



Machine learning techniques for cross-sectional equity returns' prediction

Christian Fieberg^{1,2,3} · Daniel Metko^{1,3} · Thorsten Poddig⁴ · Thomas Loy⁵

Received: 24 August 2021 / Accepted: 12 September 2022 / Published online: 28 September 2022
© The Author(s) 2022

Abstract

We compare the performance of the linear regression model, which is the current standard in science and practice for cross-sectional stock return forecasting, with that of machine learning methods, i.e., penalized linear models, support vector regression, random forests, gradient boosted trees and neural networks. Our analysis is based on monthly data on nearly 12,000 individual stocks from 16 European economies over almost 30 years from 1990 to 2019. We find that the prediction of stock returns can be decisively improved through machine learning methods. The outperformance of individual (combined) machine learning models over the benchmark model is approximately 0.6% (0.7%) per month for the full cross-section of stocks. Furthermore, we find no model breakdowns, which suggests that investors do not incur additional risk from using machine learning methods compared to the traditional benchmark approach. Additionally, the superior performance of machine learning models is not due to substantially higher portfolio turnover. Further analyses suggest that machine learning models generate their added value particularly in bear markets when the average investor tends to lose money. Our results indicate that future research and practice should make more intensive use of machine learning techniques with respect to stock return prediction.

✉ Christian Fieberg
cfieberg@uni-bremen.de

Daniel Metko
metko@uni-bremen.de

Thorsten Poddig
poddig@uni-bremen.de

Thomas Loy
thomas.loy@uni-bremen.de

- ¹ Empirical Capital Market Research and Derivatives, University of Bremen, Bremen, Germany
- ² University of Luxembourg, Luxembourg, Luxembourg
- ³ Concordia University, Montreal, Canada
- ⁴ Chair of Finance, University of Bremen, Bremen, Germany
- ⁵ Management Accounting and Information Systems, University of Bremen, Bremen, Germany

Keywords Machine learning · Finance · Stock return prediction

1 Introduction

Recently, Gu et al. (2020) conducted a much noted comparative analysis of machine learning methods for stock return prediction by synthesizing the body of empirical asset pricing literature within the field of machine learning. They provide evidence for a clear rejection of the ordinary least squares (OLS) benchmark in favor of machine learning methods in terms of statistical performance and investors' economic performance. The OLS benchmark represents the typical approach in one of two basic strands of the empirical literature on stock return prediction. Specifically, cross-sectional stock return predictability research (e.g., Fama and French 2008; Lewellen 2015; Gu et al. 2020) runs cross-sectional regressions of future stock returns on a handful of lagged stock characteristics. The second strand of literature, i.e., time-series stock return predictability research, does not forecast the cross-section but the time-series of returns. This literature typically tries to forecast stock indices employing macroeconomic predictors.¹ Attributable to the underperformance of the linear benchmark in their study, Gu et al. (2020) recommend using machine learning techniques to overcome the severe limitations of commonly applied methods.

The paper from Gu et al. (2020) is the first to comprehensively use ML methods in cross-sectional predictability research. Following this seminal paper, numerous concurrent papers have emerged that contribute to the literature by investigating the generality of the conclusions derived in Gu et al. (2020). This is achieved by applying their research design to other stock markets (see, e.g., Tobek and Hronec 2021; Drobetz and Otto 2021; Leippold et al. 2022; Liu et al. 2022; Rubesam 2022; Lalwani and Meshram 2022, for international applications and applications to individual regions or countries) or other asset classes (see, e.g., Bianchi et al. 2021, for applications to bond returns). Another strand of the literature (see, e.g., Leippold et al. 2022, for a list of exemplary papers) is dedicated to numerous additional refinements of the basic algorithms surveyed in Gu et al. (2020).² Our empirical analysis contributes to the line of research featured in the aforementioned papers by deviating from their research design in several key aspects to further investigate the generality of the conclusion to reject the OLS benchmark in favor of machine learning methods in cross-sectional predictability research. More specifically, we devote our attention to three distinct issues: (1) overfitting and irrelevant predictors, (2) missing data, and (3) a U.S. bias. In the following, we discuss these issues and how they can be overcome. First, the aforementioned literature typically relies on large predictor sets. For example, the predictor set in Gu et al. (2020) includes 94 firm characteristics and interactions of each firm characteristic with eight macroeconomic time-series from

¹ See, among others, Rapach and Zhou (2013) for an overview.

² The aforementioned papers represent just a selection of published papers. Many more working papers also build on the seminal paper by Gu et al. (2020).

Welch and Goyal (2008) and 74 industry sector dummy variables. This results in a set of 900+ predictors. Following (Gu et al. 2020; Leippold et al. 2022) even extend this predictor set (i.e., 1160 predictors) while relying on a smaller cross-section (i.e., 3,900 stocks) and time-series (i.e., a study period from 2000 to 2020). Tobek and Hronec (2021) and Bianchi et al. (2021) also use more than 100 predictors when investigating stocks internationally and bonds, respectively. Such large predictor sets and the enhanced flexibility of machine learning methods over more traditional prediction techniques such as the OLS benchmark come at the risk of overfitting the data, which may put machine learning methods at a disadvantage. However, one can also argue that some machine learning methods can handle irrelevant predictors while OLS cannot, which may put (some) machine learning methods at an advantage. For a comparison at eye level, we provide the OLS benchmark and machine learning methods with the same and a relatively sparse set of only relevant predictor variables to prevent overfitting and avoid irrelevant predictors.³ Our set of predictors consists of beta (Sharpe 1964), market capitalization (Banz 1981), the book-to-market-equity ratio (Rosenberg et al. 1985), momentum (Jegadeesh and Titman 1993), investment (Titman et al. 2004) and operating profitability (Novy-Marx 2013), all of which form the basis for well-known factor models such as the Fama and French (1993) three-factor model, the Carhart (1997) four-factor model and the Fama and French (2015) five-factor model. Second, the treatment of missing data is another potential problem affecting the findings derived in the previous literature. More specifically, if a characteristic is missing, then it is typically replaced by the cross-sectional mean or median, which is zero as stock characteristics are rank-transformed and mapped into a $[-1, 1]$ interval.⁴ In this vein, Cismondi et al. (2013) argue that in cases where missing data can range up to 50% or more, imputing the data is incorrect, as it might create unrealistic states of the process being modeled. Afifi and Elashoff (1966) even argue that imputing the mean yields unbiased estimates if and only if the data follow a multivariate normal distribution and the data are missing at random. Given that this is likely not the case for financial market and accounting data, imputing missing data has received much attention in recent research by Freyberger et al. (2021); Cahan et al. (2022); Bryzgalova et al. (2022); Beckmeyer and Wiedemann (2022). To reduce the unintended impact of missing data, we select the aforementioned predictors so that they are available over the entire period and there

³ As the OLS benchmark results reported in Gu et al. (2020) are based on just three predictors, the astonishing outperformance of the machine learning methods might also be an artifact of the large set of predictors applied in the latter. In unreported results, we find that the OLS benchmark achieves a performance similar to the best-performing machine learning method in Gu et al. (2020) (i.e., a neural network with three hidden layers) when applied to a comparable set of predictors on the original data. We thank the authors for providing their data.

⁴ For example, the analysis in Gu et al. (2020) begins in 1957, and with regard to the 94 firm characteristics from the 900+ predictors, information for the full set of predictors is available from 1985 onward. As an example, cash flow statements became mandatory in the U.S. only in the 1980s, as FAS95, after considerable discussion, was finally issued in 1987. While earlier adoption was suggested, it certainly was not the norm (Livnat and Zarowin 1990). Furthermore, even in 1985, 10% of the 94 firm characteristics have more than 30% missing data points. This means that the findings in Gu et al. (2020) might critically depend on replacing missing data with zeros.

are no large deviations in missing data between the predictors in the cross-section. Finally, another potential issue is related to the U.S. bias (and, therefore, lack of external validity) in research in economics (e.g., Das et al. 2013) and finance (e.g., Karolyi 2016).⁵ Specifically, Karolyi (2016) finds that only 16% (23%) of all empirical studies published in the top four (fourteen) finance journals examine non-U.S. markets, a fraction that is well below the measures reflecting their economic importance. This is problematic in two respects. First, generalizing conclusions solely from U.S. data can be dangerous, as research has shown that such conclusions do not necessarily hold internationally (see, e.g., Goyal and Wahal 2015; Woodhouse et al. 2017; Jacobs and Müller 2020). Therefore, every replication makes a contribution when extending existing studies out of sample (see, e.g., Harvey 2017; Hou et al. 2018). Second, a disproportionately high use of U.S. data poses the danger of widespread p-hacking, as argued by Harvey et al. (2016); Harvey (2017); Hou et al. (2018). Harvey et al. (2016) outline three ways to deal with the bias introduced by multiple testing: (1) using out-of-sample validation, (2) using a statistical framework that allows for multiple testing⁶, and (3) looking across multiple asset classes. In this paper, we choose the first approach and conduct the empirical analysis based on all European countries that are part of the MSCI Developed Europe Index. More specifically, we investigate nearly 12,000 individual stocks from 16 countries over almost 30 years from 1990 to 2019. To summarize, we contribute to the cross-sectional predictability research (e.g., Fama and French 2008; Lewellen 2015; Gu et al. 2020) by investigating whether recently derived conclusions (see, e.g., Gu et al. 2020; Tobek and Hronec 2021; Drobetz and Otto 2021; Leippold et al. 2022; Liu et al. 2022; Rubesam 2022; Lalwani and Meshram 2022; Bianchi et al. 2021) hold when we minimize or eliminate the influence of overfitting, irrelevant predictors, missing values, and U.S. bias.

We find confirmation for the results from the previous literature in the sense that the use of machine learning methods appears promising. This is also true if we exclude influences that could potentially distort the results, such as overfitting, irrelevant predictors, missing values, and U.S. bias, and compare the methods on an equal footing. Specifically, the outperformance of the best-performing individual machine learning method is approximately 0.6% per month based on the entire cross-section. This figure is impressive, given that only six predictor variables are considered. The outperformance shrinks to 0.1% per month when only the stocks of the largest ten percent of firms are considered, revealing that the outperformance of the machine learning methods is higher among stocks that are more difficult and costlier to trade. Given that the information environment on these stocks is arguably worse (e.g., in terms of media coverage or analyst following), our results are of great importance to investors seeking guidance. Additionally, we find that the superior performance of machine learning models is not due to substantially higher portfolio

⁵ Similar lines of reasoning can be found in recent research by, e.g., Tobek and Hronec (2021), Lalwani and Meshram (2022) and Hanauer (2020).

⁶ Harvey et al. (2016) and Harvey (2017) raise the claim for a higher statistical threshold (t statistic of 3 instead of 2).

turnover. The analysis of bull and bear markets suggests that machine learning models generate their added value particularly in bear markets when the average investor tends to lose money. Lastly, we find that forecast combinations provide the most robust forecasts, on average. More specifically, forecast combinations consisting of only nonlinear methods consistently outperform forecast combinations that also consider linear models. Thus, we find that investors and researchers alike should prioritize machine learning techniques over the commonly applied benchmark approach when engaging in stock return predictions.

The remainder of the paper is organized as follows. Section 2 discusses the differences between time-series and cross-sectional stock return predictability. Section 3 formulates the OLS benchmark and the machine learning methods applied. Section 4 describes the data and provides descriptive statistics on the predictor variables and the dependent variable (stock returns) from the approximately 12,000 European firms included in this study. Sections 5 and 6 report the main and additional results of the performance comparisons based on statistical (model R^2) and economic (economic gains to investors from portfolio strategies) analyses. Section 7 concludes the paper.

