Journal of Asset Management (2023) 24:419–441
https://doi.org/10.1057/s41260-023-00318-z

ORIGINAL ARTICLE

Stock market anomalies and machine learning across the globe

Vitor Azevedo[2] · **Georg Sebastian Kaiser**[3] · **Sebastian Mueller**[1]

Revised: 27 June 2023 / Accepted: 28 June 2023 / Published online: 20 July 2023

Abstract
We identify the characteristics and specifications that drive the out-of-sample performance of machine-learning models across an international data sample of nearly 1.9 billion stock-month-anomaly observations from 1980 to 2019. We demonstrate significant monthly value-weighted (long-short) returns of around 1.8–2.2%, and a vast majority of tested models outperform a linear combination of predictors (our baseline factor benchmark) by a substantial margin. Composite predictors based on machine learning have long-short portfolio returns that remain significant even with transaction costs up to 300 basis points. By comparing 46 variations of machine-learning models, we find that the models with the highest return predictability apply a feed-forward neural network or composite predictors, with extending rolling windows, including elastic net as a feature reduction, and using percent ranked returns as a target. The results of our nonlinear models are significant across several classical asset pricing models and uncover market inefficiencies that challenge current asset pricing theories in international markets.

Keywords International stock market · Anomalies · Machines learning models · Market efficiency · Publication impact

JEL Classification G12 · G29 · M41

Introduction

In recent years, top finance journals have published more than 400 anomalies[1] and exponentially expand the factor zoo (Harvey and Liu 2019; Cochrane 2011) calling for different methods and higher hurdle rates (Harvey and Liu 2014; Harvey et al. 2016; Harvey 2017; Harvey and Liu 2019). Thereby, a noticeable tendency of a "(home) bias in academic research in finance" in the United States of America (U.S.) has been highlighted by Andrew Karolyi (2016) for published anomalies. This circumstance ignores potential regional differences, such as the post-publication profitability decline of 58% for numerous anomalies in the US stock market observed by McLean and Pontiff (2016) in contrast to a mostly insignificant post-publication decline in other countries reported by Jacobs and Müller (2020).

In addition to the predominant focus on the US stock universe in the field of asset pricing, there has also been a particular emphasis on linear models, such as ordinary least squares. However, linear models do not seem to be able to handle the multidimensionality of return predictors sufficiently (e.g., Azevedo and Hoegner 2023). Unlike linear models, machine-learning algorithms, with their increased complexity, potentially have enhanced capabilities in handling these issues of anomaly-based research.

Among the recent literature that applies machine learning in asset pricing, Gu et al. (2020) compare a variety of machine-learning methods in a US sample with 94 firms'

✉ Vitor Azevedo
vitor.azevedo@rptu.de

Georg Sebastian Kaiser
georgsebastian.kaiser@rolandberger.com

Sebastian Mueller
sebastian.mueller.hn@tum.de

[1] Technical University of Munich, TUM School of Management, Center for Digital Transformation, Campus Heilbronn, Am Bildungscampus 9, Heilbronn 74076, Germany

[2] RPTU Kaiserslautern-Landau, Department of Financial Management, Gottlieb-Daimler-Straße 42, Kaiserslautern 67663, Germany

[3] Roland Berger, Sederanger 1, Munich 80538, Germany

[1] We refer to anomalies as variables which explain the cross section of stock returns as identified in previous studies. Examples include the traditional (six-months) momentum effect (Jegadeesh and Titman 1993) or market value of equity (Banz 1981). The terms anomalies, predictors, signals, characteristics, and factors are used interchangeably in this paper.

characteristics and eight macroeconomic predictors, covering a period from 1957 to 2016. They find that trees and neural networks performed the best. Tobek and Hronec (2020) conduct a similar study with an international sample from 1963 to 2018 and find that machine learning is profitable around the globe and survives on a liquid universe of stocks. Chen et al. (2023) suggest a methodology that integrates four neural networks to leverage conditioning information and predict individual stock returns. Their research employs a dataset comprising 46 stock anomalies and 178 macroeconomic time series, spanning from 1967 to 2016, to estimate stock returns. The results show that their model yields a yearly Sharpe ratio of 2.6, which surpasses the linear special case of their model, which reports a Sharpe ratio of 1.7. More recently, Azevedo and Hoegner (2023) analyze the predictability of 299 capital market anomalies enhanced by 30 machine-learning approaches for the US market. They find that risk-adjusted returns of a machine-learning-based investment strategy are significant across alternative asset pricing models. The results are robust considering transaction costs with round-trip costs of up to 2% and including only anomalies after publication.

While evidence suggests that machine learning can be used to enhance the predictability power of anomalies in international markets (e.g., Tobek and Hronec 2020; Drobetz and Otto 2021; Cakici et al. 2022; Hanauer and Kalsbach 2022; Leippold et al. 2022; Breitung 2023; Fieberg et al. 2023), previous studies do not identify the characteristics and specifications (e.g., machine-learning algorithms, target values, rolling windows, and features reduction) that make machine-learning models successful at predicting stock returns. Furthermore, with exception of the contemporaneous study by Cakici et al. (2022)[2] there is to our knowledge no other comprehensive overview of the potential return predictability of a wide variety of machine-learning models for a representative global sample.

To examine the characteristics that make machine learning successful at predicting stock returns, our study examines nonlinear relationships among anomalies across the international stock universe, applying a broad range of different machine-learning algorithms and parameters. In contrast to the data sample of Tobek and Hronec (2020) consisting of 153 unique signals, we incorporate 240 individual anomalies for our machine-learning models to avoid omitting essential factors. Even in the first year 2003 of the post-publication scenario, our dataset already contains 90 published anomalies compared to 55 in Tobek and Hronec (2020). In addition, we evaluate a comprehensive international data sample with up to 38,001 firms per month compared to 4058 stocks in the case of Tobek and Hronec (2020). While they primarily focus on 23 developed countries in the regions USA, Japan, Asia Pacific, and Europe, we take a more holistic view, including Emerging and Frontier Markets. Lastly, we do not exclude micro-caps as machine learning is insensitive to outliers (Anand et al. 2019), and we are interested in the impact these particular stocks can have on the predictability of the machine-learning models. Hence, we ensure a vast and multifaceted data sample as the foundation of our analysis within the entire international universe. Nevertheless, we test if the results are robust for economically important stocks.

This versatile foundation embodies complex nonlinear relationships among signals which can be exploited by our models. We largely follow previous studies (e.g., Azevedo and Hoegner 2023; Gu et al. 2020) in which a variety of machine-learning algorithms are applied to develop profitable trading strategies in long-short portfolios. With a larger number of algorithms compared to Tobek and Hronec (2020) and Gu et al. (2020), our set of applied models comprises one Generalized Linear Model (GLM), two trees-based approaches [e.g., Distributed Random Forest (DRF) and Gradient Boosting Machine (GBM)], and two Feedforward Neural Network (FNN) models with both a wide and a narrow architecture. These models are analyzed twice by training them with two target values, namely, the raw returns and the percent-ranked returns as input. Supplementary, we include a variant of a recurrent neural network (RNN) with a long short-term memory (LSTM), typically very suitable for time series. Surprisingly, the FNN models outperform the RNN in both raw-return and percent-ranked portfolios settings with the highest monthly value-weighted return on average of 2.24% within a percent-ranked long-short portfolio for the larger FNN.

As a major difference compared to Gu et al. (2020), we do not support their finding for neural networks in finance “that ‘shallow’ learning outperforms ‘deep’ learning" (Gu et al. 2020, p. 2269). As our results reflect mixed outcomes, we find evidence that the superiority of a FNN rather depends on the target values chosen than on the learning architecture. Our “deep” FNN with five hidden layers (99,021 parameters) seems to outperform with scaled values (i.e., with percent-ranked returns), while our “shallow" large FNN with three hidden layers (251,759 parameters) performs better in comparison with the “deep" FNN trained on raw-returns. Noteworthy, Gu et al. (2020) analyze only 94 anomalies in the US stock market. Since they focus on a substantially lower fraction of the existing factor zoo in the USA, this might be a plausible explanation for their significantly lower Sharpe ratios of up to 1.35 for their neural network forecast compared to the Sharpe ratios for our neural networks ranging between 1.87 and 2.48 for the global market.

[2] We posted our first draft on April, 23, 2022 at SSRN, while the study of Cakici et al. (2022) was posted about 2 months later on June, 28, 2022 at SSRN.

Furthermore, this study is accompanied by several additional supporting analyses expanding the research procedure of comparable meta-studies like Tobek and Hronec (2020) and Gu et al. (2020). Firstly, to reduce the influence of unnecessary noise due to correlated anomalies, we simultaneously preprocess our data with a range of feature selection methodologies based on significance levels, unsupervised machine-learning models such as regularization approaches like least absolute shrinkage and selection operator (lasso) regression and elastic net selections. Then, combined with three rolling window training techniques, we complete the set of tested models while improving the performance with some applied techniques. In total, the monthly return on average for a single model can reach up to 2.71% in the case of the percent-ranked FNN with an extending learning window. Finally, we enrich our study by analyzing round-trip costs, the upper limit for transaction costs, in which trading strategies remain significant at the 0.05 level. The round-trip costs estimation of up to 328 basis points is another demonstrative indicator of robust results which are neither traceable to data snooping nor transaction costs.

We also propose a combination of all machine-learning models. As a result, we observe significant monthly returns on average for the composite predictors, ranging between 1.85% and 2.60%, with *t*-statistics largely above the critical value of 3 proposed by Harvey et al. (2016). Consequently, these results further strengthen the improbability that significant outperformance of the models is justified by *p*-hacking.

We then identify the characteristics and specifications that make machine-learning models successful at predicting returns. We measure the long-short portfolio returns of portfolios formed on predictions from 46 machine-learning models. We find that a combination of machine-learning models performs at least as well as any single model. Among the single models, the highest returns are achieved with FNN models. Extending and 10-year rolling windows are the window training models with the highest return performance. In terms of target, we find that percent ranked returns outperform raw returns. Finally, elastic net reports the highest average return among feature reduction methods. Despite the superiority of machine-learning models over linear models, we find that specifications can play a major role in return predictability. The difference between the long-short returns of the machine-learning models with the best and worst return predictability is 171 basis points per month.

While we are mindful of reducing data snooping risk, the question is to what extent current factor models explain the return of these models. Testing our findings against eight distinct factor models, such as the two Fama–French factor models (Fama and French 1993, 2015), we find significant alphas for all tested machine-learning models. These results challenge the Efficient Market Hypothesis (EMH) in the international stock universe. Especially as our STATEW models enjoy alpha figures ranging between 1.10% and 2.64% with *t*-statistic values far above the minimum significance hurdle rate for new factors of 3.00.

Overall, we contribute to the existing literature mainly in the following three aspects. First, we highlight the tremendous potential of machine-learning algorithms for investors seeking profitable trading strategies and for scholars to understand (international) asset pricing in more detail. We offer a wide variety of 40 applied machine learning and six combinations of models using distinct algorithms, different feature reduction methods, and static and rolling training techniques. We quantify significant outperformance almost universally over single anomalies and our linear combined baseline factor benchmark. Thus, this paper extends the broad analysis of machine-learning models in the US market by Gu et al. (2020) and Azevedo and Hoegner (2023) by adding international evidence as well as identifying the characteristics and specifications that drive the out-of-sample performance of machine-learning models.

Second, we extend the literature by comparing models with different features and parameters. By doing so, it is possible to assess the impact on the predictability power of the models by changing the target value, the machine-learning algorithms, the window training, and the feature reduction. In particular, the elastic net can outperform the full feature, which is evidence that for international markets, some predictors might add some noise to the model, and feature reduction can be a solution for dealing with the multidimensionality of international data. Furthermore, we find that training the machine-learning models based on percent-rank returns shows superior results over the most common approach, which is based on raw returns (e.g., Gu et al. 2020).

