prospects of ML applications in finance, we first analyse the journals in which the existing ML applications have been published. Figure 5 illustrates the large growth in the usage of ML. In 2018, the number of publications that used ML more than tripled compared to the previous years' average. In 2019, the increase was more than fivefold. In 2020, there were almost seven times as many publications using ML than before, and in 2021 we found an almost elevenfold increase in the number of published ML papers.

While the strong growth in the number of finance publications that apply ML over the last few years shows a clear trend toward an increasing usage of ML, the question of whether ML

[14]To obtain the JEL codes of the papers in our sample, we use the EconLit database from the AEA. The JEL codes from EconLit are assigned by professional staff, ensuring systematic classification criteria and maximal coverage (Falk & Andre, 2021).

[15]JEL codes are structured hierarchically and consist of one letter and two digits (e.g., G35), where the *letter* refers to the general field in economics (e.g., G for financial economics), the *first digit* describes the subfield (e.g., G3 for corporate finance), and the *second digit* determines the specific area within a subfield (e.g., G35 for payout policy).

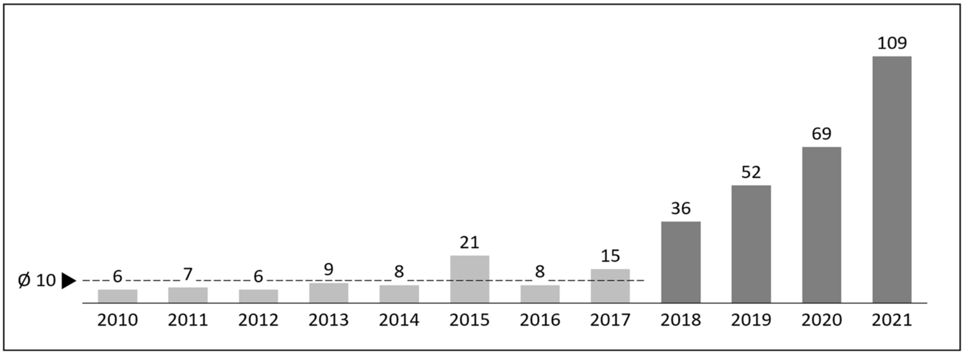

FIGURE 5 Number of relevant publications in finance that apply ML by year. This figure depicts how the number of papers that apply ML and have been published in major finance journals has evolved over time. Since 2018, we observe a strong increase in ML publications compared to the average of the previous years.

applications have the potential to be published in the most prestigious journals of the profession remains unanswered. Panel A in Table 5 shows how the number of ML publications has evolved over time by journal rank. In the years until 2017, the few early ML applications were published mostly in journals ranked as B. Since 2018, however, a significant portion of the ML publications appeared in the highest-ranked journals. To control for the fact that there exist many more lower-ranked than higher-ranked journals (and thus publications in the respective journals), Panel B reports the share of ML publications relative to the total number of publications that major finance journals of different ranks published each year. The results show that the strong increase in the number of ML publications was not driven by a general increase in the number of papers that journals of any rank have published; similar to the absolute numbers, the relative share of publications that use ML has increased similarly in total and for each journal rank.[16] In 2021, there are no meaningful differences in the relative share of ML publications across journal ranks: approximately 3%–4% of the publications used ML in 2021 independent of the journal rank.[17]

Our results in this section give two main indications of the future prospects of ML in finance. First, there is steady and robust growth in the number of finance publications that apply ML. It is likely that this trend will continue with even more ML applications in the years ahead. The benefits of ML illustrated above and the continuing increase in relevance of ML outside of academia also leave little reason to expect otherwise. Second, researchers who apply ML in finance can reasonably expect their papers to have the potential to reach the highest-ranking journals of the profession. Not only are there currently numerous examples of ML applications in such journals, but their relative share has now reached a level that is comparable to lower-ranked journals. Hence, these results may suggest a bright and promising future for ML applications in finance.

[16]We conduct two-sample *t*-tests for the differences between the 2010–2017 share of ML papers across journals of the three different journal categories (A+, A, and B). In the 2010–2017 period, we detect a statistically significant difference between the share of ML papers in B ranked journals (0.5%) and that in A+ and A ranked journals (0.2%/0.3%) at the 5% level. In the 2018-2021 period, this statistically significant difference disappears.

[17]Notably, the total number of publications also includes theory papers and other methodologies. The share of ML papers among empirical studies would be even higher.

TABLE 5 Yearly number of relevant finance publications that apply ML and their share relative to all publications in major finance journals

This table reports the number of ML publications over time by journal rank. Panel A reports the absolute number of papers that apply ML and have been published in major finance journals per year in total and by journal rank. Panel B reports the share of these ML applications relative to the number of all publications in major finance journals per year in total and by journal rank. The means in the years 2010–2017 and 2018–2021 in Panel B are weighted by the number of publications. a, b or c denote statistical significance of differences in proportions at the 5% level for the groups A+/A, A+/B and A/B, respectively.

Year	2010	2011	2012	2013	2014	2015	2016	2017	*Mean* 2010–2017	2018	2019	2020	2021	*Mean* 2018–2021	Total
Panel A: Number of ML publications in major finance journals															
Total	6	7	6	9	8	21	8	15	10	36	52	69	109	66.5	346
A+	0	0	0	2	0	2	0	1	0.6	4	8	8	14	8.5	39
A	4	1	0	1	2	6	1	1	2.0	8	7	12	19	11.5	62
B	2	6	6	6	6	13	7	13	7.4	24	37	49	76	46.5	245
Panel B: Share of ML publications relative to all publications in major finance journals															
Total	0.3%	0.3%	0.3%	0.4%	0.3%	0.9%	0.3%	0.6%	0.4%	1.4%	2.0%	2.3%	3.4%	2.3%	1.1%
A+	0.0%	0.0%	0.0%	0.7%	0.0%	0.7%	0.0%	0.3%	0.2%[b]	1.3%	2.4%	2.2%	3.4%	2.4%	1.1%
A	0.6%	0.2%	0.0%	0.1%	0.3%	0.8%	0.2%	0.2%	0.3%[c]	1.2%	1.2%	1.9%	3.2%	1.9%	0.8%
B	0.2%	0.5%	0.5%	0.5%	0.4%	0.9%	0.4%	0.8%	0.5%[b, c]	1.5%	2.1%	2.5%	3.5%	2.5%	1.3%

4.3 | Which kinds of ML applications in finance are most promising?