2 Related literature

In finance research, and even more so in non-finance research such as operations research, the time-series approach dominates the cross-sectional approach when it comes to forecasting asset returns. Against this background, in this section, we aim to shed light on the differences between the two approaches discussed in recent literature:

- Engelberg et al. (2022): “Financial researchers have examined the predictability of stock returns for over a century (e.g., Gibson 1906) and a large literature has documented evidence of predictability in the cross-section of stock returns. A separate literature has examined the predictability of the equity risk-premium using time-series predictive variables. To date, these two literatures have evolved relatively independently.”
- Dong et al. (2022): “The first examines whether firm characteristics can predict the cross-sectional dispersion in stock returns. These studies identify numerous equity market anomalies (e.g., Fama and French 2015; Harvey et al. 2016; McLean and Pontiff 2016; Hou et al. 2018). The second line of research investigates the time-series predictability of the aggregate market excess return based on a variety of economic and financial variables, such as valuation ratios, interest rates, and inflation (Nelson 1976; Campbell 1987; Fama and French 1988, 1989; Pástor and Stambaugh 2009). Studies in this vein attempt to shed light on the variables that affect the equity risk premium.”
- Gu et al. (2020): “The first strand models differences in expected returns across stocks as a function of stock-level characteristics, and is exemplified by Fama and French (2008) and Lewellen (2015). The typical approach in this literature runs cross-sectional regressions of future stock returns on a few lagged stock charac-

teristics. The second strand forecasts the time-series of returns and is surveyed by Welch and Goyal (2008), Koijen and van Nieuwerburgh (2011), and Rapach and Zhou (2013)."

In summary, prediction of the time-series of a single asset such as the aggregate market returns is related to the time-series predictability research. The typical approach is to forecast a single time-series (e.g., a specific MSCI index) based on other time-delayed time-series (e.g., economic and financial variables, such as valuation ratios, interest rates, and inflation). Predicting the cross-sectional dispersion in stock returns is related to cross-sectional predictability. The main approach is to forecast the returns of multiple assets at once and at one point in time (e.g., returns from all stocks in the US at time t) using other, time-delayed, cross-sectional information (e.g., market equity, momentum, market beta from all stocks in the US at time $t-1$). Following (Rapach and Zhou 2020) and based on a linear regression model, one can also make a more formal distinction between the two approaches:

- The time-series predictability approach considers the following time-series multiple regression model for an individual asset

$$r_t = \alpha + \sum_{j=1}^J \beta_j x_{j,t-1} + \epsilon_t, \quad (1)$$

for $t = 1, \dots, T$, where r is the return on a broad stock market index in excess of the risk-free return, T is the number of time-series observations, $x_{j,t-1}$ are the lagged predictor variables, and ϵ_t is a zero-mean disturbance term.

- The cross-sectional predictability approach considers the following cross-sectional multiple regression model for a certain month

$$r_i = \alpha + \sum_{j=1}^J \beta_j z_{i,j} + \epsilon_i, \quad (2)$$

for $i = 1, \dots, N$, where $z_{i,j}$ is the j th lagged characteristic of firm i and N is the number of available firms.⁷

In this paper, we are interested in predicting the cross-sectional dispersion in stock returns. However, in terms of cross-sectional predictability, Gu et al. (2020) is the first paper to comprehensively use ML methods and compare them against the standard OLS benchmark in this line of research (e.g., Fama and French 2008; Lewellen 2015; Gu et al. 2020). Against this background, we also rely on OLS as our "natural" benchmark.

⁷ To ensure parameter stability, cross-sectional predictability typically relies on multiple cross-sections or pooled data.

3 Methodology

In our empirical study, we are dealing with a panel of stocks, where months are indexed as $t = 1, \dots, T$ and stocks are indexed as $i = 1, \dots, N$. Accordingly, the future stock return r of asset i at month $t + 1$ can be defined in general terms as:⁸

$$r_{i,t+1} = E_t(r_{i,t+1}) + \epsilon_{i,t+1}, \quad (3)$$

with

$$E_t(r_{i,t+1}) = g(x_{i,t}), \quad (4)$$

where current expectations about future returns are expressed as a function $g(\cdot)$ of a vector of predictor variables $x_{i,t}$ (i.e., firm characteristics). Specifically, the predictor variables are defined as a P -dimensional vector $x_{i,t} = (x_{i,t,1}, \dots, x_{i,t,P})$. Hence, the aim of the following methods is to provide an estimate $\hat{g}(\cdot)$ of $g(\cdot)$.

3.1 Simple and penalized linear regression

Throughout the recent decades of cross-sectional asset pricing research, the common strand in the literature on stock return predictability has been to model conditional expectations about future returns $g(\cdot)$ as a linear function of stock-level characteristics and a parameter vector $\beta = (\beta_1, \dots, \beta_P)$ (e.g., Fama and French 2008; Lewellen 2015; Green et al. 2017):

$$\hat{g}(x_{i,t}; \beta) = x'_{i,t} \beta. \quad (5)$$

The parameter vector β can be obtained by (pooled) OLS, minimizing the l_2 objective function:

$$\min_{\beta} \mathcal{L}(\beta) = \frac{1}{N(T-1)} \sum_{i=1}^N \sum_{t=1}^{T-1} (r_{i,t+1} - \hat{g}(x_{i,t}; \beta))^2, \quad (6)$$

where N is the number of stocks in the cross-section and T is the number of months in the estimation period. We refer to this approach as the OLS benchmark. To reduce the impact of overfitting, we also add a penalty term to the objective function:

$$\min_{\beta} \mathcal{L}(\beta; \cdot) = \mathcal{L}(\beta) + \phi(\beta; \cdot). \quad (7)$$

Regularization techniques, such as penalty terms, are most commonly used in the context of large feature spaces. Although the problem of overfitting is only of minor importance for the linear model given the small feature set, it can still occur and affect performance. For the sake of completeness, we follow prior literature and append the linear model by the popular elastic net penalty, which takes the form:

⁸ For ease of notation, we assume a balanced panel.

$$\phi(\beta; \lambda; \rho) = \lambda(1 - \rho) \sum_{p=1}^P |\beta_p| + \frac{1}{2} \lambda \rho \sum_{p=1}^P \beta_p^2. \quad (8)$$

The two nonnegative hyperparameters of the elastic net penalty, λ and ρ , nest some special cases. λ controls the penalty strength; hence, for $\lambda = 0$, the model reduces to the standard OLS. The parameter ρ controls the ratio of lasso versus ridge penalty, where the cases $\rho = 0$ and $\rho = 1$ correspond to lasso and ridge regression, respectively. Lasso imposes sparsity in the coefficients and may lead to some coefficients being exactly zero. Ridge regression is a dense modeling technique that prevents coefficients from becoming unduly large in magnitude by shrinking them closer to zero without imposing a weight of exactly zero. In addition to lasso ($\rho = 0$) and ridge regressions ($\rho = 1$), we consider elastic net ($\rho = 0.5$), which represents a compromise between the two.

In our empirical application, we determine the parameter λ using grid search (see Table 11 in appendix for the ranges) on the validation data.

3.2 Support vector machine

Support vector regression (SVR) is an extension of the basic principle underlying support vector machines developed by Vapnik (1995), originally designed for classification. SVR generally estimates a linear regression function of the following form:

$$\hat{g}(x_{i,t}; w, b) = x'_{i,t} w + b, \quad (9)$$

where w is a vector of weights and b is a constant. The objective of SVR is to estimate a function that provides a good balance between model complexity and goodness of fit by introducing an error margin (ϵ -insensitive tube) to which the cost function is insensitive in approximating the returns. This may be interpreted as the tolerance for minuscule deviations up to ϵ . The ϵ -insensitive loss is defined as (cf. Vapnik 1995):

$$\mathcal{L}(r_{i,t+1}, \hat{g}(x_{i,t})) = \begin{cases} 0 & \text{if } |r_{i,t+1} - \hat{g}(x_{i,t})| \leq \epsilon \\ |r_{i,t+1} - \hat{g}(x_{i,t})| - \epsilon & \text{otherwise.} \end{cases} \quad (10)$$

By introducing two slack variables $\xi_{i,t+1}$ and $\xi_{i,t+1}^*$ for positive and negative deviations from the ϵ -insensitive loss function, respectively, the objective function can be written in primal form as:

$$\begin{aligned} \min_{w, b} \mathcal{L}(w, b) &= C \sum_{i=1}^N \sum_{t=1}^{T-1} (\xi_{i,t+1} + \xi_{i,t+1}^*) + \frac{1}{2} \sum_{p=1}^P w_p^2 \\ \text{s.t. } r_{i,t+1} - (w' x_{i,t} - b) &\leq \epsilon + \xi_{i,t+1}, \\ (w' x_{i,t} + b) - r_{i,t+1} &\leq \epsilon + \xi_{i,t+1}^*, \\ \xi_{i,t+1}, \xi_{i,t+1}^* &\geq 0, \end{aligned} \quad (11)$$

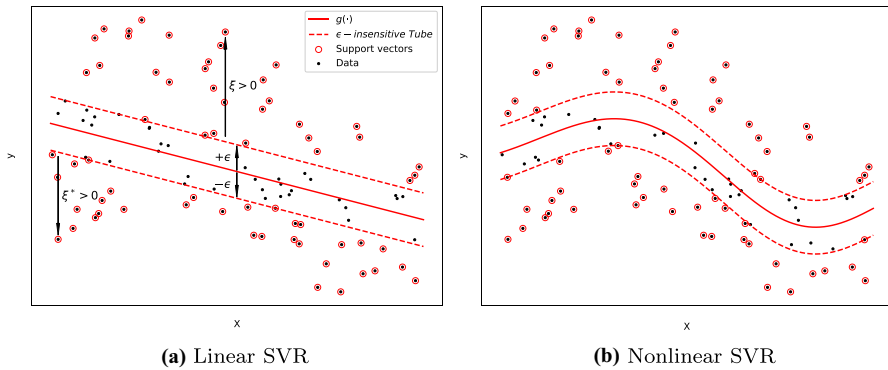


Fig. 1 This figure visualizes an exemplary SVR with a linear and a nonlinear kernel in (a) and (b), respectively. ξ and ξ^* denote errors with magnitude greater than the $|\epsilon|$ threshold

where the optimization problem is defined as the sum of aggregate errors exceeding the ϵ -boundary plus l_2 regularization that encourages a simple model to prevent overfitting of the data. The nonnegative constant C controls how strong deviations from ϵ should be penalized.⁹ Note that C can be interpreted as an inverse regularization parameter, as it controls the tradeoff between error tolerance and complexity. Two notable cases are $C = 0$, for which SVR approximates a constant function, and $C \rightarrow \infty$, $\epsilon = 0$, for which SVR essentially estimates an l_1 -regression.

By means of Lagrangian theory, the above minimization problem can be represented in dual form as:

$$\begin{aligned} \min_{\alpha, \alpha^*} \mathcal{L}(\alpha, \alpha^*) &= \frac{1}{2} \sum_{(i,t) \in SV} \sum_{(j,k) \in SV} (\alpha_{i,t} - \alpha_{i,t}^*) (\alpha_{j,k} - \alpha_{j,k}^*) x'_{i,t} x_{j,k} \\ &\quad - \epsilon \sum_{(i,t) \in SV} (\alpha_{i,t} - \alpha_{i,t}^*) + \sum_{(i,t) \in SV} y_{i,t} (\alpha_{i,t} - \alpha_{i,t}^*) \\ \text{s.t. } \sum_{(i,t) \in SV} (\alpha_{i,t} - \alpha_{i,t}^*) &= 0, \quad 0 \leq \alpha_{i,t}, \alpha_{i,t}^* \leq C, \end{aligned} \quad (12)$$

where $\alpha_{i,t}$ and $\alpha_{i,t}^*$ are nonnegative Lagrangian multipliers for each observation $x_{i,t}$. SV are the indices of nonzero Lagrangian multipliers, where the corresponding $x_{i,t}$ are the support vectors. Accordingly, the first line of equation (12) considers pairwise combinations of panel observations if they are in the set of support vectors. Support vectors are observations that correspond to nonzero Lagrangian multipliers, i.e., either $\alpha_{i,t}$ or $\alpha_{i,t}^*$ is nonzero.

The dual form enables the introduction of nonlinearity by replacing the inner product $x'_{i,t} x_{j,k}$, i.e., a linear mapping, with a kernel function $K(x_{i,t}, x_{j,k})$, which

⁹ By convention, the SVR formulation typically uses an (inverse) regularization parameter. The objective may also be formulated as $\sum_{i=1}^N \sum_{t=1}^{T-1} (\xi_{i,t+1} + \xi_{i,t+1}^*) + \lambda \frac{1}{2} \sum_{p=1}^P w_p^2$. The relation between λ and C is $C = \frac{1}{\lambda}$.

