Journal of Empirical Finance 72 (2023) 532–556

Contents lists available at ScienceDirect

Journal of Empirical Finance

journal homepage: www.elsevier.com/locate/jempfin

Automated stock picking using random forests☆

Christian Breitung

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

ARTICLE INFO

Keywords:
Stock picking
Machine learning
Random forest
Portfolio optimization

ABSTRACT

We derive a stock ranking by applying a technical features-based random forest model on an international dataset of liquid stocks. Rather than predicted return, our ranking is based on outperformance probability. By applying a decile split, we find that long–short portfolios achieve Sharpe ratios of up to 1.95 and a highly significant yearly six-factor alpha of up to 21.79%. Moreover, we show that outperformance probabilities serve as a superior measure of future returns in the context of portfolio optimization. Mean–variance portfolios using this measure are less volatile and more profitable than equally- or value-weighted portfolios. Our findings are robust to firm size, regional restrictions, and non-crisis periods and cannot be explained by limits to arbitrage.

1. Introduction

Picking stocks that outperform in the future is one of the challenges asset managers face. As active investing requires accurate return estimations, various researchers proposed respective prediction models. In line with the Efficient Market Hypothesis (EMH), most of these models fail to predict future stock price movements. In a comprehensive study, Welch and Goyal (2008) show that the vast majority of suggested models not only perform poorly out-of-sample, but also in-sample.

With the increase in computation power, researchers started to apply machine learning algorithms to predict short-term stock price trends using models like support vector machines (Liu et al., 2016), deep learning (Dos Santos Pinheiro and Dras, 2017) or ensemble models (Basak et al., 2019) with promising results. More recently, Gu et al. (2020) compare the performance of neural networks and random forests with traditional models to predict future stock prices and find that machine learning models outperform traditional ones. They argue that this overperformance may be traced back to the capability of machine learning algorithms to capture non-linear patterns. Avramov et al. (2021) find a similarly strong performance for ML-driven investment strategies. While they agree with Gu et al. (2020) that the profitability likely originates in non-linear anomalies that ML models may successfully identify, they claim that these anomalies mainly occur in stocks that are subject to serious trading frictions. As a consequence, when controlling for transaction costs and excluding stocks with a low market capitalization or missing credit rating information, the performance of these strategies barely generate alpha, questioning the applicability of ML driven investment strategies. This is not surprising, given that smaller stocks are more likely to experience extreme return behavior and thus are more likely to receive higher absolute return predictions.

We contribute to the literature by suggesting a new stock ranking approach. In contrast to absolute return predictions, we suggest a random forest classification model that is less likely to rank stocks with extreme return behavior as highest (lowest). Thus, ranking-based portfolios are more in line with the preferences of risk-averse investors since they are less likely to contain

☆ I want to thank Sebastian Müller, the editor, an anonymous referee, and participants at the 2022 Quantitative Finance and Financial Econometrics and internal research seminars at the TUM School of Management for valuable comments and feedback. I am grateful to Sebastian Müller for providing the international factor and stock market dataset used in this research project.

E-mail address: christian.breitung@tum.de.

https://doi.org/10.1016/j.jempfin.2023.05.001
Received 1 October 2021; Received in revised form 17 April 2023; Accepted 3 May 2023
Available online 12 May 2023
0927-5398/

highly volatile stocks. More specifically, we calculate outperformance probabilities by relating the number of trees that classify a stock as an outperformer to the total number of considered trees and sort the stocks accordingly. By evaluating the performance of our model using a dataset of sufficiently liquid stocks,[1] we obtain highly significant five-factor alphas of up to 21.79% per year. This outperformance is robust to various dimensions and suggests that outperforming stocks may be identified using machine learning. Moreover, we show how outperformance probabilities may be used to construct mean–variance portfolios that outperform out-of-sample. These findings point towards a weak form of market efficiency.

To train a machine learning model, we need a sufficiently large amount of data. In this paper, we construct technical indicators using international stock market data from Refinitiv. Among others, these comprise various oscillators and bollinger bands, indicators that are vastly overlooked by the financial literature. We construct more than 180 indicators on an international dataset of sufficiently liquid stocks. While it is unlikely that these indicators may predict future returns individually, it could be that there exist some patterns in the cross-section that could be identified using a machine learning model.

Using this dataset, we initially construct two different random forest models using default parameters. On the one hand, we construct a random forest classifier that should predict whether a stock is an outperformer or an underperformer. On the other hand, we construct a random forest regressor that should predict the future return of a stock. We then test the hypothesis that portfolios based on an outperformance probability ranking contain more liquid, less volatile stocks.