Third, we alleviate the data dredging risk of comparing a single machine-learning model by combining our entire set of tested machine-learning models into several overarching composite predictors. The approach of combining multiple forecasts is associated with enhanced forecast accuracy as widely proven by the statistical research (Clemen 1989; Bates and Granger 1969; Makridakis and Hibon 2000; Timmermann 2006). Inspired by Rasekhschaffe and Jones (2019), we scale the combination concept to multiple composite predictors with international evidence containing our entire set of applied model variations.

Our study is structured as follows. In section "Data and methodology", we describe the origin of our international sample and the underlying methodology of our paper in detail. Subsequently, we prepare the results of our empirical study in a twofold fashion. Starting with section "Performance evaluation of individual anomalies and the baseline factor", we discuss the performance of the individual anomalies and combine the full feature base into one overarching

baseline factor. The purpose of this baseline factor is to serve as a linear benchmark for our complex machine-learning models in section "Portfolio construction with machine learning algorithms in a static window". Here, we construct future return predictors using various training and preprocessing methodologies. Then, we compare and interpret the performance of our tested machine-learning models in section "Comparison and robustness tests of machine learning models". We discuss the findings concerning feature importance, transaction costs, and results against traditional factor models. We further combine our set of tested machine-learning models into several composite predictors. Lastly, we conclude in section "Conclusion".

Data and methodology

Our methodology consists of two phases. Within the first phase, we assess the performance of the individual anomalies (i.e., cross-sectional stock return predictors) with a classical portfolio-sort analysis. Then, we combine the single performance of our anomalies into one overarching signal, the baseline factor, intended to serve as a linear benchmark. We compare this linear benchmark with different nonlinear machine-learning models on their predictive power and additional profits in the second phase. These machine-learning models are built on the international anomaly dataset and encompass several distinct algorithms, feature reduction techniques, rolling windows modifications, and training concepts discussed in the following.

Data, preprocessing, and anomaly calculation

For the performance assessment of the individual signals and the machine-learning models, we use an updated international anomaly data sample comparable to the dataset in Jacobs and Müller (2018), which includes 240 distinct anomalies taken from Green et al. (2017); McLean and Pontiff (2016); Hou et al. (2015), and Harvey et al. (2016). Our dataset rests on three Thomson Reuters databases: Datastream supplies stock returns and other stock-related figures (e.g., unadjusted prices), Worldscope is the data source for accounting figures, and IBES provides analyst data, including recommendations and earnings forecasts.

Our sample period ranges from July 1980, which marks the first year with the availability of accounting data, to June 2019. We download stock data for all countries which belong to one of the major MSCI regional indices as of June 2019 (i.e., MSCI North America, Europe, Pacific, Emerging Markets, or Frontier Markets). By relying on the MSCI classification, we ensure that our sample includes only countries with economically important and sufficiently liquid equity markets. At the same time, the selected countries are still very heterogeneous in terms of their size and financial market development, providing us with a representative sample of global equity markets.

We implement several supplementary filters in Datastream to ensure that our dataset exclusively comprises common equity. Specifically, (i) we select only the primary share class when multiple securities exist for a company (Schmidt et al. 2019), (ii) we ensure the security-type equity (Ince and Porter 2006), (iii) we acquire solely the principal quotations for a security in instances of multiple exchange listings (Fong et al. 2017), and (iv) we include only stocks that Datastream links to one of the countries in our study (Ince and Porter 2006). Lastly, (v) to further eliminate non-common equity securities from our sample, we require all stocks to possess a non-missing Worldscope identifier.

These filtering steps result in a final sample of 9.39 million stock-month observations from more than 66,000 different firms across 68 different countries. Table 1 presents summary statistics at the country level. Around 80% of the total stock sample is from non-US stock markets, which also account for approximately 65% of the average total stock market capitalization. On average, 20,071 stocks are included in our dataset per month in a given year. The number is increasing over time, partly because stock data is not available for all countries at the beginning of our sample period. Table 1 also shows the starting dates per country.

We calculate monthly stock returns in US-Dollar using Datastream's total return index (code: RI), which includes dividends. Because there are few outliers in the return data, we winsorize returns at the 0.1% and 99.9% level, respectively. Further, we use the methodology of Ince and Porter (2006) to include delisted stocks in our analysis only up to the point of their actual delisting. To calculate the 240 cross-sectional return predictors for our sample, we follow the instructions provided in the original paper of Jacobs and Müller (2018). We list all anomalies together with their reference study in Table A.2 of the Internet appendix. For more details on the gathering, filtering, and calculation process of the anomalies, we refer to Jacobs and Müller (2018).

Due to missing values, we are left on average with 201 out of 240 anomalies for each stock-month observation. Furthermore, we categorize the anomaly set into 113 anomalies based on fundamentals, 75 market-based signals, 18 analyst-based anomalies, 19 valuation-based signals, and 15 other signals. The number of anomalies is comparable to other anomaly studies within current literature, such as Hou et al. (2015) assessing 447 anomalies, Harvey et al. (2016) analyzing 315 anomalies, Azevedo and Hoegner (2023) calculating with 299 signals, McLean and Pontiff (2016) analyzing 97 signals, and Green et al. (2017) evaluating 94 anomalies.

Table 1 Descriptive statistics

Country	Region	Start date	Number of stocks		Number of non-microcaps		Total market value	
			Monthly average	Percentage of total (%)	Monthly average	Percentage of total (%)	Monthly average	Percentage of total (%)
Brazil	Emerging markets	6/1994	128	0.55	57	0.90	325,042	1.12
Chile	Emerging markets	8/1989	132	0.57	44	0.70	122,800	0.42
China	Emerging markets	2/1991	1434	6.18	741	11.76	2,375,123	8.17
Colombia	Emerging markets	2/1992	39	0.17	16	0.25	65,959	0.23
Czech Republic	Emerging markets	6/1996	29	0.12	6	0.10	28,715	0.10
Egypt	Emerging markets	12/1997	118	0.51	18	0.29	42,029	0.14
Greece	Emerging markets	2/1988	185	0.80	28	0.44	61,346	0.21
Hungary	Emerging markets	6/1993	32	0.14	5	0.08	19,047	0.07
India	Emerging markets	2/1990	1294	5.57	154	2.44	689,468	2.37
Indonesia	Emerging markets	8/1987	270	1.16	46	0.73	157,654	0.54
Korea	Emerging markets	8/1984	868	3.74	116	1.84	492,302	1.69
Malaysia	Emerging markets	2/1981	512	2.21	76	1.21	181,245	0.62
Mexico	Emerging Markets	2/1988	93	0.40	46	0.73	184,427	0.63
Peru	Emerging markets	2/1992	85	0.37	15	0.24	38,813	0.13
Philippines	Emerging markets	6/1989	167	0.72	32	0.51	91,537	0.32
Poland	Emerging markets	6/1992	255	1.10	27	0.43	89,387	0.31
Qatar	Emerging markets	6/2004	38	0.16	23	0.37	114,922	0.40
Russia	Emerging markets	6/1997	202	0.87	56	0.89	494,198	1.70
South Africa	Emerging markets	8/1980	216	0.93	70	1.11	215,811	0.74
Taiwan	Emerging markets	12/1988	917	3.95	137	2.17	474,552	1.63
Thailand	Emerging markets	6/1988	381	1.64	57	0.90	171,070	0.59
Turkey	Emerging markets	6/1988	195	0.84	37	0.59	112,659	0.39
Austria	Europe	6/1981	66	0.28	26	0.41	58,828	0.20
Belgium	Europe	7/1980	94	0.40	34	0.54	157,097	0.54
Denmark	Europe	10/1980	132	0.57	31	0.49	111,135	0.38
Finland	Europe	2/1987	103	0.44	34	0.54	157,725	0.54
France	Europe	8/1980	563	2.43	170	2.70	1,071,857	3.69
Germany	Europe	12/1980	447	1.93	113	1.79	814,853	2.80
Ireland	Europe	7/1980	47	0.20	16	0.25	50,492	0.17
Italy	Europe	6/1981	201	0.87	90	1.43	384,235	1.32
Netherlands	Europe	3/1981	120	0.52	54	0.86	274,226	0.94
Norway	Europe	6/1981	137	0.59	37	0.59	122,277	0.42
Portugal	Europe	2/1988	56	0.24	17	0.27	49,128	0.17
Spain	Europe	2/1986	133	0.57	71	1.13	422,925	1.46
Sweden	Europe	2/1982	258	1.11	60	0.95	245,116	0.84
Switzerland	Europe	8/1980	175	0.75	79	1.25	623,456	2.15
United Kingdom	Europe	7/1980	1244	5.36	352	5.59	1,752,611	6.03
Argentina	Frontier markets	2/1988	53	0.23	13	0.21	30,054	0.10
Bahrain	Frontier markets	6/2004	32	0.14	6	0.10	14,256	0.05
Bangladesh	Frontier markets	9/2005	72	0.31	6	0.10	20,253	0.07
Bulgaria	Frontier markets	4/2006	189	0.81	2	0.03	6619	0.02
Croatia	Frontier markets	6/2006	93	0.40	5	0.08	21,709	0.07
Estonia	Frontier markets	6/2004	13	0.06	1	0.02	2587	0.01
Jordan	Frontier markets	11/2005	178	0.77	5	0.08	22,589	0.08
Kazakhstan	Frontier markets	5/2009	27	0.12	5	0.08	14,884	0.05
Kenya	Frontier markets	6/2001	37	0.16	5	0.08	11,668	0.04
Kuwait	Frontier markets	6/2004	153	0.66	27	0.43	96,589	0.33

Table 1 (continued)

Country	Region	Start date	Number of stocks		Number of non-microcaps		Total market value	
			Monthly average	Percentage of total (%)	Monthly average	Percentage of total (%)	Monthly average	Percentage of total (%)
Lebanon	Frontier markets	6/2006	9	0.04	4	0.06	6703	0.02
Lithuania	Frontier markets	6/2003	25	0.11	2	0.03	3594	0.01
Mauritius	Frontier markets	12/2005	50	0.22	3	0.05	6997	0.02
Morocco	Frontier markets	6/1998	50	0.22	16	0.25	41,798	0.14
Nigeria	Frontier markets	9/2009	110	0.47	15	0.24	45,822	0.16
Oman	Frontier markets	11/2005	102	0.44	8	0.13	18,772	0.06
Pakistan	Frontier markets	3/1991	155	0.67	12	0.19	29,623	0.10
Romania	Frontier markets	6/2006	121	0.52	7	0.11	21,262	0.07
Serbia	Frontier markets	7/2006	57	0.25	1	0.02	3342	0.01
Slovenia	Frontier markets	6/2003	24	0.10	2	0.03	3694	0.01
Sri Lanka	Frontier markets	9/1993	127	0.55	2	0.03	7633	0.03
Tunisia	Frontier markets	1/2006	55	0.24	3	0.05	7632	0.03
Ukraine	Frontier markets	4/2006	52	0.22	8	0.13	46,041	0.16
Vietnam	Frontier markets	1/2007	641	2.76	15	0.24	57,891	0.20
Canada	North America	7/1980	1195	5.15	198	3.14	804,443	2.77
USA	North America	8/1980	3773	16.25	1646	26.13	10,089,937	34.73
Australia	Pacific	7/1980	828	3.57	126	2.00	545,718	1.88
Hong Kong	Pacific	9/1980	661	2.85	162	2.57	890,917	3.07
Japan	Pacific	7/1980	2533	10.91	919	14.59	3,069,582	10.56
New Zealand	Pacific	2/1986	81	0.35	16	0.25	33,586	0.12
Singapore	Pacific	9/1980	352	1.52	68	1.08	235,758	0.81

This table reports summary statistics for the 68 countries included in our sample. The sample period ranges from 7/1980 to 6/2019, but for some (particularly emerging) countries, coverage in Refinitiv Datastream starts later. The next columns show the monthly average and the percentage of the total average number of firms per month, the number of non-microcaps (stocks in NYSE size decile larger than 2), and the total market value (in billion US dollars)

Finally, we use percent-ranked signal values instead of raw signal values as this preprocessing scaling procedure provides an effective and simple solution to deal with outliers and data errors and, therefore, might increase the performance of the linear baseline factor and the nonlinear machine-learning models. Like Jacobs (2016), we first rank stocks as underpriced and overpriced according to each predictor. Ranks are standardized in intervals from (0,1) in each country-month observation. Among the advantages of this procedure, it allows us to fill in missing values with a median of 0.5 without any forward-looking bias.