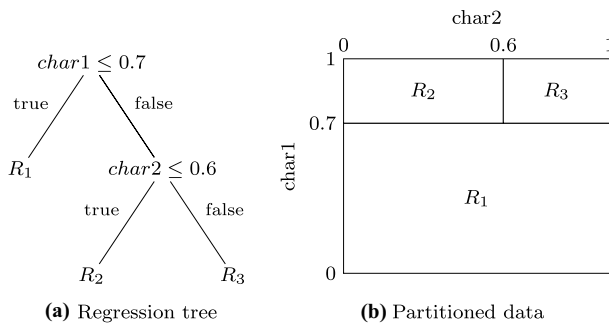


Fig. 2 This figure shows the structure of an example regression tree (a) with two characteristics (*char1* and *char2*) and its equivalent representation as sliced partitions of the predictor space (b)

performs a nonlinear mapping of the inputs. Figure 1 illustrates a SVR with a linear (a) and a nonlinear (b) kernel. Conveniently, the introduction of a kernel function allows all calculations to be performed in the original feature space. We apply the most popular radial basis function (RBF), which takes the form Lu et al. (2009), Rasekhschaffe and Jones (2019):

$$K(x_{i,t}, x_{j,k}) = \exp(-\gamma \|x_{i,t} - x_{j,k}\|^2). \quad (13)$$

This kernel function measures the similarity of firm characteristics between firm-observations and, consequently, how they affect the regression function. Here, γ is a bandwidth parameter that controls the influence of observations. In our empirical application, we determine parameters ϵ , λ and γ using grid search (see Table 11 in appendix for the ranges) on the validation data.

3.3 Tree-based methods: random forest and gradient boosted regression trees

Regression trees are fully nonparametric and based on recursive binary splits, which enables them to account for high-level interactions. With respect to our application, regression trees group observations based on firm characteristics such that observations within a group are similar in terms of future returns. A tree ‘grows’ by performing a series of binary splits based on cutoffs of the firm characteristics at each branching point. Consequently, each split adds an additional layer of depth capable of incorporating more complex interaction effects. Starting with all observations, the tree successively divides the feature space into two rectangular partitions. The firm characteristic and its cutoff value are chosen to provide the best fit in terms of forecasting error. The resulting rectangular regions in the predictor space approximate the unknown function $g(\cdot)$ as the mean of the respective partition. The splitting procedure continues on one or both regions, resulting in increasingly smaller rectangles until a stopping criterion is met, e.g., depth of the tree (J) or no improvement by additional splits.

Figure 2 illustrates a regression tree of depth $J = 2$ based on exemplary firm characteristics $char1$ and $char2$.¹⁰ Figure 2a presents the tree architecture, and Fig. 2b shows the representation in the rectangular feature space. Initially, all observations are divided based on the firm characteristic $char1$ with a cutoff value of 0.7. Firm-month observations with a value below that threshold are assigned to the left branch, while all other observations are assigned to the right branch. The left node is not split up any further, resulting in R_1 , which is also called the “leaf node”. The right node represents another decision node, where the remaining firm-month observations are additionally divided on firm characteristic $char2$ with a threshold value of 0.6, resulting in R_2 and R_3 . Thus, the observations are divided into regions R_1 , R_2 and R_3 , as shown in Fig. 2b. The prediction of a tree is the simple average of the response within a leaf node.

Formally, the unknown function $g(\cdot)$ is approximated by a regression tree as:

$$\hat{g}^{tree}(x_{i,t}; M) = \sum_{m=1}^M c_m I(x_{i,t} \in R_m), \quad (14)$$

where R_m is one of the M leaf nodes, I is an indicator function of the firm characteristics identifying observations that belong to a particular region and c_m is the associated average response. At each decision node, the algorithm aims to find the firm characteristic and its respective value that minimize the forecasting error. Finding a global optimal solution is generally computationally infeasible (cf. Hastie et al. 2017). Therefore, we follow (Breiman et al. 1984) and proceed with a greedy algorithm, in which the optimization is only performed locally at each individual decision node, i.e., solving the following optimization problem:

$$\min_{j,s} \left[\sum_{x_{i,t} \in R_l(j,s)} (r_{i,t+1} - c_l(j,s))^2 + \sum_{x_{i,t} \in R_r(j,s)} (r_{i,t+1} - c_r(j,s))^2 \right], \quad (15)$$

with

$$c_l(j,s) = \text{ave}(r_{i,t+1} | x_{i,t} \in R_l(j,s)) \\ \text{and } c_r(j,s) = \text{ave}(r_{i,t+1} | x_{i,t} \in R_r(j,s)),$$

where c_l and c_r are the average responses of the left and right partitions, which are conditional on variable j and its value s subject to the minimization. Put differently, at each decision node, the split variable j and its associated split value s are to be found with the aim of locally minimizing the forecasting error on each split.

Regression trees are widely adopted in the ML literature because (as visualized in Fig. 2) the sequence of binary decision rules allows a very simple and intuitive interpretation. Additionally, the inherent structure is ideal for representing multiway interaction effects and nonlinearities. However, regression trees are rarely used individually, as they are unstable with respect to changes in the input

¹⁰ Depth, split variable and split value are chosen for illustrative purposes.

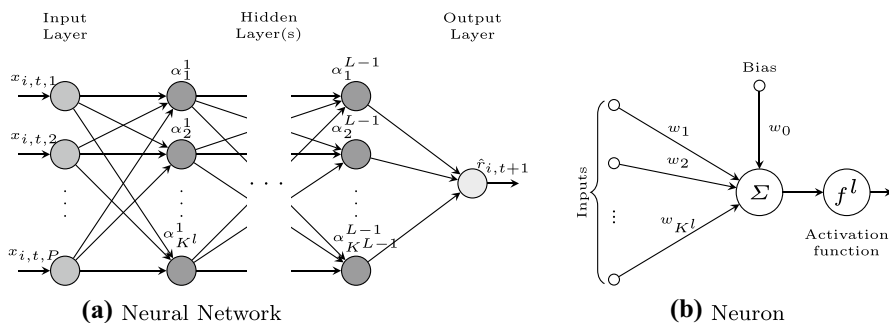


Fig. 3 This figure provides a graphical representation of the general architecture of a feedforward neural network (a) and the information processing within a hidden neuron (b)

data and subject to severe overfitting problems. Unlike OLS, a regression tree can heavily overfit even with few variables, and with sufficient depth, a perfect fit can be achieved. In the limiting case, each leaf node contains only one observation. To alleviate the drawbacks of individual regression trees, we consider popular ensemble algorithms, i.e., random forests and gradient boosting, each of which combines many individually weak trees.

Random forest (RF) is based on the concept of bootstrap aggregation, in which the variance of an estimator is reduced by averaging over multiple independent and identically distributed (i.i.d.) bootstrapped samples. Thus, in the random forest, several B trees are trained independently on bootstrapped samples. In addition, at each decision node of a tree, only a random subset of $p < P$ predictors is considered, reducing both the correlation and variance of the estimator (cf. Breiman 2001). The output of a random forest is provided by:

$$\hat{g}(x_{i,t}; B, M) = \frac{1}{B} \sum_{b=1}^B \hat{g}_b^{tree}(x_{i,t}; M). \quad (16)$$

While trees are fit independently in the random forest algorithm, gradient boosted regression trees (GBRT) are estimated in an adaptive manner to reduce the bias. Therefore, B trees are estimated sequentially and combined additively to form an ensemble prediction. At each iteration b , a new tree \hat{g}_b^{tree} is fit on the residuals from the previous iteration and added to the ensemble. However, the contribution of each individual tree is controlled by the learning rate $0 < \nu < 1$ (shrinkage) to prevent overfitting of the residuals. Finally, the estimation function of GBRT is given by:

$$\hat{g}(x_{i,t}; B, M, \nu) = \sum_{b=1}^B \nu \hat{g}_b^{tree}(x_{i,t}; M). \quad (17)$$

In our empirical application, we determine parameters (J, p) and (J, ν, B) for RF and GBRT, respectively using grid search (see Table 11 in the appendix for the ranges) on the validation data.

3.4 Neural networks

Neural networks are successfully applied in numerous scientific and practical applications. Their flexibility draws from the concept to derive features from linear combinations of the input space and to model the response as a (nonlinear) function of these features (cf. Hastie et al. 2017). We focus on traditional feed-forward neural networks, which generally consist of an input layer, one or more hidden layers, and an output layer. Each layer is represented by (computational) units called neurons. Figure 3a gives a graphical representation of a general feed-forward neural network. Starting with the input layer, which represents the firm characteristics, the information is carried through the hidden layers via interconnected neurons and finally aggregated to form an estimate of the future returns (represented by the output layer). Each of the neurons performs very simple operations individually, i.e., combining information from the previous layer and applying an (nonlinear) activation function to the aggregate signal (see Fig. 3b).

More generally, the output of an L -layer neural network (including the output layer) with $K^l \forall 1 \leq l \leq L$ neurons per layer can be thought of as a chain of functions (cf. Goodfellow et al. 2016), which recursively calculates the activations. Accordingly, the l -th layer's activations α^l , which is a vector of length equal to the number of nodes, are given by:

$$a^l = f^l(\alpha^{(l-1)}) = f^l(W^l \alpha^{(l-1)} + b^l), \quad \forall 1 \leq l \leq L, \quad (18)$$

where W^l is a $K^l \times K^{(l-1)}$ matrix of weights transmitting the outputs from layer $(l-1)$ to layer l , where they then aggregate to form the activation a . Accordingly, $l=0$ corresponds to the input layer with a^0 being the raw vector of firm characteristics, $x_{i,t}$, and f^1, \dots, f^L are univariate transformations of the weighted signal from the previous layer plus a bias. $W = \{W^1, \dots, W^L\}$ and $b = \{b^1, \dots, b^L\}$ are parameter sets of weight matrices and bias vectors, respectively, determined in the model calibration.

In a regression task, the output layer L performs a linear combination, yielding the final output:

$$\hat{g}(x_{i,t}; W, b) = f^L(\alpha^{(L-1)}; W^L, b^L) = W^L \alpha^{(L-1)} + b^L. \quad (19)$$

There are numerous architectural design options when structuring a neural network, e.g., number of neurons, number of layers, and activation functions. Zhang et al. (1998, p. 42) state that “[...] the design of an ANN is more of an art than a science”. However, there is no clear guidance on how to determine an appropriate network architecture. Therefore, we take a pragmatic approach by choosing a single hidden layer and the hyperbolic tangent as an activation function:

$$f^l(\cdot) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \quad \forall 1 \leq l \leq L-1. \quad (20)$$

Hornik et al. (1989) and Cybenko (1989) proved universal approximator properties for single hidden layer feedforward neural networks with sigmoid activations.

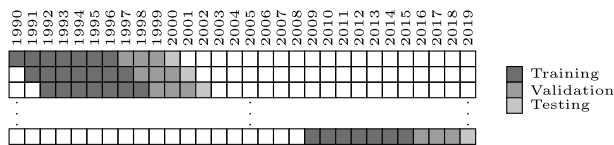


Fig. 4 This figure illustrates the rolling estimation scheme employed in this study

The number of hidden neurons is chosen based on a rule-of-thumb for single hidden layer neural networks derived by Wong (1991), i.e., twice the number of input nodes.

As stated before, in a neural network, W and b are the parameters of interest. Their estimates, however, are solutions of a nonconvex optimization problem. One of the most widely used algorithms for training neural networks are variants of stochastic gradient descent (SGD). SGD iteratively approaches the (local) minimum of the l_2 objective function by approximating its gradient on random subsets of the training data and adapting the parameters accordingly. We use the popular Adaptive Moment estimation (Adam) by Kingma and Ba (2014). A critical tuning parameter in the learning algorithm is the learning rate lr , which controls the step size of each parameter update. First, to mitigate the risk of overfitting, we impose an l_2 penalty (weight decay), which shrinks the weights between neurons toward zero without imposing exact zero. Second, we stop the training of our network early if the error on the validation set does not improve for ten consecutive iterations. Third, we build ensembles of size ten and average their outcome, as the parameter initialization and stochastic training introduce sources of randomness. Hansen and Salamon (1990) provide evidence that ensembles not only stabilize results but also improve generalization. In our empirical application, we determine the learning rate lr and l_2 penalty strength λ using grid search (see Table 11 in the appendix for the ranges) on the validation data.