In the previous section, we showed that ML applications have seen strong time-series growth in the most prestigious finance journals over the last several years. We now move on to the question of what makes certain applications more promising than others with regard to publication success. To answer this question, we first investigate differences in the distribution of ML publications by *research field* and across *journal ranks*; and then, subsequently apply the classification from our *taxonomy* (see Section 3) as a third dimension (*methodological purpose*) to the analysis.

In Table 6, we begin with examining the distribution of ML publications by *research field*. Column 1 shows that most ML publications (to date) belong to the *general financial markets* (G1) category (71.1%), which consists of asset pricing and related areas. Considerably fewer ML publications have been published in the fields of *financial institutions and services* (G2, 13.6%) and *corporate finance and governance* (G3, 14.2%). There is a very small share of ML publications in *behavioural finance* (G4, 0.9%) and *household finance* (G5, 0.3%).

To account for heterogeneity in the distribution of *all published* finance papers by research field, we compare the distribution of ML publications to that of all publications in major finance journals. This comparison is crucial if the general financial markets (G1) category also represents the largest field in major finance journals. If so, the previous result could be simply driven by a large number of publications that belong to the general financial markets category (G1). Therefore, Column 5 shows the distribution of *all* (2010–2021) publications across fields,[18] which we then compare with the distribution of ML publications across fields. Visual inspection of Columns 1 and 5 already suggests that even after accounting for research field effects, ML papers are significantly more likely in the *general financial markets* category compared to other fields. A Pearson χ^2-test, which tests for systematic differences of two distributions with categorical variables, confirms this observation at every plausible level of significance (see last row of Table 6). In additional analyses using z-tests for differences in proportions, Column 9 shows that the distribution of ML publications is much more concentrated with a substantially higher share of ML (relative to all) papers in the field of *general financial markets* (G1: 71.1% vs. 47.1%, z-stat: 8.84) and a lower share of papers in the fields of *financial institutions and services* (G2: 13.6% vs. 25.4%, z-stat: −5.03) and *corporate finance and governance* (G3: 14.2% vs. 27.3%, z-stat: −5.44). In the fields of *behavioural finance* (G4) and *household finance* (G5), the sample sizes are too small to draw any economically meaningful conclusions. We repeat our analysis for each of the three journal ranking categories (A+ , A and B) in Columns 10–12 and find qualitatively similar results.

Second, we examine the distribution of ML publications by the *methodological purpose* (see our taxonomy, Section 3). Table 7 (Panel A, Column 1) shows the distribution for the full sample of ML publications across all fields. A large majority of publications (69.1%) apply ML to reduce the prediction error in economic prediction problems. Using ML to construct superior and novel measures is much less widespread on average (25.1%). Very few finance publications (5.8%) use ML to extend the econometric toolset.[19] Columns 2-4 reveal that there is strong heterogeneity by *journal rank*. Specifically, publications in the highest-ranked journals (A+) use ML disproportionally more often to construct *superior and novel measures* compared to

[18]We obtain data for all finance publications in the 45 major finance journals (ranked as A+, A, or B according to the VHB-JOURQUAL3 rating) for the years 2010 to 2021 from EconLit. We classify each paper into one of the five JEL subfields within financial economics (G1–G5 code range) with the procedure described in Section 4.1.

[19]Note that the number of papers in our sample that apply ML to extend the econometric toolset is low mainly because we only consider papers from finance journals and therefore ignore contributions from the econometrics literature.

TABLE 6 Distribution of ML applications in finance by research field and comparison to all publications in major finance journals

This table reports the distribution of ML research applications in major finance journals by research field (single-digit JEL categories). The first column reports the results for the entire sample, while Columns 2–4 report the results separately for publications in journals ranked as A+, A, and B. Columns 5–8 report the same results for all publications in major finance journals. The last four columns report the *z*-statistics of *z*-tests for the difference in proportions. ***, ** or * denote statistical significance at the 1%, 5%, or 10% level.

	ML publications in major finance journals				**All publications in major finance journals**				**z-stat for difference**			
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(1)–(5)**	**(2)–(6)**	**(3)–(7)**	**(4)–(8)**
	All	**A+**	**A**	**B**	**All**	**A+**	**A**	**B**	**All**	**A+**	**A**	**B**
General financial markets (G1)	71.1%	66.7%	53.2%	76.3%	47.1%	38.0%	36.1%	55.5%	8.84***	3.67***	2.78***	6.49***
Financial institutions and services (G2)	13.6%	12.8%	32.3%	9.0%	25.4%	23.5%	31.3%	23.1%	−5.03***	1.56	0.15	−5.21***
Corporate finance and governance (G3)	14.2%	20.5%	14.5%	13.1%	27.3%	38.3%	32.4%	21.3%	−5.44***	−2.27**	−2.99***	−3.12***
Behavioural finance (G4)	0.9%	0.0%	0.0%	1.2%	0.0%	0.0%	0.0%	0.0%	9.72***	NA	−0.11	9.64***
Household finance (G5)	0.3%	0.0%	0.0%	0.4%	0.1%	0.3%	0.2%	0.1%	0.73	−0.35	−0.31	1.75***
Nobs	346	39	62	245	18,605	3,241	5,089	10,275	-	-	-	-
Chi-squared statistics (*p* value):	-	-	-	-	-	-	-	-	0.000***	0.004***	0.025**	0.000***

TABLE 7 Distribution of ML applications in finance by application type for the entire sample and for publications in the different journal ranks

This table reports the distribution of ML research applications in major finance journals by application type from our taxonomy. Panel A reports the results across all research fields. Panel B reports the results for each research field separately. The first column reports the results for the entire sample, while Columns 2–4 report the results separately for publications in journals ranked as A+, A, and B. a, b, or c denote statistical significance of differences in proportions at the 5% level for the groups 'A+ versus A', 'A+ versus B' and 'A versus B', respectively.