Portfolio-sort strategy and baseline factor construction

We test our signals with a portfolio-sort strategy for each month-anomaly by assessing statistical significance and performance in terms of signal profitability. Following the approach of Chen and Zimmermann (2022), we assign stocks with the best (worst) performing signals within each country into long (short) positions. We then create a decile (10) minus decile (1) long-short portfolio with a monthly return defined as the spread of these long-short positions.

In order to ensure performance comparability, we use a standardized portfolio-sort methodology with no further stock filtering like minimum price filtering, excluding micro-cap stocks, or adapting to different rebalancing and holding periods. Additionally, a standardized approach diminishes the risk of a limited selection of filters or parameters to boost research results (i.e., *p*-hacking).

We calculate equally-weighted and value-weighted long-short portfolios with a standardized signal calculation approach. However, according to the bad model problem stated in Fama (1998), an equally-weighted analysis suffers from the potential overweighting of micro-cap stocks, while a value-weighted portfolio is more influenced by stocks with a large market capitalization. Therefore, we focus on the analysis of value-weighted portfolios for our machine-learning models, which offer an intuitive interpretation of results and a foundation for potential investment decisions.

We simultaneously report the number of stock holdings in the long and short legs of the portfolios. In the context of these long and short positions, we determine the one-sided

turnover rate defined as the relative amount of shares required for the monthly rebalancing of the portfolio. We utilize this one-sided turnover rate in section "Turnover rate and acceptable transaction costs estimation" to calculate round-trip costs.

Similar to the computation of individual signals, we combine our available set of anomalies into a linear baseline factor assessing it on the same criteria. For the linear combination, we build an arithmetic average of our 240 percent-ranked anomalies for each stock-month observation by each country. We include only stock-month observations with at least 100 distinct signals to ensure the diverse foundation of our baseline factor. This baseline factor serves as a benchmark for our machine-learning models in section "Portfolio construction with machine learning algorithms in a static window".

Sample split and cross-validation of machine-learning models

As a typical prerequisite for any supervised machine-learning model, we divide our data sample into a training sample comprising the signal data from July 1980 to June 2003 and a test sample in the following periods. To increase the robustness of our training sample, we implement a threefold cross-validation.[3] After training and testing our machine-learning models, we determine model performance with a long-short portfolio-sort strategy. In general, we monitor our machine-learning models to predict the next month's returns of stocks based on the described training sample. We then use these predictions as a decision factor for our classification in long-short portfolios. The following subsections briefly summarize the machine-learning algorithms and the applied training, validation, and test mechanisms.

Machine-learning algorithms

In recent years, there has been a growing interest in finance-related machine-learning issues. Various machine-learning models have been tested with the result that tree-based models and neural networks seem to be among the most promising algorithms in finance (Gu et al. 2020). Following recent literature, we, therefore, investigate the performance of the two tree-based models GBM and DRF and compare them to a regression-based GLM. For this purpose, we use the widely spread open-source library for machine learning H2O.ai (2020). In addition to these three models, we include two FNN with different architectures and one RNN as further models in our analysis. To deploy these three neural networks, we apply the Tensorflow (2020) framework from Google DeepMind. In the following, we briefly specify the applied algorithms. We refer to the original documentation and source code for a more comprehensive description of all six machine-learning algorithms.

Nelder and Wedderburn (1972) were among the first to establish GLM as a flexible generalization of multiple regression types, including linear regression, logistic regression, and Poisson regression. GLM can be modified with different distributions and link functions. For our GLM, we apply the default identity link combined with a normal distribution for both percent-ranked and raw-returns target values. In contrast to GLM, GBM and DRF are tree-based machine-learning techniques (Hastie et al. 2009). Based on the algorithm developed by Breiman (2001), the bootstrap aggregation, called bagging (Breiman 1996), is used for the DRF. It involves randomly selecting data points for multiple decision trees and combining them by averaging all decision trees to enhance the robustness and accuracy of the machine-learning model. The GBM algorithm is based on other weak prediction models called learners, such as a decision tree. Based on these weak learners, the GBM algorithm builds multiple decision trees, weighting the predictive power of individual learners according to their performance and reducing the amount of misclassified data by using the multiple learner approach (Zhou 2012). Our GBM follows the implementation of Hastie et al. (2009) and the H2O.ai (2020) library documentation.

In addition to the three standard machine-learning models described, we also investigate the performance of neural networks. It is important to note that neural networks differ distinctly from tree-based or regression-based models. In contrast, these neural networks are constructed from a series of neuron layers that aim to simulate the mechanism of the human brain. Among various neural network algorithms developed since the 1950s, we concentrate on three distinct neural network models. Two models are based on the classical architecture of a FNN. The two models differ in the width of the neuron architecture. The smaller structure consists of five hidden layers with a decreasing number of neurons while graphically resembling a tunnel. The broader structure has three hidden layers and a more significant number of neurons per layer. The third assessed neural network model is the RNN. Previous work has shown that the RNN is particularly suitable for analyzing data along a time period, as these models build up a type of short-term memory to increase performance. This approach might be particularly advantageous for predicting future stock returns.

[3] A threefold cross-validation involves training the model in rotation with two-thirds of the in-sample data (until June 2003). The remaining one-third of the in-sample data is used for validation while performing hyperparameter optimization.

Performance evaluation of individual anomalies and the baseline factor

Performance evaluation of individual anomalies

By analyzing equally-weighted long-short portfolios based on 240 individual anomalies in our international out-of-sample data, we find an average return of 0.35% per month with a *t*-statistic of 3.10. At the minimum *t*-statistic hurdle rate of 1.96, 167 signals show significant returns accounting for 70% of total available anomalies. With an increased minimum absolute *t*-statistic of 3.00, we still list 132 significant signals throughout our international sample, representing 55% of all assessed signals.

Not surprisingly, the number of significant anomalies decreases for value-weighted portfolios because stocks with a larger market capitalization have a greater performance influence in this case. In exchange, the considerable influence of micro-cap stocks is reduced. Our analysis shows an average monthly return of 0.25% with a mean *t*-statistic of 1.47. 41% (20%) of all signals, namely 98 (49) signals, surpass the minimum *t*-statistic hurdle of 1.96 (3.00). Table A.2 in the Internet appendix presents the value-weighted returns for each anomaly.

Baseline factor as linear multi-anomaly combination benchmark

Relying on these 240 individual anomalies, we calculate the baseline factor based on percent-ranked values of all single anomalies at a stock level for each month. Previous studies find that by combining anomalies, additional profit opportunities might arise, or hidden structures might be discovered (e.g., Stambaugh et al. 2015; Green et al. 2017). Moreover, with this procedure, we alleviate the data dredging concerns associated with individual anomalies and, thus, strengthen the robustness of our anomaly research.

For our main analysis, we use a more restrictive and reliable setup with mid-2003 as a breaking point for our baseline factor analysis. In previous meta-studies, for instance, in Green et al. (2017) and Jacobs and Müller (2018), the year 2003 stands for an essential breaking point in the performance of signals. In 2003, the auditing and reporting quality in the USA significantly increased due to the ratification of the Sarbanes-Oxley Act and the new SEC filing changed (Green et al. 2017). Additionally, following the strategy of increased global standardization of reporting, the European Union accepted in 2002 the International Financial Reporting Standards as the new mandatory reporting standards for listed EU companies starting in 2005. Due to these meaningful adaptations of reporting standards, the stock data quality has arguably increased. In addition, the number of newly published anomalies has substantially risen since 2003. As McLean and Pontiff (2016) first reported, the performance of signals in terms of return becomes significantly lower after the publication of the hidden patterns. Our out-of-sample data mostly consist of anomalies found in the 2000s, largely comparable to the data sample in Azevedo and Hoegner (2023), in which 2003 marks the mean publication year of 299 individual assessed anomalies. Therefore, we initially take July 2003 as a breaking point for our baseline factor analysis and the creation of our training and testing set required by our machine-learning models in section "Portfolio construction with machine learning algorithms in a static window".

Comparing the full sample period from August 1980 to June 2019 and the period starting in August 2003 of an equally-weighted baseline factor shows similar average monthly returns and mean *t*-statistics. As can be seen in Table 2, the mean long-short return of the baseline factor increases from 2.01 to 2.20% with a stock data sample starting in August 2003, while the *t*-statistic decreases from 13.25 to 12.79. In contrast to the mean of the 240 single anomalies, the minimum *t*-statistic hurdle rate of 3.0 is surpassed by far for both time frames.

In contrast to the equally-weighted returns, we do not find an increase in average monthly returns with stock data from 2003 onwards in the case of a value-weighted baseline factor (1.36% vs. 1.02%). Focusing on the significance level of our value-weighted baseline factor, we also see a clear surpass of the minimum hurdle rate for new factors. Additionally, we report a comparable decline of the mean *t*-statistics between the two-time frames [7.48 vs. 4.93] in line with McLean and Pontiff (2016) and Azevedo and Hoegner (2023).

We observe a higher (two-sided) turnover rate for the baseline factor than for the average of our individual signals. Since the baseline factor composes the entire set of 240 anomalies, it is affected by several monthly stock ranking changes, resulting in more volatile long-short portfolios associated with higher turnover rates. We review these findings and quantify the impact on the performance of our models in section "Comparison and robustness tests of machine learning models", as higher turnover rates likely lead to higher transaction costs, which reduce the profitability for practitioners implementing our potential investment strategies.

To summarize our performance analysis, we report significant returns for our baseline factor, outperforming most of the individual signals in both equally-weighted and value-weighted portfolios. For our nonlinear machine-learning models in the subsequent chapters, we thus use the value-weighted baseline factor with a post-July 2003 sample as a linear benchmark. This benchmark generates a mean monthly return of 1.02% [*t*-statistic of 4.93], restricting ourselves to a more conservative approach.

Table 2 Performance comparison of the baseline factor

Strategy	Baseline factor			Mean of individual signals		
	Return	*t*-statistic	Turnover rate	Return	*t*-statistic	Turnover rate
Full sample period						
Original settings				0.35	3.10	47.13
Equally-weighted	2.01	13.25	65.52	0.34	3.11	46.85
Value-weighted	1.36	7.48	76.55	0.25	1.47	46.83
Until July 2003						
Original settings				0.35	2.04	47.33
Equally-weighted	1.87	8.24	66.60	0.33	2.04	47.44
Value-weighted	1.61	5.93	76.84	0.28	1.08	48.07
Post-July 2003						
Original settings				0.34	2.91	46.21
Equally-weighted	2.20	12.79	63.97	0.34	2.93	46.09
Value-weighted	1.02	4.93	76.05	0.22	1.08	45.24

This table shows the long-short value- and equally-weighted portfolio returns (in %) of the baseline factor and the average of individual anomalies. We report the results for the subsamples of August 1980 to July 2003 and August 2003 to June 2019, as well as the full sample (from 1980 to 2019). We also report the *t*-statistics and (two-sided) turnover rate

Portfolio construction with machine-learning algorithms in a static window

Machine-learning algorithms in a static window and the cross section of stock returns

After creating a linear benchmark with a significant outperforming return of the baseline factor, we explore nonlinear relationships among signals in the following chapter. We train regression-based models, tree-based approaches, and neural networks to detect hidden structures in our anomaly sample and try to exploit them profitably.