In line with the idea that a regression model focuses primarily on the stocks with the most extreme return behavior during training, we find evidence that our classification approach is less likely to suggest investments into smaller stocks than the regression model. At the same time, we observe a substantially higher return and lower risk if we pursue long-short investments into stocks suggeted by the classification model. Since more recent return data seems to contains a lower signal-to-noise ratio, we refrain from refitting our prediction model over time.

Motivated by these findings, we identify parameters that optimize predictions on a validation set and train the random forest classifier on stock market data from 1990 to 2003. We then evaluate the performance of our model by splitting the dataset into deciles based on the predicted outperformance probability and calculate the return spread between the average return per decile and the average return of the whole dataset. We find that the return spread is largest (smallest) for stocks with the highest (lowest) probability of outperformance. Moreover, the average monthly return difference between the largest and lowest decile is positive in all considered years. We also do not observe a substantial decrease within recent years, indicating that the identified abnormal return patterns remain valid.

To test the performance of ranking-based investment strategies more extensively, we construct model-based equally- and value-weighted portfolios. We find that the returns of portfolios containing the highest-ranked stocks yield a substantially higher return and lower volatility than portfolios that are long in the entire dataset. However, this coincides with large negative skewness and high excess kurtosis of the returns, suggesting a high tail exposure. We find no significant outperformance in the top portfolio if we control for tail exposure by calculating the adjusted Sharpe ratio (*ASR*).

If investors short the stocks with the lowest ranking simultaneously, they may effectively hedge against tail risk. We observe a Sharpe ratio of up to 2.49 and adjusted Sharpe ratios of up to 4.10. Value-weighted portfolios, in general do not seem to be exposed to tail risk. We obtain a Sharpe ratio (*SR*) of 1.23 for the long and 1.56 for the long–short portfolio, respectively. These results are also robust to portfolio size. While larger portfolios generate lower Sharpe ratios on average, ranking-based portfolios generate a significantly larger SR than portfolios invested in the entire dataset. This effect is driven by both excess return and portfolio volatility.

We repeat the analysis with stocks that have a high market capitalization to ensure that the results are replicable in the absence of small and mid-cap stocks. Even without short selling, we achieve a *SR* of up to 1.21, substantially larger than the one we obtain for a portfolio that is long in all stocks above a $10 billion market capitalization.

To test whether any commonly known risk factors may explain the outperformance, we run a five-factor model using developed factors provided by Kenneth R. French. We obtain highly significant yearly alphas of up to 21.79% before transaction costs for equally-weighted portfolios and 15.38% for value-weighted ones. The t-statistics are 6.56 and 4.36 respectively. These alphas barely change if we control for momentum and country-specific effects.

Next to stock picking, asset managers also try to determine weights that optimize the return-to-volatility ratio of their portfolios, a problem which has been heavily investigated within academia since the mid of the last century (Ban et al., 2018; DeMiguel et al., 2009; Jagannathan and Ma, 2003). When Markowitz (1952) proposed the concept of the mean–variance portfolio (MeanVP), practitioners quickly discovered a poor out-of-sample performance, as the approach requires good estimates of future returns which typically lack in practice. To overcome this issue, some suggest investing in the minimum variance portfolio (MinVP), as it does not consider expected returns in the optimization. However, this results in a trade-off between lower volatility and lower returns, often resulting in portfolios with lower Sharpe ratios than equally-weighted portfolios in out-of-sample tests.

We suggest an alternative proxy for future returns, a linearly shifted version of our outperformance probabilities that is centered around zero. When constructing MeanVPs based on a random selection of large stocks using our future return measure, we observe a substantial increase in the performance. On average, without short selling, the portfolio yields a SR of 1.02 which is not only substantially higher than the SR of a MeanVP using past returns, but also higher than equally or value-weighted investments. With short selling, we notice a similar pattern. On the one hand, we observe dramatically lower volatility. On the other hand, we see a higher return if we use outperformance probabilities as a measure of future return. This suggests that investors can improve their portfolio performance by constructing mean–variance portfolios using ML-driven performance forecasts.

[1] We define stocks as sufficiently liquid if they are traded on at least fifteen days per month and if their market capitalization exceeds $300 million.

We consider various dimensions to investigate which factors drive the profitability of our model-based investments. First, we test whether specific regions or countries drive the results. For example, it could be that the portfolios contain mispriced stocks that operate in markets that were not accessible to foreign investors during the evaluation period, for instance, Chinese or Indian stocks. We therefore apply model-guided decile splits on the region and country level and construct value-weighted portfolios to control for limits to arbitrage simultaneously. Overall, we observe comparably high five-factor alphas across different countries and regions, suggesting that such trading frictions are unlikely to explain the results of our global portfolios.