	All (1)	**A+** (2)	**A** (3)	**B** (4)
Panel A: Distribution of application types				
	n = 346	*n* = 39	*n* = 62	*n* = 245
Superior and novel measures	25.1%	56.4%[a, b]	32.3%[a, c]	18.4%[b, c]
Economic prediction problems	69.1%	38.5%[a, b]	62.9%[a, c]	75.5%[b, c]
New econometric tools	5.8%	5.1%	4.8%	6.1%
Panel B: Distribution of application types in each research field				
General financial markets (G1)	*n* = 246	*n* = 26	*n* = 33	*n* = 187
Superior and novel measures	22.0%	38.5%[b]	33.3%[c]	17.6%[b, c]
Economic prediction problems	71.1%	57.7%	63.6%	74.3%
New econometric tools	6.9%	3.8%	3.0%	8.0%
Financial institutions and services (G2)	*n* = 47	*n* = 5	*n* = 20	*n* = 22
Superior and novel measures	29.8%	80.0%[a, b]	30.0%[a]	18.2%[b]
Economic prediction problems	66.0%	0.0%[a, b]	65.0%[a]	81.8%[b]
New econometric tools	4.3%	20.0%[b]	5.0%	0.0%[b]
Corporate finance and governance (G3)	*n* = 49	*n* = 8	*n* = 9	*n* = 32
Superior and novel measures	32.7%	100.0%[a, b]	33.3%[a]	15.6%[b]
Economic prediction problems	65.3%	0.0%[a, b]	55.6%[a]	84.4%[b]
New econometric tools	2.0%	0.0%	11.1%	0.0%
Behavioural finance (G4)	*n* = 3	*n* = 0	*n* = 0	*n* = 3
Superior and novel measures	100.0%	NA	NA	100.0%
Economic prediction problems	0.0%	NA	NA	0.0%
New econometric tools	0.0%	NA	NA	0.0%
Household finance (G5)	*n* = 1	*n* = 0	*n* = 0	*n* = 1
Superior and novel measures	0.0%	NA	NA	0.0%
Economic prediction problems	100.0%	NA	NA	100.0%
New econometric tools	0.0%	NA	NA	0.0%

publications in lower-ranked journals (56.4% vs. 32.3% and 18.4%). These differences are statistically significant at the 5% level using z-tests for differences in proportions between journal rank categories. On the other hand, *economic prediction problems* are less prevalent in the highest-ranked journals (38.5% vs. 62.9% and 75.5%), which is again statistically significant.

To detect differences in the publication success of application types across research fields, we repeat the previous analysis for each research field separately in Panel B of Table 7. Specifically, we are interested in identifying systematic patterns across research fields, for example, if *superior and novel measures* are more likely to be successful in specific fields of finance. As Panel B, Column 1 shows, *superior and novel measures* are disproportionally more often used in the *financial institutions* (G2) and *corporate finance* (G3) literatures (29.8% and 32.7% vs. 25.1%). Interestingly, within these two fields, publications in journals ranked as A + (Column 2) almost exclusively use ML to construct *superior and novel measures* (80.0% and 100.0%).

4.3.1 | Analysis by citations

To further corroborate our findings, we analyse citations as an alternative measure of publication success.[20] We obtain the number of citations from Web of Science (as of 19 Sep 2022) for each ML publication in our sample and compare it to the average number of citations for all papers published in major finance journals. Given that a paper's number of citations (as of 19 Sep 2022) naturally depends on the time since publication, we demean the number of citations in the following way: for each ML publication in our sample, we calculate *excess citations*, which is the difference between a paper's actual number of citations and the average number of citations of all publications in major finance journals from the same year.[21] We then study differences in excess citations by research field and application type and conduct t-tests against the null hypothesis that excess citations are statistically indistinguishable from zero (i.e., there are no differences in citation counts between ML publications and all publications from a given year). Table 8 shows our results. Overall, ML publications receive 3.0 more citations than the average publication in major finance journals from the same year, which is statistically significant at the 10% level. Across application types, publications that use ML to construct superior and novel measures receive 10.2 more citations than general publications in major finance journals, which is highly significant at the 1% level. Across fields, ML publications in corporate finance/governance receive 7.6 more citations than general publications in major finance journals, which is significant at the 5% level. Finally, publications that apply ML to construct superior and novel measures related to corporate finance/governance show the highest potential with regard to citation count as they receive 24.2 more citations, which is also highly significant at the 1% level. Given that the average ML publication in our sample has been cited 16.2 times, these effects are not only statistically significant but also economically large.[22,23]

[20]We thank an anonymous referee for encouraging this analysis.

[21]We obtain citation data to calculate average citation counts per year from Web of Science.

[22]In untabulated analyses, we account for possible unobserved year-level heterogeneity in citation growth across fields (for instance, if citations after publication grow stronger in certain fields) by demeaning citation counts by year-and-field averages. Our results are qualitatively and statistically similar when conducting this alternative analysis.

[23]A second possible alternative to analysing total citation counts is to analyse the ranking of journals that cite ML publications. In untabulated analyses, we show that publications that use ML to construct superior and novel measures tend to be cited from higher-ranked journals. Again, this effect is especially pronounced in the field of corporate finance and governance. These additional analyses of citations thus support our main findings. The detailed results are available from the authors upon request. We thank an anonymous referee for suggesting this analysis.

TABLE 8 Mean excess citations of ML publications relative to all publications in major finance journals

This table reports the mean excess citations of ML publications by field and application type. Excess citations are defined as the difference between actual citations and the average number of citations for all publications in major finance journals from the same year. Citation data come from Web of Science as of 19 Sep, 2022. ***, ** or * denote statistical significance at the 1%, 5% or 10% level.