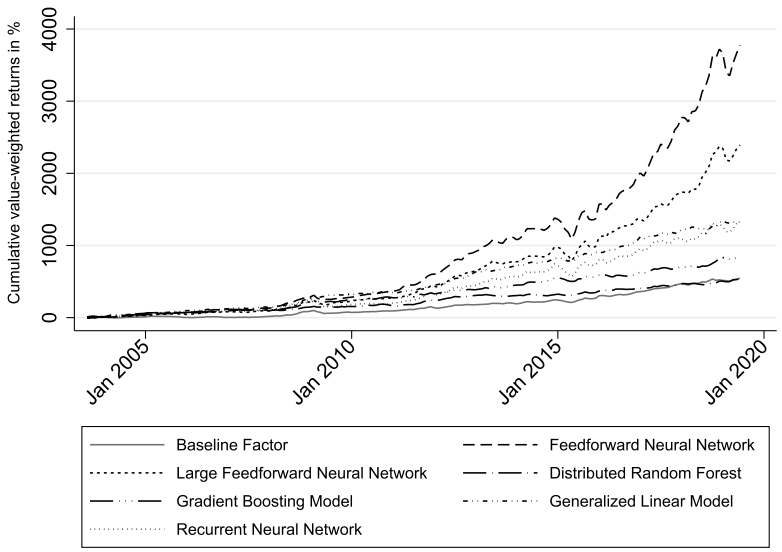
For the design of our regression and tree-based machine-learning models, given that performance-based optimization of parameters (called hyperparameter tuning) is more likely exposed to the risks of data dredging, we focus conservatively on the default H2O parameters. For our tree-based models DRF and GBM, apart from the choice of parameters, the depth definition of the tree structure influences the model performance. As recommended by Probst and Boulesteix (2017), we set up our models with 100 trees per model to achieve high performance while compensating for generalization requirements and computational limitations. For more details on these algorithms, we refer to section "Machine learning algorithms" and the H2O.ai (2020) documentation.

Besides regression and tree-based machine-learning models, artificial neural networks have become the most popular algorithms in many areas in recent years. We, therefore, expand our set of machine-learning models with three neural networks. Our first two approaches are FNN with different configurations. In general, any FNN consists of a variable number of hidden layers and neurons which are individually and directly linked. Our smaller FNN is constructed with 99,021 parameters distributed among five hidden layers with a decreasing number of neurons per layer (200, 150, 100, 50, and 10 neurons per layer). The larger FNN has a structure with 251,759 parameters distributed among three hidden layers with a constant number of 299 neurons. Inspired by previous successful meta-studies on time series, we finally enlarge our research with a RNN. As explained by Abiodun et al. (2018), RNN can enhance performance by a type of short-term memory. To implement this short-term memory, we include in each model prediction twelve so-called time steps consisting of last year's observations of all 240 anomalies.

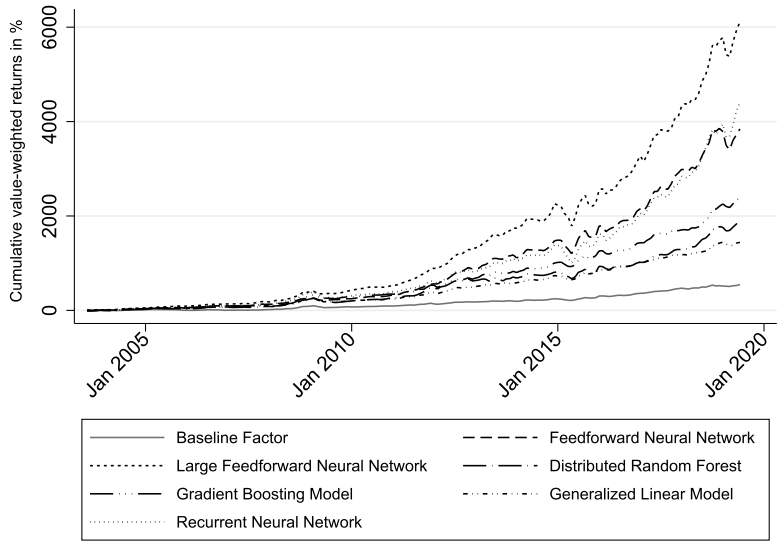
Moreover, to hold potentially common RNN back-propagated errors constantly, we upgrade our memory to a long short-term memory (LSTM) by building a long-term memory cell (Hochreiter and Schmidhuber 1997). In total, 240,449 parameters form the architecture of our RNN.

We use two variants of target values (dependent variables) to train the models. First, we train our models based on the raw next-month stock returns ($r_{t+1,i}$) (i.e., $f(anomalies_{t,i}) \rightarrow r_{t+1,i}$). The second analysis variant is based on a training method with percent-ranked next-month stock returns ($rp_{t+1,i}$) (i.e., $f(anomalies_{t,i}) \rightarrow rp_{t+1,i}$). This percent-ranking is conducted separately for every country-month. It places the returns in a data range from 0 to 1 with the advantage of having the same scaled pattern of target values as the signal values. This practice allows us to predict only the relative stock performance and distribution instead of the actual return of each stock with our portfolio-sort

(a) Cumulative value-weighted return of raw-return machine learning models



(b) Cumulative value-weighted return of percent-ranked machine learning models



The graphs illustrate the cumulative return of six machine learning algorithms compared to the benchmark, the baseline factor, during the post-July 2003 out-of-sample period. Figure (a) depicts the value-weighted return of raw-return machine learning models, while Figure (b) shows the value-weighted return of percent-ranked machine learning models.

Fig. 1 Cumulative returns of six machine-learning algorithms and the baseline factor

strategy. We expect both variants with different target values to generate a similar outcome regarding portfolio returns. Nevertheless, percent-rank returns might increase the accuracy of our predictions by being less prone to outliers.

Figure 1a shows that except for the DRF model, all other models trained with raw returns as a target clearly outperform the linear baseline factor in terms of the cumulative value-weighted return. Specifically, we find that all three neural networks exceed the performance of the regression and tree-based machine-learning models in the value-weighted setting. The best-performing model is the small FNN with a highly significant average monthly return of 1.99% for the post-July 2003 period, followed by the large FNN (1.75% average monthly return) and RNN (1.47%). Comparing the Sharpe ratios of our models in Panel A of Table 3, we detect the highest Sharpe ratio of 2.18 with our small FNN benchmarked against the baseline factor with a Sharpe ratio of 1.25. In the case of the regression and

Table 3 Portfolio metrics for machine-learning models

Panel A: Portfolio metrics for raw-return machine-learning models							
Model	Baseline	Generalized Linear Model (GLM)	Gradient Boosting Machine (GBM)	Distributed Random Forest (DRF)	Small Feedforward Neural Network (FNN)	Large Feedforward Neural Network (FNN)	Recurrent Neural Network (RNN)
Equally-weighted							
Post-July 2003 returns	2.2024	5.3920	4.2714	4.8811	4.3061	4.5159	3.1876
Post-July 2003 *t*-stat	12.7896	30.0620	29.0622	33.7009	37.2661	35.1695	18.8959
Average turnover rate	31.9856	63.0753	54.7579	64.5015	61.0663	64.1277	52.8750
Annualized return	29.4561	87.1991	64.8238	76.7838	65.6195	69.6049	45.2759
Sharpe ratio	3.5729	10.1548	9.2127	11.0735	11.8618	11.3229	5.6062
Value-weighted							
Post-July 2003 returns	1.0234	1.4308	1.2130	1.0001	1.9909	1.7505	1.4653
Post-July 2003 *t*-stat	4.9265	8.0298	6.3534	5.7989	8.0341	7.3917	5.6640
Average turnover rate	38.0247	69.6796	58.3013	72.9210	62.7168	65.3287	57.4003
Annualized return	12.4501	18.1770	15.1045	12.3121	25.8335	22.3894	18.1821
Sharpe ratio	1.2519	2.1308	1.6525	1.4912	2.1776	1.9747	1.4680
Panel B: Portfolio metrics for percent-ranked return machine-learning models							
Equally-weighted							
Post-July 2003 returns	2.2024	3.9653	4.6522	3.1022	4.4685	5.5636	4.1574
Post-July 2003 *t*-stat	12.7896	29.6059	33.7287	19.6252	35.6271	42.3406	27.7366
Average turnover rate	31.9856	63.4199	66.4727	62.8320	61.7949	63.3038	57.7469
Annualized return	29.4561	59.1624	72.2365	43.8971	68.6995	91.1649	62.6503
Sharpe ratio	3.5729	9.2265	10.9393	5.8007	11.4410	14.4918	8.7307
Value-weighted							
Post-July 2003 returns	1.0234	1.4887	1.7510	1.6642	2.0047	2.2436	2.0991
Post-July 2003 *t*-stat	4.9265	6.6748	6.8146	5.6271	7.7893	8.9897	6.9542
Average turnover rate	38.0247	69.0475	71.2350	66.1515	62.0930	67.0897	57.7593
Annualized return	12.4501	18.7459	22.2742	20.7385	25.9723	29.6202	27.0266
Sharpe ratio	1.2519	1.7556	1.8107	1.4647	2.1079	2.4791	1.8703

This table lists both model metrics and portfolio metrics for the training sample (e.g., cross-validation) and the test sample (e.g., out-of-sample) for all our H2O algorithms and the Baseline factor. We distinguish between equally-weighted and value-weighted portfolios. The target values of the models are trained on the absolute next-month return and the percent-ranked next-month return of a stock. Post-July 2003 performance and significance refer to average monthly returns. Post-July 2003 performance, Average (one-sided) turnover rate, and Annualized returns are given in %

tree-based machine-learning models, the GLM performs best with an average monthly return of 1.43% (post-July 2003 period).[4]

Compared to the raw-return variant, we see similar observations for the percent-ranked return specification. As illustrated in Fig. 1b, the three neural networks are among the three best-performing machine-learning models in terms of cumulative value-weighted return. More specifically, the large FNN returns a higher cumulative value in the post-July 2003 period than the RNN and the smaller FNN, with 2.24%, 2.10%, and 2.00% of average monthly returns, respectively. As we can see in Panel B of Table 3, the largest Sharpe ratio is 2.48 for the large FNN. Focusing on the regression and tree-based machine-learning models, we find that the GLM, with a less complex model architecture, marks the weakest machine model performance in the specification with percent-rank returns. With 1.49% of average monthly returns, it still surpasses the linear benchmark with a 1.02% monthly return on average. The best non-neural-network model is the GBM showing 1.75% of average monthly returns in the post-July 2003 period and a Sharpe ratio of 1.81. The minimum significance hurdle rate of new factors is surpassed by all models, with *t*-statistics between 5.63 and 8.99.

[4] Surprisingly, with an equally-weighted setting, this GLM beats all other models, including the neural networks, with a 5.56% average monthly return. However, we continue to focus on the value-weighted portfolios for the already-mentioned reasons.

To sum up, different nonlinear machine-learning models beat the linear baseline factor benchmark. Neural networks outperform other regression and tree-based machine-learning models. Furthermore, the minimum *t*-statistic hurdle rate of 3.00, recommended by Harvey et al. (2016), is easily surpassed by all machine-learning models.

Interpretation of the machine-learning models through relative feature importance

As the previous section described, nonlinear machine-learning models show significant performance. Related work investigates mostly the performance of single anomalies, and linear relationships among these signals. In contrast, our nonlinear machine-learning models have a larger complexity in size and the ability to observe hidden nonlinear relationships. As an illustration, we can consider the number of parameters of our small FNN, which already includes 99,021 parameters. With this larger complexity comes increased difficulty in interpreting the model results. In the literature, this issue is described as the black-box problem of Artificial Intelligence (AI) (Zednik 2021).

In order to address the black-box problem of machine learning, various approaches have been put forward. One possibility is the interpretation of the relative importance of features. As explained in the H2O.ai library documentation, the relative importance of tree-based models, like GBM, is defined as the improvement of the squared error due to the selection of this variable in the tree-building process. Likewise, the relative importance of regression-based models, such as GLM, is determined by the coefficient magnitudes. For our neural networks, we apply the permutation approach described by Breiman (2001) to compare the distribution of relative variable importance by our three reference models (raw-return GLM, percent-ranked GBM, and large percent-ranked FNN). However, it is important to note that feature importance is computed differently for each model and, therefore, the comparability between the relative importance across our models is limited. Nonetheless, it provides an informative indication of the importance of our individual signals.