3.5 Estimation, tuning and testing

We provide all models with the same data, i.e., a ten-year rolling period of monthly data, and predict returns for the months of the subsequent year for every firm. Stated differently, we use only past information to predict future (out-of-sample) returns. Unlike the OLS model, the performance of most machine learning algorithms depends heavily on the choice of hyperparameters. A commonly used strategy is to divide the in-sample data into disjoint subsets (training and validation) to mimic a pseudo out-of-sample setting. The hyperparameters are then chosen to maximize the performance on the validation set. Therefore, we split each in-sample window into seven years of training and three years of validation and perform a grid search to find an appropriate set of hyperparameters.¹¹

¹¹ By retaining the temporal ordering of the data, we mimic a realistic prediction scenario.

Figure 4 illustrates our sample splitting scheme. More precisely, we estimate the respective model with a set of hyperparameters from the parameter grid (see Table 11) on the training set. Then, we generate predictions for the validation set and evaluate the performance. After determining the best set of hyperparameters, we re-estimate the model on the full ten-year period of training and validation data to incorporate the most recent information. However, data for actual forecasting never enter the model during the estimation procedure.

To assess the predictive performance of our models, we evaluate them in three ways. We first deploy a statistical analysis to quantify the models' predictive accuracy, i.e., how well predicted returns reflect realized returns. A common metric to do so is the out-of-sample (pseudo) R^2 , which is defined as:

$$R^2 = 1 - \frac{\sum_{(i,t) \in OOS} (r_{i,t} - \hat{r}_{i,t})^2}{\sum_{(i,t) \in OOS} (r_{i,t} - \bar{r}_{i,t})^2}, \quad (21)$$

where $\hat{r}_{i,t}$ denotes the model's prediction and $\bar{r}_{i,t}$ denotes the historical average for asset i at time t . This definition measures the proportional reduction in forecasting errors of the predictive model relative to the historical average. Thus, $R^2 > 0$ indicates that the predictive model is more accurate than the historical average in terms of forecasting errors. Note that unlike the conventional R^2 , this formulation can also be negative. While the historical average is appropriate in terms of a time-series analysis, it leads to an inflated R^2 in the case of a panel of stock returns. Therefore, we set \bar{r} to zero, which yields a stricter threshold for proper forecasting models, and we follow this suggestion.¹²

R^2 represents a measure that provides information about the quality of forecasts from a predictive model compared to that from a constant model. However, if we want to compare among different models, then examining R^2 alone does not allow us to infer whether one model is significantly superior to another. Therefore, we perform the popular (Diebold and Mariano 1995) (DM) test, which is designed to compare the predictive accuracy of two models. More precisely, it tests for the null of equal prediction errors $H_0 : E[d_t] = 0$, where d_t is the loss differential between two forecasts at time t . The DM test is applicable in the case of serial correlation. However, stock returns are prone to be strongly correlated in the cross-section while only being weakly correlated across time. To alleviate this issue, we average loss differential d_t across assets, thus resulting in a time-series with weak serial correlation. The test statistic is defined as follows:

¹² We also consider other definitions of R^2 ; however, the relative order between the models remains unchanged. To remain consistent with the literature, we adopt this definition.

$$\begin{aligned}
 DM &= \frac{\bar{d}_{1,2}}{\hat{\sigma}_{\bar{d}}} \sim \mathcal{N}(0, 1), \quad \text{with} \\
 \bar{d}_{1,2} &= \frac{1}{T} \sum_{t=1}^T \hat{d}_{1,2,t}, \\
 \hat{d}_{1,2,t} &= \frac{1}{N} \sum_{i=1}^N (\hat{e}_{1,i,t}^2 - \hat{e}_{2,i,t}^2), \quad \text{and} \\
 \hat{\sigma}_{\bar{d}} &= \sqrt{\frac{\frac{1}{T} \sum_{t=1}^T (\hat{d}_t - \bar{d})^2}{T}}.
 \end{aligned} \tag{22}$$

Second, we analyze our forecasts in terms of economic gains. Leitch and Tanner (1991) argue that traditional statistical metrics are only loosely related to a forecast's achievable profit. Thus, we consider forecast-based portfolio sorts to assess the economic performance of our models. At the end of each month in the testing period (January 2001 to December 2019), we assign all available stocks with equal weight to ten portfolios using cross-sectional decile breakpoints based on the return prediction for the next month. In addition, we construct a long-short portfolio reflecting a zero-net investment by buying the top decile portfolio and selling the bottom decile portfolio to capture the aggregate effect.

Third, we use the entire cross-section of stocks to estimate our models. Stocks of small firms, however, follow different dynamics in terms of their covariances, liquidity, and expected returns (cf. Kelly et al. 2019). More specifically, Hou et al. (2018) show that much of the predictability detected in the previous literature is found in the smallest stocks. While microcaps represent only approximately 3% of the total market capitalization of the NYSE/Amex/NASDAQ universe, they account for approximately 60% of the number of stocks. Due to high costs in trading these stocks (e.g., Israel and Moskowitz 2013), predictability (in terms of economic performance) in small stocks is more apparent than it is real. To investigate the degree to which the performance depends on firm size, we successively remove stocks below the respective deciles of the cross-sectional market capitalization.

4 Data

Previous studies have documented the existence of a U.S. bias in research in economics (e.g., Das et al. 2013) and finance (e.g., Karolyi 2016). In this regard, recent literature (e.g., Harvey et al. 2016) argues for a higher statistical threshold (t-statistic of 3 rather than 2) in empirical studies on the U.S. stock market

Table 1 Descriptive statistics and characteristics of quantile portfolios formed from monthly sorts of stocks on predictor variables

	Panel: A			Panel: B						
	Avg	Std	N	1	2	3	4	5	5-1	pVal (5-1)
ret	0.0088	0.0886	1755							
mom	0.1324	0.3685	1755	0.27	0.69	0.90	1.18	1.63	1.37	0.00
beta	0.5609	0.4154	1755	1.02	1.08	1.06	0.87	0.64	− 0.37	18.24
size	5.8364	1.9083	1755	0.91	0.93	0.96	0.96	0.92	0.01	95.61
be2me	0.7408	0.6483	1755	0.68	0.85	0.88	1.04	1.24	0.56	0.16
op	0.0002	0.0003	1755	0.50	0.94	1.05	1.10	1.09	0.59	0.00
inv	18.7620	47.0595	1755	1.02	1.08	1.07	0.96	0.56	− 0.46	0.07

Panel A reports time-series averages of the cross-sectional mean (Avg), standard deviation (Std), and sample size (N) for all stocks from the countries in the MSCI Developed Markets Europe on Datastream and Worldscope from January 1990 to December 2019. The sample includes all stocks with valid monthly values for return (ret), the characteristics from the Fama and French (1993) three-factor model, the Carhart (1997) four-factor model and the Fama and French (2015) five-factor model: beta (beta), log market capitalization (size, in millions), book-to-market equity (be2me), momentum (mom) operating profitability (op), and investment (inv). Panel B reports returns of quintile portfolios formed from monthly sorts of stocks on predictor variables. The five portfolios are defined as follows: For each month from 1990 to 2019, we assign all stocks from the countries in the MSCI Developed Markets Europe to five portfolios using quintile breakpoints from the cross-section of the end-of-month predictor variable values and calculate the equal-weighted monthly percent returns for the next month. To analyze the aggregate effect of a predictor variable on stock returns, we take a long position in portfolio 5 and a short position in portfolio 1. The numbers in columns 1 to 5 report the average of the equally weighted returns. pVal(5-1) reports the p-value (in percent) from a t-test of the null hypothesis of zero average

as well as a stronger focus on other asset classes and stock markets. We follow this demand by investigating the European stock market. More specifically, we use all stocks (approximately 12,000) from the countries in the MSCI Developed Markets Europe Index¹³. We retrieve monthly data from January 1990 to December 2019 from Datastream and Worldscope. The data retrieval starts with the identification of common equity stocks using Datastream's constituent lists (research lists, Worldscope lists and dead lists). Specifically, for every country in the MSCI Developed Europe Index, we use its constituent lists and eliminate any duplicates. As a result, we obtain one remaining list for every country. To each of these lists, we apply generic as well as country-specific screens to eliminate noncommon equity stocks. Moreover, to all stocks remaining from this screening procedure, we then apply dynamic screens to account for data errors. The procedure described above is established in the academic literature and described extensively (e.g., in Ince and Porter 2006; Campbell et al. 2010; Griffin et al. 2010, 2011; Karolyi et al. 2012; Annaert et al. 2013; Fong et al. 2017).

Our predictors are beta, market capitalization, the book-to-market equity ratio, momentum, investment, and operating profitability. These predictors are calculated

¹³ These are Austria, Belgium, Denmark, Finland, France, Germany, Ireland, Italy, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and the United Kingdom.

from the Datastream and Worldscope variables (datatype in brackets) total return index (RI), market equity (MV), book equity (WC03501), total asset growth (WC08621) and operating income (WC01250). We assume that accounting data are available to the public six months after the fiscal year ends to avoid look-ahead bias. Beta (*beta*) is the exposure of a stock to the market factor as derived from the capital asset pricing model (CAPM). According to the CAPM, stock returns are expected to increase as the stock's beta increases. For each month, we estimate the beta of a stock based on the daily stock returns from the previous 12 months to avoid any look-ahead bias. Following the established literature, returns are expected to exhibit an inverse relationship with firms' market capitalization (*size*) (Banz 1981) and a positive association with the book-to-market-equity ratio (*be2me*) (Rosenberg et al. 1985). Momentum (*mom*) is defined as the total return over the prior 12 months, excluding the last month. Prior literature finds evidence of a positive relation between momentum and stock returns (Jegadeesh and Titman 1993; Carhart 1997). Investment (*inv*) is growth in total assets, and operating profitability (*op*) is operating income scaled by the book value of equity. Following (Titman et al. 2004) and (Novy-Marx 2013), stock returns are expected to increase as investment decreases and operating profitability increases, respectively. In addition to exposure to the market factor (beta), market capitalization and the book-to-market-equity ratio are proxies for sensitivities to common unobservable risk factors in the Fama and French (1993) three-factor model. Furthermore, momentum is a surrogate for the sensitivity to a common unobservable risk factor in the Carhart (1997) four-factor model. Lastly, investment and operating profitability proxy for sensitivities to common unobservable risk factors in the Fama and French (2015) five-factor model.

We report time-series averages of cross-sectional statistics, specifically, the mean, standard deviation, and number of observations, of the monthly predictor variables and stock returns in Panel A of Table 1. For comparability, we include only firm-month observations with nonmissing returns in our analysis, resulting in an average sample size of 1,755 firms per month. We find that the average monthly stock return is 0.88% and the average cross-sectional standard deviation in monthly stock returns is 8.86%. To investigate the predictive ability of an individual variable, we further assign all stocks to five portfolios for each month from 1990 to 2019 using quintile breakpoints from the cross-section of monthly predictor variable values and calculate the equal-weighted monthly percent returns for the next month. Panel B of Table 1 reports the average monthly returns of the five portfolios. To analyze the aggregate effect of the predictor variable on stock returns, we take a long position in portfolio 5 and a short position in portfolio 1. Column "pVal (5-1)" reports the p-value (as a percentage) from a t-test against the null hypothesis of zero average return of the "5-1" portfolio. As expected, we find a positive relation between stock returns and momentum, the book-to-market-equity ratio and operating profitability and a negative relation between stock returns and investment for the European stock market. Contradicting the CAPM, we find a negative relation between stock returns and beta, which is consistent with empirical evidence provided by previous studies (e.g., Frazzini and Pedersen 2014). Similarly, we find no relationship between firm size and stock returns, which is also consistent with recent empirical findings (e.g., van Dijk 2011).