		All types	Superior and novel measures	Economic prediction problems	New econometric tools
Full Sample	$n = 346$	3.0*	10.2***	1.2	−7.0*
By field					
General financial markets (G1)	$n = 246$	2.3	9.3**	1.1	−8.0**
Financial institutions and services (G2)	$n = 47$	2.4	1.0	3.7	−8.2
Corporate finance and governance (G3)	$n = 49$	7.6**	24.2***	−0.9	13.7
Behavioural finance (G4)	$n = 3$	−5.3**	−5.3**	NA	NA
Household finance (G5)	$n = 1$	−1.5	NA	−1.5	NA

In sum, the results from the citation analysis are consistent with the results from the previous analysis using journal ranks and thus provide corroborating evidence.

Our findings in this section yield three important conclusions. First, the usage of ML to construct superior and novel measures seems to be one application type with strong future potential. While most publications to date apply ML to economic prediction problems, papers that use ML for superior and novel measures have appeared in higher-ranked journals and receive more citations. Second, papers that apply ML in the field of corporate finance and governance seem to benefit from ML's ability to produce superior and new measures. Finally, the scarcity of existing research in the fields of behavioural finance and household finance indicates another attractive avenue for future ML applications.

5 | CONCLUSION

In this paper, we studied the question of how researchers can leverage ML technology in finance. First, we established that different types of ML solve different problems than traditional linear regression with OLS. While the properties of OLS are beneficial for explanation problems, supervised ML is the superior method for prediction problems. As we illustrated with a real estate asset pricing prediction problem, ML-based price predictions can achieve substantially lower pricing errors than OLS.

In the second part of this paper, we developed the following taxonomy of ML applications in finance: (1) construction of superior and novel measures, (2) reduction of prediction error in economic prediction problems, and (3) extension of the existing econometric toolset. This taxonomy serves multiple purposes. First, it enables a systematic review of the existing ML literature in finance. Second, it enables a better understanding of new contributions and how they relate to the existing literature. Finally, it may guide researchers in discovering possible applications and thus may facilitate new ML studies in finance.

In the final part, we provided indications of the future prospects of ML applications in finance by analysing the ML papers published in major finance journals. Over the last few years, there has been a strong growth in the number of ML applications in finance, and many of these applications reached the highest-ranked journals of the profession. Our results suggest that ML may become even more widespread in finance research in the coming years. They also indicate a particularly large potential of applying ML to unconventional data to construct superior and novel measures of topics related to the field of corporate finance and governance. The fields of behavioural and household finance may also offer a mostly untapped potential for ML in future research.

ACKNOWLEDGEMENTS

Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

Adämmer, P., & Schüssler, R. A. (2020). Forecasting the equity premium: Mind the news! *Review of Finance*, *24*, 1313–1355.

Adams, R. B., Akyol, A. C., & Grosjean, P. A. (2021). *Corporate gender culture* (SSRN Working Paper No. 3880650).

Adams, R. B., Kräussl, R., Navone, M., & Verwijmeren, P. (2021). Gendered prices. *The Review of Financial Studies*, *34*, 3789–3839.

Adams, R. B., Ragunathan, V., & Tumarkin, R. (2021). Death by committee? An analysis of corporate board (sub-) committees. *Journal of Financial Economics*, *141*, 1119–1146.

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on management of data* (pp. 207–216).

Akansu, A., Cicon, J., Ferris, S. P., & Sun, Y. (2017). Firm performance in the face of fear: How CEO moods affect firm performance. *Journal of Behavioral Finance*, *18*, 373–389.

Akyildirim, E., Nguyen, D. K., Sensoy, A., & Sikic, M. (2021). Forecasting high-frequency excess stock returns via data analytics and machine learning. *European Financial Management* (Forthcoming).

Alan, N. S., Karagozoglu, A. K., & Zhou, T. (2021). Firm-level cybersecurity risk and idiosyncratic volatility. *The Journal of Portfolio Management*, *47*, 110–140.

Albanesi, S., & Vamossy, D. F. (2019). *Predicting consumer default: A deep learning approach* (NBER Working Paper No. 26165).

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)*, (pp. 1–6).

Algaba, A., Ardia, D., Bluteau, K., Borms, S., & Boudt, K. (2020). Econometrics meets sentiment: An overview of methodology and applications. *Journal of Economic Surveys*, *34*, 512–547.

Amel-Zadeh, A., Calliess, J.-P., Kaiser, D., & Roberts, S. (2020). *Machine learning-based financial statement analysis* (SSRN Working Paper No. 3520684).

Amini, S., Elmore, R., Öztekin, Ö., & Strauss, J. (2021). Can machines learn capital structure dynamics? *Journal of Corporate Finance*, *70*, 102073.

Ang, Y. Q., Chia, A., & Saghafian, S. (2022). Using machine learning to demystify startups funding, post-money valuation, and success. In V. Babich, J. R. Birge, & G. Hilary (Eds.), *Innovative technology at the interface of finance and operations* (pp. 271–296). Springer.

Angrist, J. D., & Frandsen, B. (2022). Machine labor. *Journal of Labor Economics*, *40*, S97–S140.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of Internet stock message boards. *The Journal of Finance*, *59*, 1259–1294.

Athey, S., Bayati, M., Imbens, G., & Qu, Z. (2019). Ensemble methods for causal effects in panel data settings. *AEA Papers and Proceedings*, *109*, 65–70.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 7353–7360.

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

Athey, S., Imbens, G. W., Metzger, J., & Munro, E. (2021). Using Wasserstein generative adversarial networks for the design of Monte Carlo simulations. *Journal of Econometrics* (Forthcoming).

Athey, S., & Wager, S. (2019). Estimating treatment effects with causal forests: An application. *Observational Studies*, *5*, 37–51.

Azimi, M., & Agrawal, A. (2021). Is positive sentiment in corporate annual reports informative? Evidence from deep learning. *The Review of Asset Pricing Studies*, *11*, 762–805.

Aziz, S., Dowling, M., Hammami, H., & Piepenbrink, A. (2022). Machine learning in finance: A topic modeling approach. *European Financial Management*, *28*, 744–770.