First, we analyze the variables with the highest feature importance. The highest emphasis within the GBM is placed on the trendfactor with a relative importance of 10.5%. This signal can be found among the five best-performing signals in relative importance for all three reference models. The same finding of the trendfactor as a meaningful predictor is made by Jacobs and Müller (2018) for an international sample and by Gu et al. (2019) for the US market. In the case of the GLM and FNN, the most important anomaly is the trading volume over market value by Haugen and Baker (1996). Remarkably, for this signal, the relative importance in the case of FNN is 26.7%, which is more than twice as high as the second most important signal (trendfactor with 10.5%), which presumably makes this model less robust. In general, we see three (two) similarities between the most important signals for the FNN compared to the GLM (GBM). As an important common signal between the GLM and GBM, we only observe the trendfactor anomaly. An overview of the five most important signals for each reference model is shown in Fig. 2.

In order to further understand the importance of the anomalies in the machine-learning models, we take a holistic view of the weighting of categorized anomalies within the reference models shown in Fig. A.2 in the Internet appendix. We find that the category with the highest weight across all three machine-learning models is market-based anomalies. However, for our reference baseline factor, fundamental-based signals have the highest weight. This finding is consistent with Jacobs and Müller (2020), showing that accounting signals are less profitable in the global market. Hence, the larger emphasis on market-based signals instead of fundamental-based anomalies might at least partly explain the outperformance of our reference models over our baseline factor as a linear multi-signal combination.

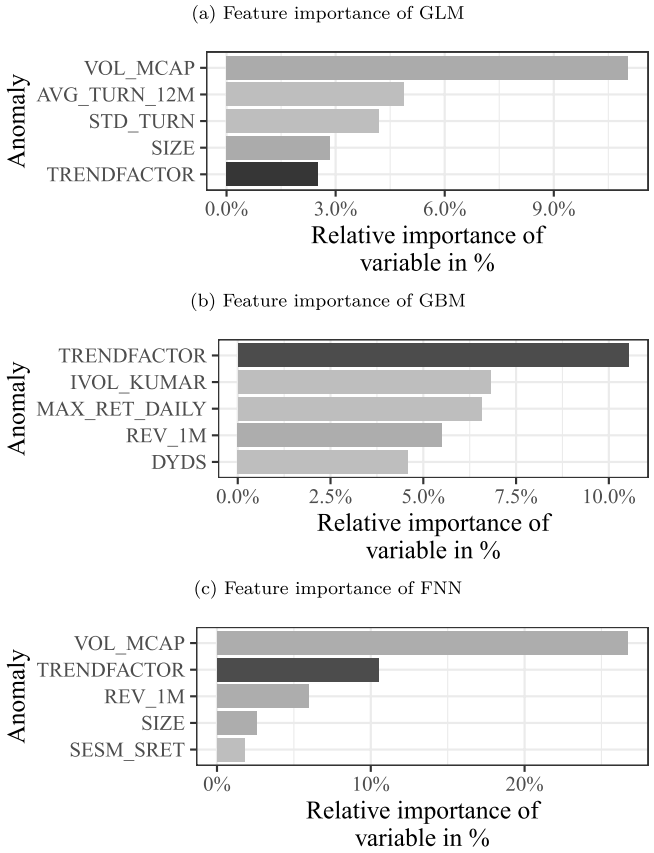
Overall, when analyzing the relative importance methods, we can identify different anomaly weights of our three reference models and, thus, better explain and interpret our models. Here, we investigate significant differences in the weighting and the degree of the weighting of individual signals. Nevertheless, the black-box problem of machine-learning models remains a challenge not only for our models. Current research is increasingly focusing on this topic to strengthen the dissemination and acceptance of machine learning in all application areas (Rudin 2019). For the following analysis, we examine typical preprocessing and training approaches to enhance performance and increase research robustness.

Comparison and robustness tests of machine-learning models

Variations of machine-learning models

Extending machine-learning models with dynamic and rolling training methodologies

In our analysis presented in section "Portfolio construction with machine learning algorithms in a static window", the machine-learning models are based on a static window from July 1980 to June 2003. New information from later stock data is not incorporated into the machine-learning models. This approach has the advantage of being less susceptible to



These graphs illustrate the feature importance of the GLM (Figure a), GBM (Figure b), and FNN (Figure c). The feature importance of the GBM is defined as the improvement of the squared error due to the selection of this variable in the tree-building process. The relative importance of GLM is determined by the coefficient magnitudes. For our FNN, we apply the permutation approach described by Breiman (2001) to compare the distribution of relative variable importance by our three reference models.

Fig. 2 Top five feature importance for machine-learning models

false-positive observations, as the number of trained models is kept low. Furthermore, this approach corresponds to a conservative approach that concentrates on stationary patterns within our data sample.

However, more recent stock data can improve the performance of our models if the predictive power of signals varies over time. For this purpose, we retrain our models on a rolling basis, including new observations over time in our training sample while preventing any forward-looking bias. In particular, we apply three separate rolling window mechanisms: a 5-year rolling window, a 10-year rolling window, and an extending rolling window that always includes all stock data from 1980 onwards. Due to limitations in computational power, the machine-learning models are trained yearly, although the dataset could generally also be updated monthly.

Employment of unsupervised learning and feature reduction techniques to preprocess the high-dimensional factor zoo

There are several preprocessing steps for machine-learning models, such as feature reduction methods, to cope with the high-dimensionality of given datasets (Ye et al. 2006). We address this preprocessing to reduce noise using two feature reduction methods and two scientific-motivated selection approaches based on the significance level.

Among the most commonly used feature reduction techniques are regularization or shrinkage methodologies. These preprocessing methods are based on regressions where additional constraints are added to the model to prevent the model from overfitting and to achieve a better generalization. One regularization form is the lasso regression, which

penalizes weak features to get zero coefficients within a loss function and eliminates no-value-adding signals. In contrast, ridge regression applies a different penalty expression where correlated features tend to get similar coefficients compared to other features, which are spread more equally. A combination of both approaches is the elastic net selection. It simultaneously reduces features by forcing some coefficients to be zero and eliminating correlated features with a similar coefficient. This results in a model that can handle correlated features and select important variables. In our paper, we utilize the elastic net and lasso selection.

Using the elastic net and lasso for feature selection can have some advantages. For example, our study applies a large number of stock return predictor variables, and many of these variables may be redundant or add some noise to the models. By using these feature-reduction techniques, we can reduce the noise of the models and prevent overfitting (i.e., when a model is too complex and fits the training data too closely, leading to poor out-of-sample performance). In addition to these two approaches, we downsize our signal sample to only anomalies with *t*-statistics above 1.96 and 3.00 (Harvey et al. 2016) targeting only significant signals without any noise of less important anomalies.

Analysis of 40 machine-learning models

Following Azevedo and Hoegner (2023), we estimate the returns for 40 relevant models with variations on the machine-learning algorithms applied, the target values, the training window, and feature reduction methods. Table 4 compares the performance of these methods. We find that the highest monthly return on average of 2.71% (*t*-statistic of 9.48) is reported by the percent-ranked small FNN with an extending window and a full feature base. Furthermore, we consistently see high *t*-statistic values and comparable performance across models with different feature reduction methods and rolling windows, which indicate a lower risk of false-positive observations.

Overall, 39 out of 40 implemented machine-learning models report a higher mean monthly value-weighted return for the post-July 2003 period than our baseline factor as a linear benchmark (1.02%). Moreover, for 30 models, the return difference to the baseline factor is positive and statistically significant at the 0.05 level. Here, the small FNN as the best-performing model indicates a performance difference of 1.68% compared to the baseline factor (*t*-static of 6.87).

Combinations of machine-learning models

To avoid data-dredging concerns, where the models are selected ex-post, we combine these 40 models according to their target values to get an aggregated predictor. Analyzed by Timmermann (2006) and Rasekhschaffe and Jones (2019), forecast combinations might enhance the information level while reducing noise, particularly in the case of relatively uncorrelated forecast biases and various tested model methodologies.

Our set of tested machine-learning models incorporates six algorithms based on two target values, optimized by six distinct feature selection methods, and trained on three rolling windows approaches. Therefore, our set of models has a high degree of diversity, which can be advantageous to achieve greater accuracy and robustness.

To create our machine-learning combinations, we treat our 40 machine-learning models separately based on their target values. In general, there are multiple combinations approaches available derived from the literature, including methodologies based on statistics (such as the Bates and Granger combination method (e.g., Bates and Granger 1969), regressions (e.g., Granger and Ramanathan 1984), eigenvectors (e.g., Hsiao and Wan 2014), as well as more enhanced approaches considering volatility, mean-variance, or idiosyncratic-return adjustments (e.g., Hanauer and Windmüller 2020). While an in-depth application of each mentioned combination methodology would be out of the scope of our paper, we focus on one specific Statistic-based equal-weighted machine-learning combination (STATEW). First, we average all machine-learning models with the same weight for each month. Afterwards, we apply the same portfolio-sort strategy we have already used throughout the paper. As mentioned above, this intuitive combination can be used as a benchmark for more complex combinations for future research.

Inspired by our results in section "Portfolio construction with machine learning algorithms in a static window", we extend our two combination models based on all machine-learning models with the best-performing feature reduction methodologies elastic net selection and lasso regression. With these feature selections, we can make intelligent model choices that might improve our performance.

As shown in Table 5, our six equal-weighted combinations demonstrate promising results, including a significant improvement compared to our linear baseline factor benchmark. On average, the combinations based on percent-ranked target values achieve higher performance with a monthly average return of 2.53% than the combinations based on our raw-return machine-learning models with only 1.93%. In both variants, the feature selection methods improve the performance compared to the full feature case. For the percent-ranked combinations, the lasso selection of 12 out of 27 models slightly outperforms the elastic net selection of 15 models (2.60% compared to 2.53% average monthly return). Here, both selection techniques mainly focus on selecting models based on neural networks with a share between 67% and 75%. In contrast, for the raw-return combinations, the elastic net selection and lasso regression choose the same

Table 4 Analysis of machine-learning models compared to the linear baseline factor