Table 2 This table reports monthly predictive panel R^2 as a percent

Size > q	OLS	Lasso	Ridge	Enet	SVR	RF	GBRT	NN
0.00	0.43	0.40	0.42	0.37	0.14	0.58	0.60	0.67
0.10	0.36	0.33	0.34	0.28	0.15	0.36	0.44	0.63
0.20	0.36	0.32	0.34	0.27	0.19	0.36	0.45	0.66
0.30	0.36	0.30	0.34	0.27	0.20	0.35	0.46	0.68
0.40	0.35	0.29	0.33	0.26	0.20	0.34	0.47	0.70
0.50	0.31	0.24	0.30	0.23	0.17	0.27	0.46	0.70
0.60	0.26	0.18	0.25	0.19	0.12	0.20	0.45	0.67
0.70	0.17	0.08	0.17	0.11	-0.01	0.09	0.43	0.63
0.80	- 0.03	- 0.10	- 0.01	- 0.06	- 0.26	- 0.17	0.32	0.50
0.90	- 0.32	- 0.39	- 0.29	- 0.35	- 0.58	- 0.45	0.14	0.24

Monthly predictions are formed for our testing sample from January 2000 to December 2019 for different models and levels of market liquidity in terms of market capitalization. Columns correspond to the models. Rows correspond to different levels of market liquidity, which we establish by removing stocks with market capitalization (Size) less than the cross-sectional quantile q

In our empirical study, we further account for missing values, different scales and extreme observations. To address missing values and maintain the number of return observations, we replace missing characteristics with their cross-sectional median if the respective returns are available. However, in contrast to much of the current cross-sectional predictability literature, we have observations on each characteristic in every month of our sample and do not fill entire cross-sections with values of zero. Different scales and extreme observations are dealt with by cross-sectionally rank normalizing firm characteristics.

5 Results

5.1 Predictive accuracy

Table 2 presents monthly out-of-sample R^2 for the employed methods and different levels of market liquidity. Column 1 displays the levels of market liquidity established by removing stocks with the prior month's market values below the cross-sectional decile breakpoints corresponding to q . Columns 2 to 9 correspond to the R^2 values as a percentage achieved by each individual model. Positive values indicate better predictive accuracy compared to a naive constant zero forecast.

For the full cross-section of stocks (first row), OLS generates an R^2 of 0.43%. After dropping firms with market values below cross-sectional thresholds in increments of ten percent, the predictive power gradually decreases to - 0.32% for the largest stocks. Penalized linear methods, i.e., Lasso, Ridge and Enet, perform on a very similar, yet slightly worse, level than OLS. This result is not surprising, as we are using only a small set of well-established, and essentially uncorrelated firm characteristics. The predictive accuracy of SVR is consistently below that of OLS. A potential source of this underperformance is that distance-based algorithms such

Table 3 This table reports p values of pairwise Diebold-Mariano (DM) tests comparing stock-level return forecasts of various models against the null hypothesis of equal predictive accuracy

	OLS	Lasso	Ridge	Enet	SVR	RF	GBRT	NN
<i>Panel A: Full sample</i>								
OLS	–							
Lasso	0.84	–						
Ridge	0.58	0.19	–					
Enet	0.95	0.92	1.00	–				
SVR	0.98	0.96	0.96	0.93	–			
RF	0.00	0.00	0.01	0.00	0.00	–		
GBRT	0.02	0.02	0.02	0.01	0.01	0.41	–	
NN	0.02	0.01	0.01	0.00	0.01	0.24	0.17	–
<i>Panel B: Size > q_{90}</i>								
OLS	–							
Lasso	0.65	–						
Ridge	0.27	0.20	–					
Enet	0.49	0.27	0.66	–				
SVR	0.85	0.77	0.95	0.90	–			
RF	0.73	0.60	0.90	0.78	0.22	–		
GBRT	0.04	0.04	0.03	0.03	0.00	0.01	–	
NN	0.06	0.05	0.05	0.04	0.01	0.02	0.25	–

Panel A shows test results comparing predictions for the full cross-section of stocks, and Panel B uses only predictions for the largest 10% of stocks. A p value of less than 0.1 (bold figures) indicates inferior performance of the column model against the row model

as kernel SVR are particularly fragile in the presence of low-impact predictors. Panel B of Table 1 indicates that this may be the case for size, as the return differences across portfolios are only of minuscule amplitude. Additionally, SVRs are highly sensitive to the choice of hyperparameters, as Probst et al. (2019) identify a much higher tunability of SVR compared to that of RF or GBRT. RF, GBRT and NN, however, exhibit a remarkable improvement over OLS for the broad market (0.58%, 0.60% and 0.67%, respectively). With respect to persistence, they exhibit a diminishing pattern in predictability toward larger stocks similar to that of OLS. This pattern is in line with Hou et al. (2018) showing that much of the predictability detected in the previous literature is attributable to the smallest stocks. GBRT and NN, however, are able to achieve both (1) consistently greater R^2 than OLS and (2) consistently positive values for R^2 . The differences range between 0.46 and 0.56 percentage points for GBRT and NN, respectively, indicating that the predictive accuracy can be improved by (choosing the appropriate) machine learning methods.

Table 3 presents the p values for the DM test on the full sample of stocks (Panel A) and for the top decile of market capitalization (Panel B). We display the lower triangle only, in which row models correspond to forecast errors \hat{e}_1 in equation (22); thus, significant p -values indicate that the row model outperforms the column model.¹⁴

¹⁴ The upper triangle is one minus the lower triangle.

Table 4 This table reports the performance characteristics of prediction-based decile portfolios formed from different models

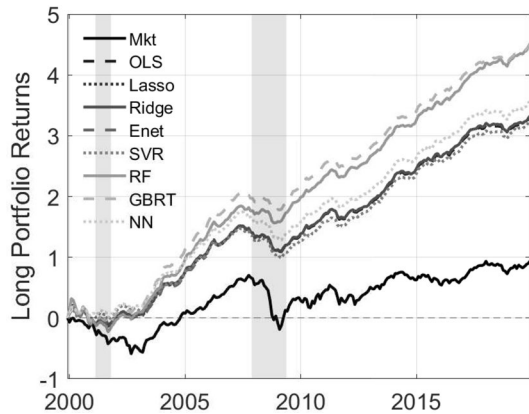
	OLS				Lasso				Ridge			
	Avg	Std	Shp	Pred	Avg	Std	Shp	Pred	Avg	Std	Shp	Pred
L	-0.29	6.54	-0.04	-0.22	-0.21	6.51	-0.03	0.06	-0.27	6.64	-0.04	-0.05
2	0.16	5.25	0.03	0.17	0.01	5.34	0.00	0.32	0.13	5.29	0.02	0.27
3	0.34	4.40	0.08	0.42	0.24	4.50	0.05	0.52	0.35	4.39	0.08	0.48
4	0.53	4.10	0.13	0.62	0.45	4.17	0.11	0.69	0.53	4.09	0.13	0.65
5	0.65	3.83	0.17	0.79	0.62	3.87	0.16	0.84	0.66	3.80	0.17	0.80
6	0.88	3.61	0.24	0.95	0.84	3.62	0.23	0.97	0.89	3.60	0.25	0.94
7	1.03	3.45	0.30	1.12	0.90	3.55	0.25	1.11	1.02	3.46	0.30	1.08
8	1.21	3.35	0.36	1.30	1.10	3.47	0.32	1.26	1.20	3.32	0.36	1.23
9	1.25	3.40	0.37	1.52	1.12	3.57	0.31	1.44	1.26	3.41	0.37	1.42
H	1.45	3.58	0.40	1.85	1.38	3.83	0.36	1.70	1.46	3.56	0.41	1.70
H-L	1.74	5.09	0.34	2.07	1.79	5.36	0.33	1.70	1.73	5.21	0.33	1.75
pVal(H-L)	0.04				0.08				0.04			
pVal(MR)	0.00	51.10	0.20		0.40	57.70	0.50		0.00	39.30	0.10	
	Enet				SVR				RF			
	Avg	Std	Shp	Pred	Avg	Std	Shp	Pred	Avg	Std	Shp	Pred
L	-0.10	6.54	-0.01	0.14	-0.35	6.68	-0.05	-0.57	-0.40	6.83	-0.06	-0.22
2	0.04	5.53	0.01	0.31	0.18	5.12	0.03	-0.04	0.09	5.06	0.02	0.29
3	0.28	4.68	0.06	0.51	0.45	4.30	0.10	0.23	0.39	4.40	0.09	0.51
4	0.47	4.15	0.11	0.68	0.61	3.87	0.16	0.42	0.53	4.03	0.13	0.67
5	0.65	3.99	0.16	0.83	0.75	3.78	0.20	0.57	0.67	3.81	0.18	0.83
6	0.84	3.63	0.23	0.97	0.90	3.65	0.25	0.70	0.87	3.46	0.25	0.97
7	0.97	3.58	0.27	1.11	0.97	3.61	0.27	0.83	0.89	3.30	0.27	1.08
8	1.10	3.46	0.32	1.26	1.10	3.53	0.31	0.97	1.02	3.34	0.31	1.18
9	1.13	3.51	0.32	1.45	1.20	3.45	0.35	1.14	1.20	3.57	0.34	1.32

Table 4 (continued)

	Enet				SVR				RF			
	Avg	Std	Shp	Pred	Avg	Std	Shp	Pred	Avg	Std	Shp	Pred
H	1.41	3.74	0.38	1.71	1.42	3.38	0.42	1.40	1.97	4.00	0.49	1.76
H-L	1.78	5.60	0.32	1.72	1.77	4.96	0.36	1.97	2.38	4.90	0.48	1.98
pVal(H-L)	0.24				0.03				0.00			
pVal(MR)	0.10	51.50	0.30		0.00	0.40	0.00		0.10	71.60	0.00	
	GBRT				NN							
	Avg	Std	Shp	Pred	Avg	Std	Shp	Pred				
L	− 0.36	6.35	− 0.06	− 0.22	− 0.27	6.28	− 0.04	0.03				
2	0.10	4.91	0.02	0.19	0.11	5.13	0.02	0.36				
3	0.23	4.79	0.05	0.36	0.36	4.47	0.08	0.55				
4	0.46	3.99	0.12	0.51	0.46	4.08	0.11	0.70				
5	0.70	3.67	0.19	0.67	0.66	3.91	0.17	0.83				
6	0.70	3.52	0.20	0.78	0.86	3.58	0.24	0.96				
7	0.84	3.38	0.25	0.89	0.98	3.50	0.28	1.08				
8	1.08	3.35	0.32	1.01	1.15	3.44	0.34	1.22				
9	1.24	3.65	0.34	1.15	1.33	3.28	0.41	1.39				
H	1.97	3.67	0.54	1.44	1.56	3.39	0.46	1.64				
H-L	2.33	4.16	0.56	1.66	1.83	4.38	0.42	1.61				
pVal(H-L)	0.00				0.01							
pVal(MR)	0.00	70.20	0.00		0.00	38.50	0.00					

The ten portfolios are defined as follows: At every month in our testing sample from January 2000 to December 2019, we sort all stocks by their expected return for the next month and assign them to the corresponding decile portfolio. To analyze the aggregate effect of a certain model on stock returns, we take a long position in portfolio 10 (High) and a short position in portfolio 1 (Low). Avg (Std, Shp) is the average (standard deviation, Sharpe ratio) of equally weighted returns. Pred is the average expected portfolio return of a model. pVal(H-L) reports the p value (in percent) from a t -test against the null hypothesis of zero average return (column Avg). pVal(MR) reports the p value (in percent) from a one-tailed monotonic relationship test in averages (Avg), standard deviations (Std), and Sharpe ratios (Shp) of the ten portfolios. The direction of the test depends on the sign of the difference between the values in the High and Low portfolios

Fig. 5 This figure shows cumulative logarithmic returns of prediction-based portfolios for different models and the market (Mkt). The strategy is implemented as follows: In every month of our testing sample from January 2000 to December 2019, we sort all stocks by their expected return for the next month and take a long position in stocks corresponding to the top decile of expected returns



The DM test provides support for the empirical results derived from the R^2 values presented in Table 2. The first conclusion from Panel A of Table 3 is that predictions from penalized linear methods and SVR do not provide statistically significant improvements over OLS. More specifically, SVR and Enet even exhibit a significantly inferior predictive ability at the 5% level. Tree-based methods and NN produce statistically significant improvements over linear methods and SVR at least at the 5% level. For any pairwise comparison among tree-based methods and NN, the null of equal predictive accuracy cannot be rejected. For Panel B of Table 3, we find that the p-values of penalized linear methods and SVR generally decrease compared to OLS, indicating that they experience relative improvement when predicting returns of large cap stocks. Conversely, the superiority of RF over linear models and SVR disappears, indicating that its predictive performance is driven by picking up small-scale inefficiencies present in microcap dynamics. For GBRT and NN, the null hypothesis against not only OLS but also against all other models can still be rejected, indicating that their outperformance is present throughout all market segments.