Aziz, S., & Dowling, M. (2019). Machine learning and AI for risk management. In T. Lynn, J. G. Mooney, P. Rosati, & M. Cummins (Eds.), *Disrupting finance: FinTech and strategy in the 21*st *century* (pp. 33–50). Palgrave.

Bandiera, O., Prat, A., Hansen, S., & Sadun, R. (2020). CEO behavior and firm performance. *Journal of Political Economy*, *128*, 1325–1369.

Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. *Journal of Accounting Research*, *58*, 199–235.

Barbon, A., Di Maggio, M., Franzoni, F., & Landier, A. (2019). Brokers and order flow leakage: Evidence from fire sales. *The Journal of Finance*, *74*, 2707–2749.

Barth, A., Mansouri, S., & Woebbeking, F. (2020). *'Let me get back to you'—A machine learning approach to measuring non-answers* (SSRN Working Paper No. 3567724).

Bartov, E., Faurel, L., & Mohanram, P. S. (2018). Can twitter help predict firm-level earnings and stock returns? *The Accounting Review*, *93*, 25–57.

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, *80*, 2369–2429.

Beneish, M. D. (1999). The detection of earnings manipulation. *Financial Analysts Journal*, *55*, 24–36.

Bertomeu, J., Cheynel, E., Floyd, E., & Pan, W. (2021). Using machine learning to detect misstatements. *Review of Accounting Studies*, *26*, 468–519.

von Beschwitz, B., Keim, D. B., & Massa, M. (2020). First to 'read' the news: News analytics and algorithmic trading. *The Review of Asset Pricing Studies*, *10*, 122–178.

Bianchi, D., Büchner, M., & Tamoni, A. (2021). Bond risk premiums with machine learning. *The Review of Financial Studies*, *34*, 1046–1089.

van Binsbergen, J. H., Han, X., & Lopez-Lira, A. (2020). *Man versus machine learning: The term structure of earnings expectations and conditional biases* (NBER Working Paper No. 27843).

Björkegren, D., & Grissen, D. (2018). The potential of digital credit to bank the poor. *AEA Papers and Proceedings*, *108*, 68–71.

Björkegren, D., & Grissen, D. (2020). Behavior revealed in mobile phone usage predicts credit repayment. *The World Bank Economic Review*, *34*, 618–634.

Boudoukh, J., Feldman, R., Kogan, S., & Richardson, M. (2019). Information, trading, and volatility: Evidence from firm-specific news. *The Review of Financial Studies*, *32*, 992–1033.

Breaban, A., & Noussair, C. N. (2018). Emotional state and market behavior. *Review of Finance*, *22*, 279–309.

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

Brown, N. C., Crowley, R. M., & Elliott, W. B. (2020). What are you saying? Using topic to detect financial misreporting. *Journal of Accounting Research*, *58*, 237–291.

Bubb, R., & Catan, E. M. (2021). The party structure of mutual funds. *The Review of Financial Studies*, *35*, 2839–2878.

Bubna, A., Das, S. R., & Prabhala, N. (2020). Venture capital communities. *Journal of Financial and Quantitative Analysis*, *55*, 621–651.

Buehlmaier, M. M. M., & Whited, T. M. (2018). Are financial constraints priced? Evidence from textual analysis. *The Review of Financial Studies*, *31*, 2693–2728.

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, *70*, 245–317.

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking & Finance*, *72*, 218–239.

Calomiris, C. W., & Mamaysky, H. (2019). How news and its context drive risk and returns around the world. *Journal of Financial Economics*, *133*, 299–336.

Campbell, D. W., & Shang, R. (2022). Tone at the bottom: Measuring corporate misconduct risk from the text of employee reviews. *Management Science*, *68*, 7034–7053.

Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics*, *170*, 383–398.

Cathcart, L., Gotthelf, N. M., Uhl, M., & Shi, Y. (2020). News sentiment and sovereign credit risk. *European Financial Management*, *26*, 261–287.

Chen, L., Pelger, M., & Zhu, J. (2019). *Deep learning in asset pricing* (SSRN Working Paper No. 3350138).

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/Neyman machine learning of treatment effects. *American Economic Review*, *107*, 261–265.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*, C1–C68.

Chinco, A., Clark-Joseph, A. D., & Ye, M. (2019). Sparse signals in the cross-section of returns. *The Journal of Finance*, *74*, 449–492.

Chinco, A., Neuhierl, A., & Weber, M. (2021). Estimating the anomaly base rate. *Journal of Financial Economics*, *140*, 101–126.

Choudhury, P., Wang, D., Carlson, N. A., & Khanna, T. (2019). Machine learning approaches to facial and text analysis: discovering CEO oral communication styles. *Strategic Management Journal*, *40*, 1705–1732.

Coffinet, J., & Kien, J.-N. (2019). Detection of rare events: A machine learning toolkit with an application to banking crises. *The Journal of Finance and Data Science*, *5*, 183–207.

Colombo, E., Forte, G., & Rossignoli, R. (2019). Carry trade returns with support vector machines. *International Review of Finance*, *19*, 483–504.

Cowden, C., Fabozzi, F. J., & Nazemi, A. (2019). Default prediction of commercial real estate properties using machine learning techniques. *The Journal of Portfolio Management*, *45*, 55–67.

Croux, C., Jagtiani, J., Korivi, T., & Vulanovic, M. (2020). Important factors determining fintech loan default: Evidence from a Lendingclub consumer platform. *Journal of Economic Behavior & Organization*, *173*, 270–296.

Dávila, A., & Guasch, M. (2022). Managers' body expansiveness, investor perceptions, and firm forecast errors and valuation. *Journal of Accounting Research*, *60*, 517–563.

Dixon, M. F., Halperin, I., & Bilokon, P. (2020). *Machine learning in finance: From theory to practice*. Springer International Publishing.

Dobbie, W., Liberman, A., Paravisini, D., & Pathania, V. (2021). Measuring bias in consumer lending. *The Review of Economic Studies*, *88*, 2799–2832.