Second, we investigate the argument of Avramov et al. (2021) that ML-driven investment strategies particularly excel during crisis periods. We split the evaluation period into crisis and non-crisis periods. Indeed, we obtain substantially lower and less significant alphas in periods of low volatility and low bid–ask spreads. Value-weighted long portfolios yield a mere 3.86% annual five-factor alpha. We also find that value-weighted investments into stocks with low model rankings do not yield significantly negative alphas in non-crisis periods.

Third, we consider whether transaction costs might explain the price inefficiencies we observe. We therefore calculate the net profitability of our investment portfolios by estimating trading costs using bid–ask spreads. We observe transaction costs of around two percentage points, leading to net returns of around 12% for an equally-weighted investment into high-ranked stocks with a market capitalization above $10 billion. Thus, we argue that transaction costs are unlikely to explain why stocks are not priced more efficiently.

Fourth, we test whether the studied portfolios load on some unobserved risk factors by constructing RP-PCA factors as suggested by Lettau and Pelger (2020). We find some evidence that the constructed portfolios indeed load on some unobserved risk factors. However, if we consider an equally-weighted long–short investment into the hundred highest (lowest)-ranked stocks, we nevertheless observe a significant alpha of more than 6% annually, suggesting that unobserved risk factors may not entirely explain our results.

Overall, we find that the combination of limits to arbitrage and the exposure to unobserved risk factors explain a significant share of the abnormal return. We trace back the remaining alpha to the technical indicators that were largely overlooked in the finance literature. We argue that investors might have been unaware of the patterns identified by our model and therefore did not timely correct these mispricings.

The remainder of this paper is organized as follows. First, we introduce the identification strategy in Section 2. We then present the dataset and some constructed features in Section 3. We then construct the model in Section 4 and evaluate the results in Section 5. Finally, we apply robustness checks in Section 6 and conclude our findings in Section 7.

2. Identification strategy

Even though markets are generally perceived as efficient, there exists a large literature that documents the existence of anomalies across the globe. While some of them are interpreted as risk factors (e.g. Fama and French (2015)), others are associated with behavioral biases (e.g. disposition effect, home bias, and limited attention of investors). Among other techniques, machine learning models may detect abnormal return patterns in stock market data. Just recently, Cakici et al. (2022), Azevedo et al. (2022) and Hanauer and Kalsbach (2023) construct superior portfolios using various machine learning models, questioning the efficiency of international markets. However, as outlined by Avramov et al. (2021), machine learning models that predict future performances tend to suggest investments into less liquid, volatile stocks of firms under financial distress. This is unsurprising, given that these machine learning models focus on stocks with the most extreme return during training. Since trading these stocks is more costly, investors often cannot realize the observed performance in practice.

We break down the regression into a classification problem to mitigate this issue. More specifically, we predict whether a stock outperforms a certain performance benchmark. By doing so, our model should focus less on stocks that experienced extreme return behavior in the past. Rather than fixing some absolute return threshold, we suggest labeling stocks according to their relative performance in the consecutive month. We define a stock as an *outperformer* (labeled 1) if its next month's stock return is higher than the median next month's stock return of all stocks in the dataset and an *underperformer* (labeled 0) otherwise.[2] By doing so, we ensure a balanced training set.

In general, there are two major forms of stock market analysis. Technical analysts try to predict future stock prices by analyzing historical prices and trade volumes (Lo et al., 2000) while fundamental analysts believe good investment opportunities may be discovered by analyzing the fundamental ratios of a company (Abarbanell and Bushee, 1998). Additionally, with the latest advancements in Natural Language Processing, researchers started to investigate whether textual data may reveal information regarding the future performance of equities (Cohen et al., 2020; Breitung and Müller, 2022).

In this paper, we entirely focus on technical indicators that are often manually applied by technical analysts but largely overlooked in academia. We calculate technical indicators like *Moving Average*, *Relative Strength Index*, and others. We also added features like volatility, skewness, and kurtosis. In total, we calculate 96 features that can broadly be categorized as either momentum, trend, volatility, or volume-based.[3] To get rid of the price levels some indicators contain by construction, we further relate all indicators to their last available share price such that a comparison among different stocks is possible. The complete list of calculated features may be found in Tables A.1 and A.2 in the Appendix.