Model specifications				Performance		Baseline factor improvement	
Algorithm	Return target	Feature set	Rolling learning	Return in %	*t*-stat.	Add. return in %	*t*-stat.
FNN	Percent-ranked	Full	Extending	2.71	9.48	1.68	6.87
FNN	Percent-ranked	Full	10y-rolling	2.70	9.73	1.68	7.19
FNN	Percent-ranked	Elastic net	Static	2.55	8.65	1.52	6.2
FNN (Larger)	Percent-ranked	Elastic net	Static	2.55	9.09	1.53	7.54
FNN (Larger)	Percent-ranked	Full	Extending	2.55	9.32	1.53	6.49
FNN	Percent-ranked	Lasso	Static	2.41	8.85	1.38	6.57
FNN (Larger)	Percent-ranked	Full	10y-rolling	2.38	8.44	1.36	5.62
FNN (Larger)	Percent-ranked	Lasso	Static	2.30	7.53	1.27	5.95
FNN	Percent-ranked	Full	5y-rolling	2.28	7.45	1.26	5.15
FNN (Larger)	Percent-ranked	Full	Static	2.24	8.99	1.22	6.16
FNN (Larger)	Percent-ranked	*t*-stat. > 1.96	static	2.24	7.89	1.21	5.29
RNN	Percent-ranked	Full	Static	2.10	6.95	1.08	5.22
FNN	Percent-ranked	*t*-stat. > 1.96	Static	2.07	7.2	1.05	4.81
GBM	Percent-ranked	Full	10y-rolling	2.01	6.67	0.99	4.95
FNN	Percent-ranked	Full	Static	2.00	7.79	0.98	5.03
FNN (Larger)	Percent-ranked	Full	5y-rolling	2.00	6.21	0.98	3.4
FNN	Raw	Full	Static	1.99	8.03	0.97	5.24
GBM	Percent-ranked	Full	Extending	1.86	6.63	0.84	4.11
GBM	Percent-ranked	Lasso	Static	1.79	7.01	0.77	4.38
FNN (larger)	Raw	Full	Static	1.75	7.39	0.73	4.11
GBM	Percent-ranked	Full	Static	1.75	6.81	0.73	4.09
FNN	Percent-ranked	*t*-stat. > 3	static	1.70	5.45	0.67	3.28
GBM	Percent-ranked	Elastic net	Static	1.68	6.66	0.65	3.6
DRF	Percent-ranked	Full	Static	1.66	5.63	0.64	2.95
GBM	Percent-ranked	Full	5y-rolling	1.65	5.3	0.62	2.75
GLM	Raw	Full	Extending	1.63	8.21	0.61	2.27
FNN (larger)	Percent-ranked	*t*-stat. > 3	Static	1.60	4.87	0.57	2.66
GBM	Percent-ranked	*t*-stat. > 1.96	Static	1.58	6.24	0.56	2.99
GLM	Percent-ranked	Full	Static	1.49	6.67	0.47	3.14
GLM	Raw	Full	10y-rolling	1.49	7.58	0.46	1.69
GLM	Raw	Elastic net	Static	1.48	8.06	0.46	1.83
RNN	Raw	Full	Static	1.47	5.66	0.44	2.54
GLM	Raw	Full	Static	1.43	8.03	0.41	1.77
GLM	Raw	Lasso	Static	1.43	7.74	0.41	1.63
GBM	Percent-ranked	*t*-stat. > 3	Static	1.40	5.01	0.38	1.88
GLM	Raw	Full	5y-rolling	1.28	6.19	0.25	0.94
GLM	Raw	*t*-stat. > 1.96	Static	1.27	6.34	0.25	1.03
GBM	Raw	Full	Static	1.21	6.35	0.19	0.9
GLM	Raw	*t*-stat. > 3	Static	1.07	5.3	0.05	0.23
DRF	Raw	Full	Static	1.00	5.8	−0.02	−0.1

The table above compares the relevant set of our applied machine-learning models. We distinguish between machine-learning algorithm types, raw return, and percent-ranked target values, the applied feature selection processes, and the rolling training techniques. In addition to the absolute model performance, we note the differences compared to our baseline factor, defined as the average differences in monthly returns. Moreover, the latter is described by the *t*-statistic indicating the statistical significance of the differences

Table 5 Performance comparison for machine-learning combinations

Model specifications				Performance		Baseline factor improvement	
Machine-learning combination	Return target	Feature set	Rolling learning	Return in %	*t*-stat.	Add. return in %	*t*-stat.
Equal-weighted	Percent-ranked	Lasso	Static	2.60	8.54	1.57	6.58
Equal-weighted	Percent-ranked	Elastic net	Static	2.53	8.59	1.51	6.58
Equal-weighted	Percent-ranked	Full	Static	2.47	8.47	1.45	6.46
Equal-weighted	Raw	Elastic net	Static	1.97	8.47	0.95	5.37
Equal-weighted	Raw	Lasso	Static	1.97	8.47	0.95	5.37
Equal-weighted	Raw	Full	Static	1.85	8.7	0.82	4.9

The table above compares the performance of our applied machine-learning combinations. We distinguish between the types of combinations, raw-return and percent-ranked target values, the feature selection processes, and the rolling training techniques. Moreover, we note the improvement compared to our baseline factor, defined as the average differences in monthly returns. Furthermore, the latter is described by the *t*-statistic indicating the statistical significance of the differences

six features out of the set of 13 raw-return machine-learning models, namely the static GBM, the static large FNN, the static small FNN, the static RNN, and the static and rolling 10-year GLM (all six models are based on a full feature base). Therefore, both machine-learning combinations have a monthly return of 1.97% on average.

In summary, we combine a set of tested machine-learning models separated by raw-return and percent-ranked models into composite predictors. These machine-learning combinations can be classified as diverse predictors based on a relatively uncorrelated input. Our 40 models cover a broad range of used target values, algorithms, feature reduction methods, and training techniques. These characteristics increase the stability and robustness of our models through a higher level of included information and a reduction of noise. We use the elastic net and lasso feature selections to improve the promising performance already shown to achieve additional gains.

Comparison among all machine-learning models

Next, we identify the characteristics and specifications that make machine-learning models successful at predicting stock returns. We assess the effect of changing machine-learning algorithms, window training, target prices, and feature reduction method. We perform regressions of monthly returns on time fixed-effects and dummies for each parameter (e.g., the machine-learning approach, window, forecast variable, and feature reduction). The results are shown in Table 6.

In specification 1, we first analyze the impact of changing the machine-learning algorithm. The intercept (baseline) of 1.40% (*t*-statistic of 7.53) refers to the performance of the GLM model. When comparing the machine-learning approaches, the FNN model reports a monthly return of 87 basis points higher than the GLM model with a *t*-stat of 3.30. In comparison, the combination model is around 83 basis points higher with a *t*-stat of 3.48. These results indicate that nonlinear models and combinations seem to outperform a linear model, such as the GLM model.

In specification 2, we analyze the impact of using a 5-year and 10-year rolling window and extending the window compared to a static window (intercept). While using a 5-year rolling window is not statistically different from a static window, we find that extending and 10-year rolling windows are the best-performing training windows. For instance, extending windows leads to almost 35 basis points higher monthly returns than the static window, which is evidence that using more recent data rather than a static window can improve the predictability power of the models. However, a 5-year rolling window seems to be too short to deal with the high complexity of stock returns.

Concerning the target return, we find in specification 3 that using percent-ranked returns leads to monthly returns around 58 basis points (*t*-stat of 3.22) higher than raw returns (intercept). These results indicate that using raw returns in the objective function is too noisy, and percent-ranked returns might be an effective alternative to deal with noisy returns. This is a very important finding given that most of the literature on machine learning uses raw returns as a target (e.g., Gu et al. 2020).

In specification 4, we test feature reduction methods. When analyzing the baseline, which is a model with the full feature set (i.e., including all predictors), we find an average monthly return of 1.883%. However, when we use elastic net (lasso), we see a return increase of approximately 0.25% (0.20%) with a *t*-statistic of 4.16 (3.76) compared to the baseline. These results indicate that these two feature reduction methods are important preprocessing tools to deal with the high dimensionality of global equity markets. In particular, these feature selection methods seem to do a

Table 6 The effect of machine-learning approach, window, forecast variable, and feature reduction on returns

Baseline:	(1) Model GLM	(2) Window static	(3) Target raw returns	(4) Feature full	(5) GLM, static, raw returns, full feature
FNN	0.8703***				0.5231***
	(3.30)				(2.71)
RNN	0.3848				0.2580
	(1.47)				(1.15)
DRF	−0.0652				−0.1921
	(−0.36)				(−1.25)
GMB	0.2616				−0.0856
	(1.21)				(−0.57)
FNN (larger)	−0.0882**				−0.0882**
	(−2.14)				(−2.14)
Combination	0.8347***				0.6005***
	(3.48)				(2.84)
Extending_Window		0.3472***			0.3348***
		(5.00)			(4.32)
Rolling_5Y		−0.0406			−0.0531
		(−0.32)			(−0.41)
Rolling_10Y		0.3043***			0.2918***
		(3.68)			(3.37)
Percent Rank			0.5764***		0.4464***
			(3.22)		(3.55)
Elastic Net				0.2450***	0.183
				(4.16)	(3.76)
Lasso				0.1995***	0.1383***
				(3.76)	(3.29)
Only_Likely				−0.0926	−0.0636
				(−1.09)	(−0.88)
Only_Very_Likely				−0.4407***	−0.4116***
				(−3.92)	(−4.03)
Intercept	1.3974***	1.8415***	1.5187***	1.8830***	1.3010***
	(7.53)	(88.31)	(13.00)	(79.36)	(6.56)
Observations	8,786	8,786	8,786	8,786	8,786
R-squared	0.633	0.624	0.628	0.625	0.638
Regtype	Fixed-effects	Fixed-effects	Fixed-effects	Fixed-effects	Fixed-effects
Date dummy	Yes	Yes	Yes	Yes	Yes
Cluster	Month	Month	Month	Month	Month
Region	World	World	World	World	World
Period	2003–2019	2003–2019	2003–2019	2003–2019	2003–2019

This table reports the regressions of monthly returns on time fixed-effects and dummies for the machine-learning approach, window, forecast variable, and feature reduction. Specification 1 shows the results according to the machine-learning model, and the baseline model is the GLM approach. Specification 2 shows the results according to the window, where the baseline model is a static window. Specification 3 classifies the models according to the forecasting objective, and the baseline model uses raw returns. Specification 4 analyzes feature reduction methods and the baseline is the full feature (i.e., including all predictors without feature reduction). Specification 5 shows the results based on the machine-learning approach, window, forecast variable, and feature reduction on returns. The baseline is based on the GLM model, static window, raw returns, and full feature predictors. Standard errors are clustered by time

Table 7 Round-trip costs of the top-30 performing models compared to the baseline factor

Model name	Return [*t*-stat]	Turnover rate (%)	Round-trip costs (%)
FNN.PERCENTRANK.FULL.ROLLING10Y	2.7% [9.73]	65.72	3.28
STATEW.PERCENTRANK.LASSO.STATIC	2.6% [8.54]	61.17	3.27
FNN.PERCENTRANK.FULL.ROLLINGEXT	2.71% [9.48]	65.98	3.26
STATEW.PERCENTRANK.ELASTICNET.STATIC	2.53% [8.59]	60.44	3.24
STATEW.PERCENTRANK.FULL.STATIC	2.47% [8.47]	59.71	3.18
FNN (larger).PERCENTRANK.ELASTICNET.STATIC	2.55% [9.09]	65.33	3.07
FNN (larger).PERCENTRANK.FULL.ROLLINGEXT	2.55% [9.32]	66.44	3.04
FNN.PERCENTRANK.ELASTICNET.STATIC	2.55% [8.65]	64.95	3.03
STATEW.RETURN.ELASTICNET.STATIC	1.97% [8.47]	51.37	2.95
STATEW.RETURN.LASSO.STATIC	1.97% [8.47]	51.37	2.95
FNN.PERCENTRANK.LASSO.STATIC	2.41% [8.85]	65.83	2.85
STATEW.RETURN.FULL.STATIC	1.85% [8.7]	50.42	2.84
FNN (larger).PERCENTRANK.FULL.ROLLING10Y	2.38% [8.44]	66.08	2.76
FNN (larger).PERCENTRANK.LASSO.STATIC	2.3% [7.53]	61.59	2.76
FNN (larger).PERCENTRANK.LIKELY.STATIC	2.24% [7.89]	62.97	2.67
FNN (larger).PERCENTRANK.FULL.STATIC	2.24% [8.99]	67.09	2.62
RNN.PERCENTRANK.FULL.STATIC	2.1% [6.95]	57.76	2.61
FNN.PERCENTRANK.FULL.ROLLING5Y	2.28% [7.45]	64.88	2.59
FNN.PERCENTRANK.FULL.STATIC	2% [7.79]	62.09	2.42
FNN.PERCENTRANK.LIKELY.STATIC	2.07% [7.2]	62.61	2.41
FNN.RETURN.FULL.STATIC	1.99% [8.03]	62.72	2.4
GBM.PERCENTRANK.FULL.ROLLING10Y	2.01% [6.67]	66.3	2.14
FNN (larger).PERCENTRANK.FULL.ROLLING5Y	2% [6.21]	66.11	2.07
FNN (larger).RETURN.FULL.STATIC	1.75% [7.39]	65.33	1.97
GBM.PERCENTRANK.FULL.ROLLINGEXT	1.86% [6.63]	67.41	1.94
FNN.PERCENTRANK.VERYLIKELY.STATIC	1.7% [5.45]	58.48	1.86
GBM.PERCENTRANK.LASSO.STATIC	1.79% [7.01]	70.53	1.83
FNN (larger).PERCENTRANK.VERYLIKELY.STATIC	1.6% [4.87]	53.56	1.78
GBM.PERCENTRANK.FULL.STATIC	1.75% [6.81]	71.23	1.75
GLM.RETURN.FULL.ROLLINGEXT	1.63% [8.21]	71.38	1.74
BASELINEFACTOR	1.02 [4.93]	38.02	1.62

This table reports the maximum round-trip costs each machine-learning model could have to remain statistically significant at the 0.05 significance level. We show results for 30 machine-learning models and the baseline factor, including the one-sided turnover rate

good job at reducing the noise of the models and eliminating redundant predictors. In other words, these results indicate that not all predictors used in our study necessarily add predictability power to our machine-learning models.