Domingues, R., Filippone, M., Michiardi, P., & Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognition*, *74*, 406–421.

Du, Q., Jiao, Y., Ye, P., & Fan, W. (2019). *When mutual fund managers write confidently* (SSRN Working Paper No. 3513288).

Easley, D., López de Prado, M., O'Hara, M., & Zhang, Z. (2021). Microstructure in the machine age. *The Review of Financial Studies*, *34*, 3316–3363.

Erel, I., Stern, L. H., Tan, C., & Weisbach, M. S. (2021). Selecting directors using machine learning. *The Review of Financial Studies*, *34*, 3226–3264.

Erkek, M., Cayirli, K., & Hepsen, A. (2020). Predicting house prices in Turkey by using machine learning algorithms. *Journal of Statistical and Econometric Methods*, *9*, 31–38.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD*, *96*, 226–231.

Falk, A., & Andre, P. (2021). *What's worth knowing? Economists' opinions about economics* (SSRN Working Paper No. 3885426).

Feng, G., Giglio, S., & Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance*, *75*, 1327–1370.

Frank, M. Z., & Goyal, V. K. (2009). Capital structure decisions: Which factors are reliably important? *Financial Management*, *38*, 1–37.

Freyberger, J., Neuhierl, A., & Weber, M. (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies*, *33*, 2326–2377.

Fudenberg, D., Kleinberg, J., Liang, A., & Mullainathan, S. (2019). *Measuring the completeness of theories* (SSRN Working Paper No. 3018785).

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance*, *77*, 5–47.

Gathergood, J., Mahoney, N., Stewart, N., & Weber, J. (2019). How do individuals repay their debt? The balance-matching heuristic. *American Economic Review*, *109*, 844–875.

Giannini, R., Irvine, P., & Shu, T. (2018). Nonlocal disadvantage: An examination of social media sentiment. *The Review of Asset Pricing Studies*, *8*, 293–336.

Goodell, J. W., Kumar, S., Lim, W. M., & Pattnaik, D. (2021). Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance*, *32*, 100577.

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

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*, 139–144.

Gow, I. D., Kaplan, S. N., Larcker, D. F., & Zakolyukina, A. A. (2016). *CEO personality and firm policies* (NBER Working Paper No. 22435).

Grammig, J., Hanenberg, C., Schlag, C., & Sönksen, J. (2020). *Diverging roads: Theory-based vs. machine learning-implied stock risk premia* (SSRN Working Paper No. 3536835).

Gu, C., & Kurov, A. (2020). Informational role of social media: Evidence from Twitter sentiment. *Journal of Banking & Finance*, *121*, 105969.

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

Gu, S., Kelly, B., & Xiu, D. (2021). Autoencoder asset pricing models. *Journal of Econometrics*, *222*, 429–450.

Gulen, H., Jens, C., & Page, T. B. (2020). *An application of causal forest in corporate finance: How does financing affect investment?* (SSRN Working Paper No. 3583685).

Guliker, E., Folmer, E., & van Sinderen, M. (2022). Spatial determinants of real estate appraisals in The Netherlands: A machine learning approach. *ISPRS International Journal of Geo-Information*, *11*, 125.

Hanley, K. W., & Hoberg, G. (2019). Dynamic interpretation of emerging risks in the financial sector. *The Review of Financial Studies*, *32*, 4543–4603.

Hansen, C., & Kozbur, D. (2014). Instrumental variables estimation with many weak instruments using regularized JIVE. *Journal of Econometrics*, *182*, 290–308.

Hartford, J., Lewis, G., Leyton-Brown, K., & Taddy, M. (2017). Deep IV: A flexible approach for counterfactual prediction. *Proceedings of the 34th International Conference on Machine Learning, Australia*, *70*, 1414–1423.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer Science & Business Media.

Ho, W. K. O., Tang, B.-S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, *38*, 48–70.

Hrazdil, K., Novak, J., Rogo, R., Wiedman, C., & Zhang, R. (2020). Measuring executive personality using machine-learning algorithms: A new approach and audit fee-based validation tests. *Journal of Business Finance & Accounting*, *47*, 519–544.

Hsieh, T.-S., Kim, J.-B., Wang, R. R., & Wang, Z. (2020). Seeing is believing? Executives' facial trustworthiness, auditor tenure, and audit fees. *Journal of Accounting and Economics*, *69*, 101260.

Hu, A., & Ma, S. (2021). *Persuading investors: A video-based study* (NBER Working Paper No. 29048).

Huang, A. H., Zang, A. Y., & Zheng, R. (2014). Evidence on the information content of text in analyst reports. *The Accounting Review*, *89*, 2151–2180.

Hutchinson, J. M., Lo, A. W., & Poggio, T. (1994). A nonparametric approach to pricing and hedging derivative securities via learning networks. *The Journal of Finance*, *49*, 851–889.

Jacobsen, B., Jiang, F., & Zhang, H. (2019). *Equity premium prediction with bagged machine learning* (SSRN Working Paper No. 3310289).

Jones, S., Johnstone, D., & Wilson, R. (2015). An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes. *Journal of Banking & Finance*, *56*, 72–85.

Kamiya, S., Kim, Y. H. A., & Park, S. (2019). The face of risk: CEO facial masculinity and firm risk. *European Financial Management*, *25*, 239–270.

Ke, Z., Kelly, B. T., & Xiu, D. (2019). *Predicting returns with text data* (NBER Working Paper No. 26186).

Kelly, B. T., Pruitt, S., & Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, *134*, 501–524.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*, 2767–2787.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the age of algorithms. *Journal of Legal Analysis*, *10*, 113–174.

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., & Smith, N. A. (2009). Predicting risk from financial reports with regression. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics, USA* (pp. 272–280).

Kozak, S., Nagel, S., & Santosh, S. (2020). Shrinking the cross-section. *Journal of Financial Economics*, *135*, 271–292.