5.2 Economic gains

From an investor's perspective, statistical performance is only of subordinate importance. In addition, Leitch and Tanner (1991) argue that statistical metrics are only loosely related to a forecast's achievable profit. Next, we assess the economic performance of our models by portfolio sorts. At the end of each month, we generate return forecasts for the next month with each model. After that, we assign stocks to one of ten portfolios based on decile breakpoints of the forecasts. Finally, to capture the aggregate effect, we construct a zero-net-investment portfolio that simultaneously buys (sells) the stocks with the highest (lowest) expected returns, i.e., decile 10 and 1, respectively. Table 4 depicts the results using the full sample. Note that

Table 5 This table reports the return of a prediction-based zero-investment strategy formed from different models

Size > q	OLS	Lasso	Ridge	Enet	SVR	RF	GBRT	NN
0.00	1.74	1.79	1.73	1.78	1.77	2.38	2.33	1.83
0.10	1.79	1.89	1.79	1.84	1.95	1.79	1.83	1.87
0.20	1.84	1.92	1.84	1.89	2.00	1.85	1.84	1.93
0.30	1.76	1.81	1.77	1.82	1.93	1.76	1.79	1.87
0.40	1.67	1.70	1.66	1.73	1.80	1.66	1.70	1.78
0.50	1.51	1.53	1.48	1.51	1.64	1.46	1.53	1.65
0.60	1.32	1.32	1.31	1.32	1.44	1.33	1.40	1.51
0.70	1.13	1.10	1.15	1.15	1.27	1.10	1.19	1.24
0.80	0.77	0.74	0.76	0.75	0.93	0.67	0.91	0.91
0.90	0.53	0.48	0.55	0.58	0.73	0.46	0.59	0.70

The strategy is implemented as follows: At every month in our testing sample from January 2000 to December 2019, we sort all stocks by their expected return for the next month and take a long position in stocks corresponding to the top decile and a short position in stocks corresponding to the bottom decile of predictions. We repeat this procedure for several levels of liquidity by removing stocks with market capitalization (Size) less than the cross-sectional quantile q . Columns 2–9 then show realized returns of a zero-investment strategy based on predictions formed by the model shown in the column header

all models share a monotonically increasing relationship in realized returns (column ‘Avg’) and in the Sharpe ratio (column ‘Shp’), which is statistically significant at the 1% level after testing for monotonicity between portfolios 1 to 10 using the Patton and Timmermann (2010) test (row ‘pVal(MR)’). From an investor’s perspective, this property is highly desirable, as higher returns are, thus, relatively cheaper in terms of risk than low returns. Moreover, we see significantly positive spreads (row ‘H-L’) at the 1% significance level for all models, indicating that, generally, all models are effective in capturing the cross-sectional dispersion of returns (row ‘pVal(H-L)’). Upon examining the return spreads, we find penalized linear methods and SVR to be on eye-level with OLS, with Lasso, Enet and SVR resulting in slightly larger spreads. NN and tree-based methods outperform OLS by a quite substantial margin, with a surplus in return spreads of up to 64 basis points for RF. The zero-net-investment strategy for the predictions of RF yields an extraordinary return of 2.38% per month (32.61% annualized), while GBRTs provide the best risk-return-tradeoff with a Sharpe ratio of 0.56 per month (1.94 annualized). Even if SVR exhibits a negligible improvement in terms of return spread, it has a unique characteristic from a risk perspective, as it is the only model with a significantly monotonic increasing relationship in returns while having a significantly monotonic decreasing relationship with risk, i.e., higher returns are associated with lower risk, which is extremely compelling from an investor’s perspective. Figure 5 presents cumulative performance plots corresponding to the long portfolios (portfolio ‘H’) of Table 4 along with the cumulative market return. The market return for Europe is retrieved from Kenneth French’s website.¹⁵ Note that all models outperform the market index and

¹⁵ <https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/>.

Table 6 This table reports average monthly percentage changes in holdings, i.e., turnover, for different models and levels of market liquidity in terms of market capitalization (Size)

Size > q	OLS	Lasso	Ridge	Enet	SVR	RF	GBRT	NN
0.00	125	124	124	125	133	154	144	132
0.10	125	125	126	126	131	152	143	131
0.20	124	126	123	126	131	152	141	133
0.30	125	126	124	125	131	153	140	131
0.40	124	126	124	124	129	154	140	133
0.50	125	128	125	126	129	151	140	133
0.60	127	129	128	129	127	150	141	131
0.70	128	129	128	129	127	147	140	127
0.80	127	129	128	129	127	144	138	127
0.90	126	129	127	130	129	142	140	127

that tree-based methods dominate all other models by a large margin. Consistent with Leitch and Tanner (1991), we find that the statistical performance reported in the previous subsection does not necessarily translate into economic performance.

Table 5 presents return spreads analogous to row ‘H-L’ in Table 4 for the different levels of market capitalization. In Table 2, the predictive accuracy gradually decreases for larger stocks. For the return spreads we generally find a similar pattern; however, most models exhibit a peak (economic) performance after cutting off the smallest 20%. This level of market liquidity is often considered free of microcaps (e.g., Lewellen 2015). A possible interpretation might be that the dynamics of those stocks are neither very persistent nor predictable. Tree-based methods, on the other hand, experience a sharp drop in economic performance after cutting off the bottom decile of market capitalization, again indicating that their outperformance, to a large extent, relies on capturing microcap dynamics. Further increasing levels of market capitalization are accompanied by decreasing return spreads, a pattern consistent with the concept of market efficiency, assuming that the information environment for large firms is richer and more efficient than that for small firms. Penalized linear methods are again close to OLS. The outperformance of tree-based methods is mostly present when considering the full sample. While outperformance vanishes for RF afterward, GBRTs maintain a slight surplus over OLS toward reducing the sample to the largest stocks. NN and SVR consistently achieve higher return spreads than OLS across all market segments. While the results for NN are in line with the previous findings, the results for SVR come as quite of a surprise given its low predictive accuracy (cf. Table 2). Specifically, after cutting off the bottom decile of market capitalization, it exhibits the highest returns for the zero-net-investment strategy overall with remarkable stability. The evaluation thus far provides evidence that machine learning can lead to better results than the traditional OLS model, not only statistically but also economically.

Table 7 This table reports monthly predictive panel R^2 as a percent

Size > q	OLS	Lasso	Ridge	Enet	SVR	RF	GBRT	NN
<i>Panel A: Bull</i>								
0.00	1.22	1.18	1.22	1.16	0.74	1.33	1.19	1.29
0.10	1.29	1.24	1.29	1.22	0.83	1.23	1.12	1.37
0.20	1.35	1.29	1.35	1.27	0.91	1.28	1.19	1.44
0.30	1.38	1.31	1.38	1.29	0.97	1.30	1.22	1.46
0.40	1.42	1.35	1.42	1.35	1.05	1.37	1.29	1.52
0.50	1.45	1.37	1.46	1.38	1.13	1.39	1.32	1.55
0.60	1.44	1.37	1.45	1.38	1.21	1.40	1.34	1.50
0.70	1.40	1.32	1.42	1.35	1.22	1.36	1.36	1.48
0.80	1.22	1.18	1.26	1.21	1.15	1.21	1.35	1.36
0.90	0.99	0.95	1.06	1.00	1.10	1.18	1.36	1.23
<i>Panel B: Bear</i>								
0.00	- 0.40	- 0.41	- 0.42	- 0.46	- 0.48	- 0.20	- 0.01	0.03
0.10	- 0.54	- 0.55	- 0.57	- 0.61	- 0.51	- 0.47	- 0.21	- 0.08
0.20	- 0.57	- 0.60	- 0.61	- 0.66	- 0.49	- 0.51	- 0.24	- 0.08
0.30	- 0.58	- 0.63	- 0.62	- 0.68	- 0.52	- 0.53	- 0.24	- 0.04
0.40	- 0.60	- 0.65	- 0.64	- 0.70	- 0.55	- 0.58	- 0.25	- 0.01
0.50	- 0.67	- 0.73	- 0.70	- 0.75	- 0.66	- 0.69	- 0.28	- 0.02
0.60	- 0.73	- 0.81	- 0.75	- 0.81	- 0.80	- 0.81	- 0.29	- 0.02
0.70	- 0.87	- 0.97	- 0.89	- 0.94	- 1.03	- 0.99	- 0.35	- 0.10
0.80	- 1.03	- 1.13	- 1.02	- 1.07	- 1.38	- 1.27	- 0.50	- 0.18
0.90	- 1.27	- 1.37	- 1.27	- 1.33	- 1.80	- 1.63	- 0.75	- 0.48

Monthly predictions are formed for our testing sample from January 2000 to December 2019 for different models and levels of market liquidity in terms of market capitalization. Columns correspond to the models. Rows correspond to different levels of market liquidity, which we establish by removing stocks with market capitalization (Size) less than the cross-sectional quantile q . The results are presented separately for bull and bear markets, i.e., when the trailing cumulative 12-month excess return on the market portfolio in months $t-12$ to $t-1$ is positive and negative, respectively

6 Additional results

6.1 Transaction costs

Transaction costs impair the profitability of any investment. Higher turnover rates of investment strategies can easily nullify their overperformance in terms of raw returns. Therefore, we aim to examine whether the stronger performance of machine learning models compared to OLS is driven by substantially higher turnovers. However, realistic transaction costs are not always available, cumbersome to use or expensive to acquire (cf. Lesmond et al. 1999). Additionally, Collins and Fabozzi

(1991) state that true transaction costs are inherently unobservable. Consequently, any good-faith estimate of transaction costs would probably be wrong. We consider the average monthly turnover of the spread portfolios ('H-L') calculated as:¹⁶

$$Turnover = \frac{1}{T} \sum_{t=1}^T \left(\sum_i \left| w_{i,t+1} - \frac{w_{i,t}(1 + r_{i,t+1})}{\sum_j w_{j,t}(1 + r_{j,t+1})} \right| \right), \quad (23)$$

where $w_{i,t}$ is the weight of stock i in the portfolio at time t .

Table 6 reports the average monthly turnover in percent for all models and levels of market liquidity. In line with Gu et al. (2020), we find that turnover rates for tree-based methods are generally approximately 20% higher than those for OLS. As the outperformance of tree-based methods appeared to be meaningful only when including microcaps, we conclude that their potential is limited by the higher turnover since microcaps are inherently more difficult and costlier to trade. Penalized linear methods, SVR and NN, on the other hand, have turnover rates comparable to those of OLS. Especially on the largest stocks, where transaction costs are generally smaller, SVR and NN still achieve a spread of up to 20 basis points above OLS. We conclude that the marginally higher turnovers are unlikely to explain the pronounced outperformance or, to put it differently, outperformance is unlikely to disappear in the light of transaction costs.

6.2 Performance within subperiods

Inspired by a number of studies suggesting that returns may differ across market states (e.g., Cooper et al. 2004; Wang and Xu 2015), we evaluate the performance of the models separately for bull and bear markets. Following (Zaremba et al. 2020), we define these states as subperiods of positive (bull market) and negative (bear market) total excess return on the market portfolio during the last 12 months. The results for the predictive accuracy and economic gains from the individual models are reported in Tables 7 and 8, respectively.