Finally, in specification 5, we include all different parameters in the same regression. Overall, the inferences are mostly unchanged, but the magnitudes of the differences are slightly smaller. Furthermore, among the machine-learning approaches, after controlling for all parameters, the combination reports the highest return, with 60 basis points higher return compared to the baseline model with a GLM approach, static window, raw returns, and using the full feature set.

Overall, these results indicate that the right choice of specifications can play a major role in the predictability of the machine-learning model. Compared to the baseline (GLM, static, raw returns, and full feature set), the regression in specification 5 indicates that it is possible to have an increase of up to 156 basis points by applying a combination of machine learning with extending rolling-window, using percent-ranked returns as a target and elastic net as a feature reduction.

Turnover rate and acceptable transaction costs estimation

Thus far, our machine-learning models have been trained to predict the next month's return for each stock to optimize the return of the portfolio-sort strategy by maximizing the spread of the long-short positions. This strategy does not consider the monthly relative rebalancing amount of shares entered or removed from the portfolio. As each rebalancing execution of a single stock is associated with transaction costs, our current approach does not illustrate the real profitability of possible trading strategies. The real profitability is rather significantly influenced by the relative rebalancing amount, also framed as the one-sided turnover rate. To address this issue, we integrate a measure for potential transaction costs into our assessment in the following.

While single anomaly strategies greatly vary in turnover rates, the relative rebalancing amount by machine-learning models is relatively high, around 50% to 75%. These findings highlight the potential relevance of transaction costs, especially as the machine-learning models with high turnover rates might display a substantially lower return after costs.

Nevertheless, these analyzed turnover rates do not directly reflect the associated transaction costs for a real trading strategy implementation. Therefore, we calculate the round-trip costs to estimate the upper limit of acceptable transaction costs (Grundy and Martin 2001; Barroso and Santa-Clara 2015; Hanauer and Windmüller 2020). Current literature assumes 50 basis points as transaction cost parameter (Lassance and Vrins 2021). This measure enables us to analyze whether enhanced model returns can offset increased transaction costs associated with larger turnover rates. The latter applies a significance level of 5% for a Z-score and is calculated as follows (Hanauer and Windmüller 2020):

$$\text{Round-trip costs}_{\alpha=5\%} = \left(1 - \frac{1.96}{T_S}\right) \times \frac{\bar{\mu_S}}{T\bar{O}_S} \tag{1}$$

where S = Portfolio strategy S, T_S= t-statistic of strategy S, μ_S = Average monthly return of strategy S, TO_S = One-sided turnover rate of strategy S.

As shown in Table 7, the set of our top-30 performing machine-learning models compensates with their increased performance for the higher turnover rate. The round-trip costs for all tested machine-learning models range between 1.74 and 3.28%, implying a realistic upper limit for acceptable transaction costs to sustain a profitable trading strategy for our models. Compared to the baseline factor model, the machine-learning models seem to beat the linear benchmark performance in monthly returns and acceptable transaction costs. Simultaneously, we find that long-short strategies based on our six composite predictors remain significant at the 0.05 level with round-trip costs between 284 and 327 basis points, which collectively proves that the outperformance of the models is not explained by transaction costs. From practitioners' perspective, these results underline the profitability for a real implementation of the training strategies based on the analysis of nonlinear relationships across the factor zoo with our machine-learning models.

Classification of machine-learning returns as a mispricing or risk components of established factor models

In this section, we test the return of our models against classical factor models using linear regressions to determine whether the returns are due to common (risk) components. Throughout the asset pricing theory, there are multiple established factor models. To this end, we regress the long-short portfolio returns resulting from selected machine-learning models against eight distinct factor models: the Carhart (1997) four-factor model, the Capital Asset Pricing Model (CAPM), the behavioral factor model of Daniel et al. (2020) (DHS), the three- and five-factor models of Fama and French (1992) (FF-3 and FF-5), both the Q-factor model and the augmented Q-factor, and the mispricing factor model of Stambaugh and Yuan (2017).

Our factor regression tests comprise the following models: a GLM with raw returns and a static window, a large FNN with a percent-ranked target and a static window, a GBM model with a percent-ranked target and a static window, and two combinations of models (STATEW). Furthermore, we add a GBM model with a percent-ranked target and rolling window, including only anomalies post-publication. An insignificant alpha value would suggest that machine-learning model returns could be (fully) explained by factor models.

The results in Table 8 show that the three reference models enjoy significant alpha values ranging between 0.87 and 2.43%, with t-statistics firmly above the critical value of 3.0. The post-publication GBM has lower but still statistically significant alphas between 0.60 and 1.95%. More precisely, for the two Q-Factor models and the DHS model, t-statistics of 2.68, 2.85, and 2.38 are reported for the post-publication case as the only three values slightly below three across all tested models. Noteworthy, our STATEW models embody significant alphas ranging between 1.10 and 2.69%. Moreover, we find that models including neural network approaches show more prominent alphas (i.e., the FNN or the STATEWs).

In summary, we highlight significant alphas across the entire range of different tested machine-learning models. Consequently, the returns of our models are not satisfactorily explainable by common asset pricing models. In addition to our analysis in the previous sections, we

Table 8 Analysis of machine-learning models against common factor models

Factor model	GLM (raw, static, full)	Large FNN (perc.-ranked, static, full)	GBM (perc.-ranked, static, full)	GBM (perc.-ranked, rolling, post-publication)	STATEW (raw, static, full)	STATEW (perc.-ranked, static, full)
Carhart Four-Factor	1.28 %	2.25 %	1.8 %	1.55 %	1.83 %	2.42 %
	[6.97]	[9.66]	[8.59]	[6.5]	[10.18]	[8.95]
Capital Asset Pricing	1.44 %	2.41 %	2.03 %	1.82 %	2.06 %	2.64 %
	[7.99]	[10.08]	[9.07]	[6.62]	[10.97]	[9.32]
DHS	1.21 %	1.55 %	0.87 %	0.6 %	1.1 %	1.69 %
	[6.44]	[6.69]	[4.12]	[2.38]	[6.13]	[6.11]
Fama-French Five-Factor	1.4 %	2.08 %	1.73 %	1.43 %	1.81 %	2.27 %
	[7.26]	[9.06]	[8.18]	[5.86]	[9.69]	[8.26]
Fama-French Three-Factor	1.46 %	2.43 %	2.08 %	1.95 %	2.1 %	2.69 %
	[8.01]	[10.66]	[9.75]	[7.66]	[11.32]	[10.02]
Q-Factor (Augmented)	1.06 %	1.66 %	1.01 %	0.73 %	1.27 %	1.56 %
	[4.69]	[6.16]	[4.29]	[2.68]	[6.12]	[4.99]
Q Factor	1.04 %	1.79 %	1.14 %	0.74 %	1.25 %	1.71 %
	[4.81]	[6.87]	[4.98]	[2.85]	[6.31]	[5.65]
Mispricing Factors	1.14 %	1.78 %	1.31 %	0.9 %	1.39 %	1.99 %
	[5.56]	[7.23]	[5.93]	[3.6]	[7.47]	[6.86]

The table above tests our three reference models, thepost-publication GBM, and two composite predictor models against eight classical factor models. The alphas and the *t*-statistics are calculated based on a linear regression within the time horizon from August 2003 to June 2019

underline the improbable cause of data dredging for the discovered outperformance of our models. Therefore, these findings challenge the EMH in the global stock universe. In the past, comparable findings defined as arbitrage possibilities were usually exploited fast after publication by investors seeking profitable trading strategies. In contrast, exploiting the returns of our machine-learning models might be more challenging due to the increased complexity and limited interpretation of these algorithms. This circumstance possibly explains our findings for a static frame with a training set consisting of pre-July 2003 stock information. As the interpretability of machine-learning models increases, the exploitation of these profitable nonlinear relationships might become more achievable by practitioners.

Practical implications and considerations

Our analyses suggest that machine-learning models would have outperformed passive buy-and-hold strategies as well as a linear combination of individual anomalies by a substantial margin during our sample period. Since our calculations are based on value-weighted portfolios, and the long-short outperformance remains intact even after accounting for significant transaction costs, our findings should be of interest to portfolio managers and other investment professionals.

Nevertheless, when implementing such models, we urge practitioners to follow a structured approach, which takes into account the challenges and limitations of these techniques. In our experience, careful data preprocessing, which includes proper handling of outliers and missing values, is extremely important, particularly when working with international data. For example, we propose a percent-ranking of input variables as an effective and simple solution to deal with outliers and data errors. For feature selection, practitioners should take advantage of the fact that there is a large number of academic studies providing a rich ground for identifying potentially relevant predictors. The reliance on other work, which is published and typically also peer-reviewed, to select input variables may also mitigate the risk of overfitting, which is a common problem in machine learning. We have taken additional steps such as cross-validation and forecast combinations to reduce overfitting and data mining risk further, and we encourage practitioners to use similar techniques.

In light of the evolving nature of financial markets, it seems also likely that the relevance of individual stock predictors changes over time. This can be accounted for by regularly updating the list of features and using techniques such as rolling- or extending-window estimation, which work well in our study. Finally, portfolio managers may seek a cost-efficient implementation of machine-learning approaches by limiting portfolio turnover or focusing on stocks with high liquidity.

In addition to these practical considerations essential for the effective deployment of machine-learning models, investment professionals should also examine potential ethical and regulatory concerns associated with using these

models. For example, from a regulatory and compliance perspective, the complexity and opacity of machine learning may make it difficult to properly monitor and assess the risks in larger asset portfolios. Feature importance analysis is one solution to address this black box problem of machine learning as shown in section "Interpretation of the machine learning models through relative feature importance". In addition, one may reduce the opacity of the models by continuously monitoring (risk) factor and industry exposures.

Market stability and efficiency can also be impacted if machine learning models are used at a larger scale in the asset management industry. For example, if multiple institutions relied on similar models, they could be exposed to correlated risks, which may lead to market instability during periods of financial stress, similar to the "quant meltdown" experience of hedge funds in August 2007 (Khandani and Lo 2011).

A related concern is that the performance of machine-learning models could be more sensitive to sudden changes in market conditions, potentially leading to elevated crash risk as has been observed for individual anomalies like stock price momentum (Daniel and Moskowitz 2016). However, in our study, we do not find evidence that the performance of the long-short portfolios is dependent on market conditions. In unreported results, we test whether the return of long-short portfolios of machine-learning models are different in periods with above average investor sentiment and CBOE Volatility Index (VIX), or during NBER-dated recessions in the U.S.[5] In our analysis, none of the variables had a statistically significant impact on portfolio performance.

Conclusion

Our study analyzes the performance of machine-learning models in a global stock universe to predict stock returns. Our tested machine learning models enjoy a significant monthly average value-weighted return of up to 2.71%, illustrating the superiority over the baseline factor. It is worth focusing on our composite machine learning predictors to avoid any forward-looking bias in selecting the best-performing model. These composite predictors demonstrate promising returns of up to 2.60% and underline the impact of nonlinear effects on asset pricing.