Lahmiri, S., & Bekiros, S. (2019). Can machine learning approaches predict corporate bankruptcy? Evidence from a qualitative experimental design. *Quantitative Finance*, *19*, 1569–1577.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*, 337–346.

Li, B., & Rossi, A. G. (2020). S*electing mutual funds from the stocks they hold: A machine learning approach* (SSRN Working Paper No. 3737667).

Li, K., Liu, X., Mai, F., & Zhang, T. (2021a). The role of corporate culture in bad times: Evidence from the COVID-19 pandemic. *Journal of Financial and Quantitative Analysis*, *56*, 2545–2583.

Li, K., Mai, F., Shen, R., & Yan, X. (2021b). Measuring corporate culture using machine learning. *The Review of Financial Studies*, *34*, 3265–3315.

Liew, J. K.-S., & Wang, G. Z. (2016). Twitter sentiment and IPO performance: A cross-sectional examination. *The Journal of Portfolio Management*, *42*, 129–135.

Lima, A. Q., & Keegan, B. (2020). Chapter 3—Challenges of using machine learning algorithms for cybersecurity: A study of threat-classification models applied to social media communication data. In V. Benson, & J. Mcalaney (Eds.), *Cyber Influence and Cognitive Threats* (pp. 33–52). Academic Press.

Loh, W.-Y. (2011). Classification and regression trees. *WIREs Data Mining and Knowledge Discovery*, *1*, 14–23.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, *66*, 35–65.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, *54*, 1187–1230.

Ludwig, J., Mullainathan, S., & Spiess, J. (2019). Augmenting pre-analysis plans with machine learning. *AEA Papers and Proceedings*, *109*, 71–76.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, USA (pp. 281–297).

Manela, A., & Moreira, A. (2017). News implied volatility and disaster concerns. *Journal of Financial Economics*, *123*, 137–162.

Martin, I. W. R., & Nagel, S. (2022). Market efficiency in the age of big data. *Journal of Financial Economics*, *145*, 154–177.

Mazrekaj, D., Titl, V., & Schiltz, F. (2021). *Identifying politically connected firms: A machine learning approach* (SSRN Working Paper No. 3860029).

McInish, T. H., Nikolsko-Rzhevska, O., Nikolsko-Rzhevskyy, A., & Panovska, I. (2019). Fast and slow cancellations and trader behavior. *Financial Management*, *49*, 973–996.

Medsker, L. R., & Jain, L. C. (2001). Recurrent neural networks. *Design and Applications*, *5*, 64–67.

Milunovich, G. (2020). Forecasting Australia's real house price index: A comparison of time series and machine learning methods. *Journal of Forecasting*, *39*, 1098–1118.

Moritz, B., & Zimmermann, T. (2016). *Tree-based conditional portfolio sorts: The relation between past and future stock returns* (SSRN Working Paper No. 2740751).

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

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the Unied States of America*, *116*, 22071–22080.

Nagel, S. (2021). *Machine learning in asset pricing*. Princeton University Press.

Nauhaus, S., Luger, J., & Raisch, S. (2021). Strategic decision making in the digital age. *Journal of Management Studies*, *58*, 1933–1961.

Nazemi, A., & Fabozzi, F. J. (2018). Macroeconomic variable selection for creditor recovery rates. *Journal of Banking & Finance*, *89*, 14–25.

O'Malley, T. (2021). The impact of repossession risk on mortgage default. *The Journal of Finance*, *76*, 623–650.

Obaid, K., & Pukthuanthong, K. (2022). A picture is worth a thousand words: Measuring investor sentiment by combining machine learning and photos from news. *Journal of Financial Economics*, *144*, 273–297.

Oh, K. J., Kim, T. Y., & Kim, C. (2006). An early warning system for detection of financial crisis using financial market volatility. *Expert Systems*, *23*, 83–98.

Osterrieder, J., Kucharczyk, D., Rudolf, S., & Wittwer, D. (2020). Neural networks and arbitrage in the VIX. *Digital Finance*, *2*, 97–115.

Ozbayoglu, A. M., Gudelek, M. U., & Sezer, O. B. (2020). Deep learning for financial applications: A survey. *Applied Soft Computing*, *93*, 106384.

Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, *42*, 2928–2934.

Peng, L., Teoh, S. H., Wang, Y., & Yan, J. (2022). Face value: Trait impressions, performance characteristics, and market outcomes for financial analysts. *Journal of Accounting Research*, *60*, 653–705.

Pérez-Rave, J. I., Correa-Morales, J. C., & González-Echavarría, F. (2019). A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. *Journal of Property Research*, *36*, 59–96.

Peysakhovich, A., & Naecker, J. (2017). Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. *Journal of Economic Behavior & Organization*, *133*, 373–384.

Philippon, T. (2019). *On fintech and financial inclusion* (NBER Working Paper No. 26330).

De Prado, M. L. (2018). *Advances in financial machine learning*. John Wiley & Sons.

Rambachan, A., Kleinberg, J., Ludwig, J., & Mullainathan, S. (2020). An economic perspective on algorithmic fairness. *AEA Papers and Proceedings*, *110*, 91–95.

Rambachan, A., Kleinberg, J., Mullainathan, S., & Ludwig, J. (2020). *An economic approach to regulating algorithms* (NBER Working Paper No. 27111).

Rasekhschaffe, K. C., & Jones, R. C. (2019). Machine learning for stock selection. *Financial Analysts Journal*, *75*, 70–88.

Rasmussen, C. (1999). The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems*, *12*, 554–560.

Renault, T. (2017). Intraday online investor sentiment and return patterns in the U.S. stock market. *Journal of Banking & Finance*, *84*, 25–40.

Rico-Juan, J. R., & Taltavull de La Paz, P. (2021). Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. *Expert Systems with Applications*, *171*, 114590.