The first conclusion from Table 7 is that return predictability is generally higher in bull markets (Panel A) than in bear markets (Panel B). While we encounter decreasing predictability with firm size in the full panel (cf. Table 2), we find that in bull markets, the predictability in terms of R^2 is greater when excluding very small stocks. For bear markets, the R^2 values are generally negative and monotonically decreasing in firm size. Among the machine learning methods, NN produces the highest R^2 and shows the highest outperformance compared to OLS in both bull and bear markets.

With regard to the economic gains reported in Table 8, we find that differences are visible, although not dramatic, during times of a general upward trend (Panel A). In

¹⁶ The turnover of a zero-net-investment portfolio falls between 0% (no turnover at all) and 200% (a full reallocation of both the high and low portfolios). To put it in perspective, a strategy based only on *mom* generates a turnover of approximately 160% over the sample period.

bear markets, however, the differences become more apparent. In particular, SVR and NN exhibit consistently higher returns compared to OLS. Even for the 10% largest stocks, the difference of approximately 40 basis points is quite staggering. The results presented here suggest that machine learning does not add exceptional value over OLS during periods of a general upward trend. However, in downward markets—which is when investors usually lose money—these models generate considerable value.

6.3 Forecast combinations

Lastly, we consider forecast combinations. In general, forecast combination acts as a tool for risk diversification across individual models while, in many cases, simultaneously improving performance (Timmermann 2006). Some of the models presented herein (i.e., RF, GBRT, and NN) utilize the benefits of building ensembles as an integral part of the model definition itself. Several studies have highlighted the benefits of forecast combinations and proposed different methods with respect to choosing optimal model weights (e.g., Bates and Granger 1969; Winkler and Makridakis 1983). However, here, we use the forecast combinations as a simple proxy to test whether linear methods contain information complementary to that of the machine learning models. To do this, we form combinations from 1) all models and 2) only nonlinear methods (SVR, RF, GBRT, NN). If there is additional value, we would expect 1), i.e., the combination of all models, to provide superior predictions. Therefore, we restrict ourselves to simple combination schemes that often show better results than more sophisticated methods involving additional parameter estimates (Timmermann 2006).

The forecast combination takes the general form:

$$\hat{r}_{i,t+1}^{\text{COMB}} = \sum_{h=1}^H w_t^h \hat{r}_{i,t+1}^h, \quad (24)$$

with

$$w_t^h = \begin{cases} \frac{1}{H} & \text{Equal-weighted (EW)} \\ \frac{1/MSE_t^h}{\sum_{m=1}^M 1/MSE_t^m} & \text{Inverse MSE-weighted (IMSE)} \\ \frac{1/R_t^h}{\sum_{m=1}^M 1/R_t^m} & \text{Inverse MSE rank-weighted (IRANK),} \end{cases} \quad (25)$$

where w_t^h is the weight of model h at time t and H is the total number of models. *MSE* refers to the in-sample mean-squared error, and *R* refers to the ranking of MSE (ascending order) among all models.

Table 9 reports results on the predictive accuracy of forecast combinations. The left-hand panel considers all models in the combination, while the right-hand panel excludes linear methods. We find that the predictive accuracy is consistently higher when relying only on nonlinear methods, irrespective of the concrete

Table 8 This table reports the return of a prediction-based zero-investment strategy from the models

Size > q	OLS	Lasso	Ridge	Enet	SVM	RF	GBRT	NN
<i>Panel A: Bull</i>								
0.00	1.77	1.73	1.75	1.68	1.41	2.33	2.21	1.76
0.10	1.84	1.80	1.82	1.73	1.60	1.85	1.83	1.84
0.20	1.85	1.80	1.83	1.74	1.66	1.91	1.84	1.90
0.30	1.74	1.72	1.78	1.73	1.67	1.76	1.77	1.82
0.40	1.66	1.60	1.65	1.65	1.53	1.75	1.74	1.75
0.50	1.56	1.53	1.54	1.52	1.47	1.59	1.59	1.70
0.60	1.31	1.33	1.36	1.37	1.33	1.46	1.42	1.45
0.70	1.12	1.04	1.18	1.12	1.16	1.09	1.14	1.17
0.80	0.77	0.72	0.78	0.78	0.90	0.74	0.93	0.91
0.90	0.66	0.57	0.66	0.72	0.69	0.47	0.66	0.68
<i>Panel B: Bear</i>								
0.00	1.69	1.86	1.71	1.89	2.23	2.43	2.48	1.91
0.10	1.74	2.01	1.74	1.96	2.40	1.71	1.83	1.91
0.20	1.81	2.07	1.86	2.05	2.42	1.76	1.83	1.97
0.30	1.78	1.93	1.76	1.93	2.26	1.75	1.81	1.93
0.40	1.68	1.84	1.67	1.82	2.14	1.55	1.66	1.83
0.50	1.44	1.54	1.40	1.49	1.86	1.29	1.45	1.58
0.60	1.32	1.30	1.24	1.27	1.59	1.15	1.37	1.59
0.70	1.14	1.17	1.11	1.19	1.41	1.11	1.26	1.32
0.80	0.77	0.76	0.72	0.71	0.97	0.58	0.88	0.92
0.90	0.37	0.37	0.40	0.43	0.77	0.44	0.51	0.73

The strategy is implemented as follows: At every month in our testing sample from January 2000 to December 2019, we sort all stocks by their expected return for the next month and take a long position in stocks corresponding to the top decile and a short position in stocks corresponding to the bottom decile of predictions. We repeat this procedure for several levels of liquidity by removing stocks with market capitalization (Size) less than the cross-sectional quantile q . Columns 2–9 then show the realized returns of a zero-investment strategy based on predictions formed by the model shown in the column header. The results are presented separately for bull and bear markets, i.e., when the trailing cumulative 12-month excess return on the market portfolio in months $t-12$ to $t-1$ is positive and negative, respectively

combination scheme used. With respect to the weighting-schemes, inverse rank-weighting appears to provide slightly worse results, while the other two are on equal footing. Most importantly, we exhibit two key take-aways: First, all combinations outperform the single use of OLS. Second, and in contrast, none of the combinations is superior to the strongest machine learning model, NN. However, the best model may not be known *ex ante*; thus, the combination of various models might still reduce model selection risk.

Table 10 reports return spreads based on the forecast combinations. The results generally support the statistical findings. However, we find that the economic performance improves over the best individual model, SVR, up to $Size > 0.7$. Overall, the performance of the conventional OLS approach can be improved when combined with models from the machine learning literature. Moreover, our results raise the

Table 9 This table reports monthly predictive panel R^2 as a percent

Size > q	Including linear methods			Excluding linear methods		
	EW	IRANK	IMSE	EW	IRANK	IMSE
0.00	0.55	0.60	0.55	0.64	0.67	0.64
0.10	0.45	0.47	0.45	0.51	0.51	0.51
0.20	0.46	0.48	0.46	0.53	0.52	0.53
0.30	0.46	0.47	0.46	0.54	0.52	0.54
0.40	0.45	0.47	0.45	0.54	0.53	0.54
0.50	0.42	0.43	0.42	0.52	0.49	0.52
0.60	0.37	0.37	0.37	0.47	0.44	0.47
0.70	0.29	0.28	0.29	0.39	0.36	0.39
0.80	0.11	0.08	0.11	0.21	0.16	0.21
0.90	-0.16	-0.19	-0.16	-0.05	-0.10	-0.05

Monthly predictions are formed for our testing sample from January 2000 to December 2019 for different models (columns) and levels of market liquidity in terms of market capitalization (Size, rows). Model predictions are then combined by simple averaging ('EW'), inverse rank-weighting ('IRANK') and inverse MSE-weighting ('IMSE')

Table 10 This table reports the return of a prediction-based zero-investment strategy formed from different forecast combination schemes, i.e., simple averaging ('EW'), inverse rank-weighting ('IRANK') and inverse MSE-weighting ('IMSE')

Size > q	Including linear methods			Excluding linear methods		
	EW	IRANK	IMSE	EW	IRANK	IMSE
0.00	2.27	2.32	2.27	2.46	2.47	2.46
0.10	1.97	1.94	1.98	2.05	2.02	2.05
0.20	2.02	1.97	2.02	2.09	2.05	2.09
0.30	1.94	1.85	1.94	1.99	1.93	2.00
0.40	1.87	1.78	1.87	1.90	1.83	1.90
0.50	1.67	1.63	1.67	1.74	1.68	1.74
0.60	1.48	1.41	1.48	1.51	1.46	1.51
0.70	1.24	1.21	1.24	1.33	1.23	1.34
0.80	0.83	0.78	0.83	0.89	0.81	0.89
0.90	0.62	0.61	0.62	0.67	0.66	0.66

The strategy is implemented as follows: At every month in our testing sample from January 2000 to December 2019, we sort all stocks by their expected return for the next month and take a long position in stocks corresponding to the top decile and a short position in stocks corresponding to the bottom decile of predictions. We repeat this procedure for several levels of liquidity by removing stocks with market capitalization less than the cross-sectional quantile q . Columns 2–4 and 5–7 then show the realized returns of a zero-investment strategy based on predictions formed by the forecast combination shown in the header

question of whether the commonly applied purely linear model is even necessary for building expectations about future stock returns, as the combination of nonlinear methods provides superior performance both statistically and economically while individual model selection risk is reduced.

7 Conclusion

In this paper, we provide an eye-level comparison of the commonly applied linear regression approach in stock return predictions (Lewellen 2015) and machine learning methods, i.e., penalized linear methods, support vector regression, tree-based methods, and neural networks. We find that the nonlinear machine learning methods in particular can provide both statistically and economically meaningful performance gains over the conventional OLS approach, thereby revealing that machine learning methods are not more vulnerable than OLS to model breakdowns.¹⁷ In our analysis, GBRTs exhibit the highest outperformance, of approximately 0.6% per month, based on the entire cross-section of European stocks, which decreases to 0.1% per month when only stocks of the largest ten percent of firms are considered. Overall, we find that NN provides the most reliable predictions in terms of both statistical and economic performance, as its superior performance is robust over all market segments. In addition, we find that the economic performance gains are not attributable to substantially higher turnover of stocks. Analyzing bull and bear markets separately, we find that the value of NN is mostly generated in downward trending markets when investors generally lose money. As forecast combinations have frequently been found to produce better forecasts than methods based on the *ex ante* best individual forecasting models (Timmermann 2006), we investigate the performance of forecast combinations and find that the combination of only nonlinear models exhibits the performance of the best individual model, which is good news, as it allows investors to reduce model selection risk by diversification. We further find that all forecast combinations in this study outperform the sole use of OLS. Our results are derived based on a small set of well-established predictor variables borrowed from the asset pricing literature. However, as the performance of the (nonlinear) machine learning methods is generally higher than that of OLS, we strongly recommend a more intensive use of machine learning models to predict stock returns in future research and practice. Beyond our considerations, future research should also examine the use of machine learning techniques with respect to long-term stock returns (as considered in Kyriakou et al. 2021; Mammen et al. 2019) more closely.

Appendix

See Table 11.

¹⁷ At this point, we return to the cross-sectional vs. time-series predictability discussion introduced in Sect. 2. A benchmark method commonly found in the time-series predictability literature is local linear smoothing (LLS). We have also used this method for cross-sectional predictability purposes and can report from unreported results that LLS is outperformed by the machine learning methods used here.