Additionally, our study extends the existing anomaly research about machine learning with an international view and strengthens measures against *p*-hacking, leading to greater study robustness. Furthermore, with different types of algorithms, training approaches, and feature selections, we enlarge the set of tested machine-learning models as the basis of the creation of several composite predictors. Due to these conducted measures, our findings cannot be merely traced back to data dredging. The outperformance of our models is not explained by common factor models, and likely suggests market inefficiencies and mispricing.

Consequently, researchers might focus more closely on nonlinear relationships across anomalies within the factor zoo and investigate individual anomalies and linear connections. Thereby, nonlinear hidden patterns identified by complex machine-learning algorithms might provide further insights into international asset pricing. To enhance robustness, more complex composite predictors of multiple machine-learning models, as well as reinforcement learning algorithms, might be applied by subsequent scholars. Finally, for practitioners, our findings might offer new possibilities for profitable trading strategies supplementary to the mostly exhausted patterns of individual anomalies.

Our study emphasizes the importance of nonlinear relationships within the factor zoo and their impact on international asset pricing. With continued progress in research, enhanced interpretation measures, and greater computational power, smart machine-learning algorithms have the capabilities to broaden our knowledge of asset pricing. In the future, models based on these smart algorithms may announce a new development of improved asset pricing models incorporating nonlinear effects. In the meantime, these hidden patterns might yield arbitrage opportunities for practitioners who can navigate these complexities.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1057/s41260-023-00318-z.

Acknowledgments We thank Lucas Hahn, Christopher Hoegner, Christoph Riedersberger, and Heiko Jacobs for insightful discussions andhelpful comments. None of the authors has conflicts of interest to disclose.

Funding Open Access funding enabled and organized by Projekt DEAL.

[5] The data for investor sentiment was taken from Jeffrey Wurgler's data library, while the data from VIX was taken from Michael W. McCracken's data library of macroeconomic variables.

References

Abiodun, O.I., A. Jantan, A.E. Omolara, K.V. Dada, N.A. Mohamed, and H. Arshad. 2018. State-of-the-art in artificial neural network applications: A survey. *Heliyon* 4 (11): 1–41.

Anand, V., R. Brunner, K. Ikegwu, T. Sougiannis, 2019. Predicting profitability using machine learning. *SSRN Electronic Journal*, pp. 1–63.

Andrew Karolyi, G. 2016. Home bias, an academic puzzle. *Review of Finance* 20 (6): 2049–2078.

Azevedo, V., and C. Hoegner. 2023. Enhancing anomalies with machine learning. *Review of Quantitative Finance and Accounting* 60 (1): 195–230.

Banz, R.W. 1981. The relationship between return and market value of common stocks. *Journal of Financial Economics* 9 (1): 3–18.

Barroso, P., and P. Santa-Clara. 2015. Momentum has its moments. *Journal of Financial Economics* 116 (1): 111–120.

Bates, J., C. W. J. Granger, 1969. The combination of forecasts. operations research Quaterly, v. 20. *Operations Research Quarterly*, 20(4):451–468.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24 (2): 123–140.

Breiman, L. 2001. Random forests. *Machine Learning* 45 (1): 5–32.

Breitung, C. 2023. Automated stock picking using random forests. *Journal of Empirical Finance*, Forthcoming, pp. 1–51.

Cakici, N., C. Fieberg, D. Metko, A. Zaremba, 2022. Machine learning goes global: Cross-sectional return predictability in international stock markets. *SSRN Electronic Journal*, pp. 1–59.

Carhart, M.M. 1997. On persistence in mutual fund performance. *The Journal of Finance* 52 (1): 57–82.

Chen, A. Y., T. Zimmermann, 2022. Open source cross-sectional asset pricing. *Critical Finance Review*, 11 (2): 207–264.

Chen, L., M. Pelger, J. Zhu, 2023. Deep learning in asset pricing. *Management Science*, Forthcoming.

Clemen, R.T. 1989. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting* 5 (4): 559–583.

Cochrane, J.H. 2011. Presidential address: Discount rates. *The Journal of finance* 66 (4): 1047–1108.

Daniel, K., D. Hirshleifer, and L. Sun. 2020. Short-and long-horizon behavioral factors. *The Review of Financial Studies* 33 (4): 1673–1736.

Daniel, K., and T.J. Moskowitz. 2016. Momentum crashes. *Journal of Financial Economics* 122 (2): 221–247.

Drobetz, W., T. Otto, 2021. Empirical asset pricing via machine learning: Evidence from the European stock market. *SSRN Electronic Journal*, pp. 1–60.

Fama, E.F. 1998. *Market efficiency, long-term returns, and behavioral finance*. Chicago: University of Chicago Press.

Fama, E.F., and K.R. French. 1992. The cross-section of expected stock returns. *The Journal of Finance* 47 (2): 427–465.

Fama, E.F., and K.R. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33 (1): 3–56.

Fama, E.F., and K.R. French. 2015. A five-factor asset pricing model. *Journal of Financial Economics* 116 (1): 1–22.

Fieberg, C., D. Metko, T. Poddig, T. Loy, 2023. Machine learning techniques for cross-sectional equity returns' prediction. *OR Spectrum*, pp. 289–323.

Fong, K.Y.L., C.W. Holden, and C.A. Trzcinka. 2017. What are the best liquidity proxies for global research? *Review of Finance* 21 (4): 1355–1401.

Granger, C.W., and R. Ramanathan. 1984. Improved methods of combining forecasts. *Journal of Forecasting* 3 (2): 197–204.

Green, J., J.R. Hand, and X.F. Zhang. 2017. The characteristics that provide independent information about average US monthly stock returns. *The Review of Financial Studies* 30 (12): 4389–4436.

Grundy, B.D., and J.S.M. Martin. 2001. Understanding the nature of the risks and the source of the rewards to momentum investing. *The Review of Financial Studies* 14 (1): 29–78.

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

Gu, S., B.T. Kelly, and D. Xiu. 2019. Autoencoder asset pricing models. *Journal of Econometrics* 222 (429–450): 19–24.

H2O.ai. 2020. H2O.Ai programming library. https://www.h2o.ai.

Hanauer, M.X., T. Kalsbach. 2022. Machine learning and the cross-section of emerging market stock returns. *SSRN Electronic Journal*, pp. 1–89.

Hanauer, M.X., S. Windmüller, 2020. Enhanced momentum strategies. *SSRN Electronic Journal*, pp. 1–65.

Harvey, C.R. 2017. Presidential address: The scientific outlook in financial economics. *The Journal of Finance* 72 (4): 1399–1440.

Harvey, C.R., and Y. Liu. 2014. Evaluating trading strategies. *The Journal of Portfolio Management* 40 (5): 108–118.

Harvey, C. R., Y. Liu, 2019. A census of the factor zoo. *SSRN Electronic Journal*, pp. 1–7.

Harvey, C. R., Y. Liu, and H. Zhu. 2016. ...and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. Boosting and additive trees. In *The Elements of Statistical Learning*, pp. 337–387. Springer.

Haugen, R.A., and N.L. Baker. 1996. Commonality in the determinants of expected stock returns. *Journal of Financial Economics* 41 (3): 401–439.

Hochreiter, S., and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9 (8): 1735–1780.

Hou, K., C. Xue, and L. Zhang. 2015. Digesting anomalies: An investment approach. *Review of Financial Studies*, pp. 650–705.

Hsiao, C., and S.K. Wan. 2014. Is there an optimal forecast combination? *Journal of Econometrics* 178: 294–309.

Ince, O.S., and R.B. Porter. 2006. Individual equity return data from Thomson Datastream: Handle with care! *Journal of Financial Research* 29 (4): 463–479.

Jacobs, H. 2016. Market maturity and mispricing. *Journal of Financial Economics* 122 (2): 270–287.

Jacobs, H., and S. Müller. 2018. ... And nothing else matters? On the dimensionality and predictability of International Stock Returns. *SSRN Electronic Journal*, pp. 1–44.

Jacobs, H., and S. Müller. 2020. Anomalies across the globe: Once public, no longer existent? *Journal of Financial Economics* 135 (1): 213–230.

Jegadeesh, N., and S. Titman. 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance* 48 (1): 65–91.

Khandani, A.E., and A.W. Lo. 2011. What happened to the quants in August 2007? Evidence from factors and transactions data. *Journal of Financial Markets* 14 (1): 1–46.

Lassance, N., and F. Vrins. 2021. Portfolio selection with parsimonious higher comoments estimation. *Journal of Banking & Finance* 126: 106–115.

Leippold, M., Q. Wang, and W. Zhou. 2022. Machine learning in the Chinese stock market. *Journal of Financial Economics*, 145(2, Part A):64–82.

Makridakis, S., and M. Hibon. 2000. The M3-competition: Results, conclusions and implications. *International Journal of Forecasting* 16 (4): 451–476.

McLean, R.D., and J. Pontiff. 2016. Does academic research destroy stock return predictability? *The Journal of Finance* 71 (1): 5–32.

Nelder, J.A., and R.W. Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* 135 (3): 370–384.

Probst, P., and A.-L. Boulesteix. 2017. To tune or not to tune the number of trees in random forest. *J. Mach. Learn. Res.* 18 (1): 6673–6690.

Rasekhschaffe, K.C., and R.C. Jones. 2019. Machine learning for stock selection. *Financial Analysts Journal* 75 (3): 70–88.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1 (5): 206–215.

Schmidt, P.S., U. von Arx, A. Schrimpf, A.F. Wagner, and A. Ziegler. 2019. Common risk factors in international stock markets. *Financial Markets and Portfolio Management* 33: 213–241.

Stambaugh, R.F., J. Yu, and Y. Yuan. 2015. Arbitrage asymmetry and the idiosyncratic volatility puzzle. *The Journal of Finance* 70 (5): 1903–1948.

Stambaugh, R.F., and Y. Yuan. 2017. Mispricing factors. *The Review of Financial Studies* 30 (4): 1270–1315.

Tensorflow. 2020. TensorFlow. https://www.tensorflow.org/?hl=de.

Timmermann, A. 2006. *Forecast combinations. Handbook of Economic Forecasting* 1: 135–196.

Tobek, O. and M. Hronec. 2020. Does it pay to follow anomalies research? Machine learning approach with international evidence. *Journal of Financial Markets*, pp. 1–63.

Ye, J., R. Janardan, Q. Li, and H. Park. 2006. Feature reduction via generalized uncorrelated linear discriminant analysis. *IEEE Transactions on Knowledge and Data Engineering* 18 (10): 1312–1322.

Zednik, C. 2021. Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology* 34 (2): 265–288.

Zhou, Z.-H. 2012. *Ensemble Methods: Foundations and Algorithms*. Chapman and Hall/CRC.

Vitor Azevedo is a Professor at the Chair of Financial Management at the RPTU Kaiserslautern - Landau. He holds a Ph.D. in finance from the Technical University of Munich (TUM) and also worked as a postdoctoral researcher in the same institution from 2018 until 2021. He focuses his research mainly on empirical asset pricing, behavioral finance, and quantitative finance, including artificial intelligence and machine learning. Before joining the TUM, Vitor worked in the financial markets as a portfolio manager and as a stockbroker.

Georg Sebastian Kaiser works as a Senior Consultant in the Consumer Goods and Retail sector at Roland Berger. He holds a Master of Science in Finance and Informatics from the Technical University of Munich (TUM). During his studies, he mainly conducted research on empirical asset valuation and quantitative finance with a focus on artificial intelligence and machine learning. Before coming to TUM, Georg Sebastian completed his Bachelor of Science in International Business Administration at WHU - Otto Beisheim School of Management.

Sebastian Mueller is a Professor of Finance at TUM School of Management, Campus Heilbronn. His research is quantitative and empirically oriented, primarily focusing on asset pricing, asset management, and behavioral finance. He takes a special interest in the price formation process of financial markets, the investment decisions of market participants, and the effects of digitalization and sustainability on corporations and markets. Mr. Müller received several awards for his research which has been published in leading finance journals.