Rish, I. (2001). An empirical study of the Naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. *3*, pp. 41–46).
Rossi, A. G. (2018). *Predicting stock market returns with machine learning* (Working Paper). Retrieved December 7, 2022, from https://mendoza.nd.edu/wp-content/uploads/2019/07/2018-Alberto-Rossi-Fall-Seminar-Paper-1-Stock-Market-Returns.pdf
Rossi, A. G., & Timmermann, A. (2015). Modeling covariance risk in Merton's ICAPM. *The Review of Financial Studies*, *28*, 1428–1461.
Rossi, A. G., & Utkus, S. P. (2020). *Who benefits from robo-advising? Evidence from machine learning* (SSRN Working Paper No. 3552671).
Routledge, B. R. (2019). Machine learning and asset allocation. *Financial Management*, *48*, 1069–1094.
Sadhwani, A., Giesecke, K., & Sirignano, J. (2021). Deep learning for mortgage risk. *Journal of Financial Econometrics*, *19*, 313–368.
Samuelson, P. A., & Nordhaus, W. D. (2009). *Economics* (19th ed.). McGraw Hill/Irwin.
Sermpinis, G., Tsoukas, S., & Zhang, Y. (2022). Modelling failure rates with machine-learning models: Evidence from a panel of UK firms. *European Financial Management* (Forthcoming).
Settles, B. (2009). *Active learning literature survey* (Computer Science Technical Report No. 1648). Retrieved December 7, 2022, from https://minds.wisconsin.edu/bitstream/handle/1793/60660/TR1648.pdf?sequence=1
Sigrist, F., & Hirnschall, C. (2019). Grabit: Gradient tree-boosted Tobit models for default prediction. *Journal of Banking & Finance*, *102*, 177–192.
De Spiegeleer, J., Madan, D. B., Reyners, S., & Schoutens, W. (2018). Machine learning for quantitative finance: fast derivative pricing, hedging and fitting. *Quantitative Finance*, *18*, 1635–1643.
Sprenger, T. O., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management*, *20*, 926–957.
Stock, J. H., & Watson, M. W. (2020). *Introduction to econometrics* (4th ed.). Pearson.
Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.
Tang, V. W. (2018). Wisdom of crowds: Cross-sectional variation in the informativeness of third-party-generated product information on twitter. *Journal of Accounting Research*, *56*, 989–1034.
Tantri, P. (2021). FinTech for the poor: Financial intermediation without discrimination. *Review of Finance*, *25*, 561–593.
Tchuente, D., & Nyawa, S. (2022). Real estate price estimation in French cities using geocoding and machine learning. *Annals of Operations Research*, *308*, 571–608.
Tian, S., Yu, Y., & Guo, H. (2015). Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance*, *52*, 89–100.
Vamossy, D. F. (2021). Investor emotions and earnings announcements. *Journal of Behavioral and Experimental Finance*, *30*, 100474.
Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*, 3–28.
Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, *21*, 1455–1508.
Xiang, G., Zheng, Z., Wen, M., Hong, J., Rose, C., & Liu, C. (2012). A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on TechCrunch. In *Proceedings of the sixth international AAAI conference on weblogs and social media*, Ireland (vol. pp. 607–610).
Yao, J., Li, Y., & Tan, C. L. (2000). Option price forecasting using neural networks. *Omega*, *28*, 455–466.
Yu, Y., Lu, J., Shen, D., & Chen, B. (2021). Research on real estate pricing methods based on data mining and machine learning. *Neural Computing and Applications*, *33*, 3925–3937.
Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: An efficient data clustering method for very large databases. *ACM SIGMOD Record*, *25*, 103–114.
Zhu, X. (2005). *Semi-supervised learning literature survey* (Computer Science Technical Report No. 1530). Retrieved December 7, 2022, from https://minds.wisconsin.edu/bitstream/handle/1793/60444/TR1530.pdf?sequence=1
Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*, 301–320.

SUPPORTING INFORMATION

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

How to cite this article: Hoang, D., & Wiegratz, K. (2022). Machine learning methods in finance: Recent applications and prospects. *European Financial Management*, 1–45. https://doi.org/10.1111/eufm.12408

APPENDIX

TABLE A1 Selection of public announcements of large financial institutions using ML in their day-to-day business operations

This table reports a selection of newswires and press releases from Nexis Uni that contain public announcements of large financial institutions using ML in their day-to-day business operations.

Company	Release date	Source	Extract
Axa	21 Jul 2021	MarketLine NewsWire	*'AXA UK has launched a new machine learning tool to accelerate as well as improve the accuracy of complex property claims'*
Bank of America	13 Jan 2022	PR Newswire	*'Bank of America today announced the launch of CashPro Forecasting, a tool that uses artificial intelligence (AI) and machine learning (ML) technology to more accurately predict future cash positions across clients' accounts'*
Blackrock	11 Apr 2016	ENP Newswire	*'BlackRock investment teams [...] utilise technology-based tools and research methodologies such as machine learning, natural language processing, scientific data visualisation and distributed computing to produce sustainable alpha.'*
Deutsche Bank	23 Sep 2022	MarketLine NewsWire	*'The solution leverages artificial intelligence and specified rules to calculate the risk value for each transaction. [...] Our worldwide network and the use of machine learning techniques allow us to deploy a global data set to reduce fraud.'*
HSBC	6 Nov 2019	Malaysia Economic News	*'HSBC has been able to deal promptly with any anomalous or suspicious transaction through the adoption of new technologies namely Artificial Intelligence (AI) and machine learning.'*

TABLE A1 (Continued)

Company	Release date	Source	Extract
J.P. Morgan Asset Management	17 Dec 2021	PR Newswire	*'J.P. Morgan Asset Management has recently launched its first mutual fund employing a data science-driven investment process [...]. The investment process is driven by machine learning [...]'*
State Street	18 Jul 2018	Business Wire	*'State Street Corporation (NYSE: STT) today announced the launch of State Street VerusSM, a mobile-first application that makes connections between news coverage and investors' holdings through the application of big data, machine learning, natural language processing and human intelligence. Verus is designed to help investment professionals in the front office gain greater insights, mitigate risk, and generate alpha'.*
State Street	22 Jun 2021	Business Wire	*'State Street Corporation today announced it will implement a cloud-based, machine learning technology to transform private markets processing and document management'.*