Table 11 This table describes the parameter ranges considered in the grid search

Lasso	Ridge	Enet	SVR
$\lambda \in [10^{-4}, 10^{-1}]$	$\lambda \in [10^{-4}, 10^{-1}]$	$\alpha = 0.5$ $\lambda \in [10^{-4}, 10^{-1}]$	$\gamma \in [10^{-1}, 10^3]$ $C \in [10^{-1}, 10^3]$ $\epsilon \in [10^{-3}, 10^{-1}]$
RF	GBRT	NN	
$B = 300$	$B = [1, 2500]$	$L = 2$	
$J \in [1, 10]$	$J \in \{1, 5\}$	$K = 12$	
$p \in \{2, 5\}$	$v \in \{10^{-2}, 10^{-1}\}$	$act. = tanh$ $lr \in \{10^{-3}, 10^{-2}\}$ $\lambda \in [10^{-5}, 10^{-3}]$	

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Afifi AA, Elashoff RM (1966) Missing observations in multivariate statistics I. Review of the literature. *J Am Stat Assoc* 61(315):595–604
- Annaert J, de Ceuster M, Versteegen K (2013) Are extreme returns priced in the stock market? European evidence. *J Bank Financ* 37(9):3401–3411
- Banz RW (1981) The relationship between return and market value of common stocks. *J Financ Econ* 9(1):3–18
- Bates JM, Granger CWJ (1969) The combination of forecasts. *J Op Res Soc* 20(4):451–468
- Beckmeyer H, Wiedemann T (2022) Recovering missing firm characteristics with attention-based machine learning. *SSRN Electron J*. <https://doi.org/10.2139/ssrn.4003455>
- Bianchi D, Büchner M, Tamoni A (2021) Bond risk premiums with machine learning. *Rev Financ Stud* 34(2):1046–1089
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. Chapman & Hall, New York, NY
- Collins Bruce M, Faziozi Frank J (1991) A methodology for measuring transaction costs. *Financ Anal J* 47(2):27–36

- Bryzgalova S, Lerner S, Lettau M, Pelger M (2022) Missing financial data. SSRN Electron J. <https://doi.org/10.2139/ssrn.4106794>
- Cahan E, Bai J, Ng S (2022) Factor-based imputation of missing values and covariances in panel data of large dimensions. *J Econom* (forthcoming). <https://doi.org/10.1016/j.jeconom.2022.01.006>
- Campbell CJ, Cowan AR, Salotti V (2010) Multi-country event-study methods. *J Bank Financ* 34(12):3078–3090
- Campbell JY (1987) Stock returns and the term structure. *J Financ Econ* 18(2):373–399
- Carhart MM (1997) On persistence in mutual fund performance. *J Financ* 52(1):57–82
- Cismondi F, Fialho AS, Vieira SM, Reti SR, Sousa JMC, Finkelstein SN (2013) Missing data in medical databases: Impute, delete or classify? *Artif Intell Med* 58(1):63–72
- Cooper MJ, Gutierrez RC Jr, Hameed A (2004) Market states and momentum. *J Financ* 59(3):1345–1365
- Cybenko G (1989) Approximation by superpositions of a sigmoidal function. *Math Control Signals Syst* 2(4):303–314
- Das J, Do QT, Shaines K, Srikant S (2013) U.S. and them: the geography of academic research. *J Dev Econ* 105:112–130
- Diebold FX, Mariano RS (1995) Comparing predictive accuracy. *J Bus Econ Stat* 13(3):253–263
- Dong X, Li Y, Rapach DE, Zhou G (2022) Anomalies and the expected market return. *J Financ* 77(1):639–681
- Drobtetz W, Otto T (2021) Empirical asset pricing via machine learning: evidence from the European stock market. *J Asset Manag* 22(7):507–538
- Engelberg J, McLean RD, Pontiff J, Ringgenberg MC (2022) Do cross-sectional predictors contain systematic information?. *J Financ Quant Anal* (forthcoming)
- Fama EF, French KR (1988) Dividend yields and expected stock returns. *J Financ Econ* 22(1):3–25
- Fama EF, French KR (1989) Business conditions and expected returns on stocks and bonds. *J Financ Econ* 25(1):23–49
- Fama EF, French KR (1993) Common risk factors in the returns on stocks and bonds. *J Financ Econ* 33(1):3–56
- Fama EF, French KR (2008) Dissecting anomalies. *J Financ* 63(4):1653–1678
- Fama EF, French KR (2015) A five-factor asset pricing model. *J Financ Econ* 116(1):1–22
- Fong KYL, Holden CW, Trzcinka CA (2017) What are the best liquidity proxies for global research? *Rev Financ* 21(4):1355–1401
- Frazzini A, Pedersen LH (2014) Betting against beta. *J Financ Econ* 111(1):1–25
- Freyberger J, Höppner B, Neuhierl A, Weber M (2021) Missing data in asset pricing panels. SSRN Electron J. <https://doi.org/10.2139/ssrn.3932438>
- Gibson T (1906) The pitfalls of speculation. The Moody Corporation, New York
- Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge, Massachusetts and London, England
- Goyal A, Wahal S (2015) Is momentum an echo? *J Financ Quant Anal* 50(6):1237–1267
- Green J, Hand JRM, Zhang XF (2017) The characteristics that provide independent information about average US monthly stock Returns. *Rev Financ Stud* 30(12):4389–4436
- Griffin JM, Kelly PJ, Nardari F (2010) Do market efficiency measures yield correct inferences? A comparison of developed and emerging markets. *Rev Financ Stud* 23(8):3225–3277
- Griffin JM, Hirschey NH, Kelly PJ (2011) How important is the financial media in global markets? *Rev Financ Stud* 24(12):3941–3992
- Gu S, Kelly B, Xiu D (2020) Empirical asset pricing via machine learning. *Rev Financ Stud* 33(5):2223–2273
- Hanauer MX (2020) A comparison of global factor models. SSRN Electron J. <https://doi.org/10.2139/ssrn.3546295>
- Hansen LK, Salamon P (1990) Neural network ensembles. *IEEE Trans Pattern Anal Mach Intell* 12(10):993–1001
- Harvey CR (2017) Presidential address: the scientific outlook in financial economics. *J Financ* 72(4):1399–1440
- Harvey CR, Liu Y, Zhu H (2016) ... and the cross-section of expected returns. *Rev Financ Stud* 29(1):5–68
- Hastie TJ, Tibshirani RJ, Friedman JH (2017) The elements of statistical learning: Data mining, inference, and prediction, second edition, corrected at 12th printing, 2017th edn. Springer Series in Statistics. Springer, New York, NY

- Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. *Neural Netw* 2(5):359–366
- Hou K, Xue C, Zhang L (2018) Replicating anomalies. *Rev Financ Stud* 33(5):2019–2133
- Ince OS, Porter RB (2006) Individual equity return data from Thomson Datastream: handle with care! *J Financ Res* 29(4):463–479
- Israel R, Moskowitz TJ (2013) The role of shorting, firm size, and time on market anomalies. *J Financ Econ* 108(2):275–301
- Jacobs H, Müller S (2020) Anomalies across the globe: Once public, no longer existent? *J Financ Econ* 135(1):213–230
- Jegadeesh N, Titman S (1993) Returns to buying winners and selling losers: implications for stock market efficiency. *J Financ* 48(1):65–91
- Karolyi GA (2016) Home bias, an academic puzzle. *Rev Financ* 20(6):2049–2078
- Karolyi GA, Lee KH, van Dijk MA (2012) Understanding commonality in liquidity around the world. *J Financ Econ* 105(1):82–112
- Kelly BT, Pruitt S, Su Y (2019) Characteristics are covariances: a unified model of risk and return. *J Financ Econ* 134(3):501–524
- Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. In: conference proceeding
- Koijen RSJ, van Nieuwerburgh S (2011) Predictability of returns and cash flows. *Annu Rev Financ Econ* 3(1):467–491
- Kyriakou I, Mousavi P, Nielsen JP, Scholz M (2021) Forecasting benchmarks of long-term stock returns via machine learning. *Ann Oper Res* 297(1):221–240
- Lalwani V, Meshram VV (2022) The cross-section of Indian stock returns: evidence using machine learning. *Appl Econ* 54(16):1814–1828
- Leippold M, Wang Q, Zhou W (2022) Machine learning in the Chinese stock market. *J Financ Econ* 145(2):64–82
- Leitch G, Tanner JE (1991) Economic forecast evaluation: profits versus the conventional error measures. *Am Econ Rev* 81(3):580–590
- Lesmond DA, Ogden JP, Trzcinka CA (1999) A new estimate of transaction costs. *Rev Financ Stud* 12(5):1113–1141
- Lewellen J (2015) The cross-section of expected stock returns. *Crit Financ Rev* 4(1):1–44
- Liu Q, Tao Z, Tse Y, Wang C (2022) Stock market prediction with deep learning: the case of China. *Financ Res Lett* 46(102):209
- Livnat J, Zarowin P (1990) The incremental information content of cash-flow components. *J Account Econ* 13(1):25–46
- Lu CJ, Lee TS, Chiu CC (2009) Financial time series forecasting using independent component analysis and support vector regression. *Decis Support Syst* 47(2):115–125
- Mammen E, Nielsen JP, Scholz M, Sperlich S (2019) Conditional variance forecasts for long-term stock returns. *Risks* 7(4):113
- McLean RD, Pontiff J (2016) Does academic research destroy stock return predictability? *J Financ* 71(1):5–32
- Nelson CR (1976) Inflation and rates of return on common stocks. *J Financ* 31(2):471–483
- Novy-Marx R (2013) The other side of value: the gross profitability premium. *J Financ Econ* 108(1):1–28
- Pástor L, Stambaugh RF (2009) Predictive systems: living with imperfect predictors. *J Financ* 64(4):1583–1628
- Patton AJ, Timmermann A (2010) Monotonicity in asset returns: new tests with applications to the term structure, the CAPM, and portfolio sorts. *J Financ Econ* 98(3):605–625
- Probst P, Bischl B, Boulesteix AL (2019) Tunability: importance of hyperparameters of machine learning algorithms. *J Mach Learn Res* 20(53):1–32
- Rapach DE, Zhou G (2013) Forecasting stock returns. In: Timmermann A, Elliott G (eds) *Handbook of economic forecasting*, vol 2, part A, *Handbooks in economics*, vol 2. Elsevier North-Holland, Amsterdam and Boston, pp 328–383
- Rapach DE, Zhou G (2020) Time-series and cross-sectional stock return forecasting: New machine learning methods. In: Jurczenko E (ed) *Machine learning for asset management: New developments and financial applications*. Iste and Wiley, London and Hoboken, pp 1–33
- Rasekhschaffe KC, Jones RC (2019) Machine learning for stock selection. *Financ Anal J* 75(3):70–88
- Rosenberg B, Reid K, Lanstein R (1985) Persuasive evidence of market inefficiency. *J Portf Manag* 11(3):9–16

- Rubesam A (2022) Machine learning portfolios with equal risk contributions: evidence from the Brazilian market. *Emerg Mark Rev* 51(100):891
- Sharpe WF (1964) Capital asset prices: a theory of market equilibrium under conditions of risk. *J Financ* 19(3):425–442
- Timmermann A (2006) Forecast combinations. In: Timmermann A, Granger CWJ, Elliott G (eds) *Handbook of economic forecasting, Handbooks in economics*, vol 1. Elsevier North-Holland, Amsterdam and Boston, pp 135–196
- Titman S, Wei KCJ, Xie F (2004) Capital investments and stock returns. *J Financ Quant Anal* 39(4):677–700
- Tobek O, Hronec M (2021) Does it pay to follow anomalies research? Machine learning approach with international evidence. *J Financ Mark* 56(100):588
- van Dijk MA (2011) Is size dead? A review of the size effect in equity returns. *J Bank Financ* 35(12):3263–3274
- Vapnik VN (1995) *The nature of statistical learning theory*. Springer, New York
- Wang KQ, Xu J (2015) Market volatility and momentum. *J Empir Financ* 30:79–91
- Welch I, Goyal A (2008) A comprehensive look at the empirical performance of equity premium prediction. *Rev Financ Stud* 21(4):1455–1508
- Winkler RL, Makridakis S (1983) The combination of forecasts. *J R Stat Soc: Ser A (General)* 146(2):150–157
- Wong FS (1991) Time series forecasting using backpropagation neural networks. *Neurocomputing* 2(4):147–159
- Woodhouse K, Mather P, Ranasinghe D (2017) Externally reported performance measures and benchmarks in Australia. *Account Financ* 57(3):879–905
- Zaremba A, Umutlu M, Maydybura A (2020) Where have the profits gone? Market efficiency and the disappearing equity anomalies in country and industry returns. *J Bank Financ* 121(105):966
- Zhang G, Patuwo BE, Hu MY (1998) Forecasting with artificial neural networks. *Int J Forecast* 14(1):35–62

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.