We construct the training and test set as follows. First, we filter out insufficiently liquid stocks from our dataset of international stocks by dropping those with on average less than 15 trading days per month in the previous year and a market capitalization of

[2] Note that we do not incorporate the volatility of a security during the labeling process. The reason is that a portfolio of highly volatile stocks might experience low volatility if the correlation among the stocks is sufficiently low.

[3] Note that some of the features could be associated with more than one category, therefore we decided to allocate those to a fifth category called "other".

less than $300 million. We refrain from adding a liquidity threshold for the average daily traded amount as the variable is often missing in our dataset and tends to be highly positively correlated with market capitalization. Second, we calculate the technical indicators at the end of each month and assign a one (zero) label to the stocks with an above-median (below-median) return next month. We apply this to all months ranging from 1990 until the end of 2003. Finally, we concatenate these observations and obtain a large dataset with nearly three million observations. The test set is created similarly using stock market data from 2004 to 2018.

As a next step, we need to select an appropriate machine learning algorithm. In principle, we may differentiate between models capable of identifying non-linear relationships and those that are not. One possibility to test whether a problem requires a non-linear algorithm is to test for the existence of a separating hyperplane (De la Fuente, 2000) between the two classes we try to differentiate among. If such a hyperplane exists, we may argue that some linear algorithm might solve the problem.

To do so, we follow Basak et al. (2019) and reduce the dimensionality of the feature set to two dimensions using the so-called Principal Component Analysis (PCA) and plot the convex hulls of both classes. We find that the convex hulls intersect such that a separating hyperplane does not exist.[4] Thus, we need to employ a non-linear algorithm for the classification problem. Among others, neural networks, support vector machines, and random forests are the most prominent ones.

In this paper, we construct a random forest suggested by Breiman (2001). Random forests combine multiple *decision trees* whereas each tree is trained on a random subsample of the dataset. A decision tree itself is capable of learning non-linear relationships but tends to overfit the training data. Random forests may reduce this problem by averaging the results of multiple decision trees (Horning et al., 2010). Random forests have been successfully applied within quantitative finance (Emerson et al., 2019). They are robust, automatically handle missing values and work well on discrete and continuous variables (Obthong et al., 2020). Moreover, we may interpret the relation between the number of trees that predict an outperformance and the total number of trees in the forest as an *outperformance probability*, which may be used to construct a stock ranking that active investors may rely on.

3. Data

We collect international stock market data from Refinitiv (Datastream), covering stocks from 68 countries between 1990 and 2018. All technical indicators are calculated by considering not more than one year of previous stock price data. We restrict the dataset to sufficiently liquid stocks as defined in Section 2. To reduce the probability that the dataset contains data errors that significantly affect our results, we further drop the 0.1% of stocks with the highest (lowest) return each month. In total, our dataset comprises 18 973 stocks across 67 countries. Note that not all of those stocks fulfill the required liquidity constraints throughout the entire investment horizon. We therefore restrict investments to those stocks that match the criteria at the month of investment. As a result, the number of possible investment opportunities ranges between 3000 and 10 000 over time.

Table 1 provides an overview of the stocks contained within our dataset. We observe that most stocks are traded in the US, China, and Japan. However, our dataset also contains stocks from smaller countries like Bahrain or Serbia. The average market capitalization within our dataset is $3.74 billion. We also provide the average monthly return on the country level and observe a substantial heterogeneity.

4. Model construction

4.1. Model construction

As argued in Section 2, portfolios constructed based on absolute stock return forecasts tend to suffer from several weaknesses that are only exacerbated in the context of machine learning. In particular, regression models tend to focus on the tiniest, illiquid, and volatile stocks during training as they usually have a higher absolute return in the subsequent month. Consequently, ranking-based portfolios should be typically associated with higher volatility, larger maximum drawdowns, and a higher kurtosis of the portfolio returns, which is not in line with the preferences of risk-averse investors.

To test whether our alternative suggestion to rank stocks according to their *outperformance probability* probability may indeed mitigate these effects, we compare equally-weighted long–short investments based on our classification approach with investments based on a random forest regressor that predicts future returns.[5]

We present performance metrics like *Mean*, *volatility*, *maximum drawdown*, *skewness* and *kurtosis* for portfolios based on non-refitted models from 2004 until the end of 2018 in Panel A of Table 2. As predicted, we find evidence that a classification-based investment is less biased toward smaller stocks. The average market capitalization in the long–short portfolio is approximately $500 million higher than the average in regression-based investments. We also observe lower volatility, maximum drawdown, and kurtosis, suggesting that ranking according to outperformance probability leads to less risky portfolios. Despite the lower risk, we observe a portfolio return of 18.92%, which is roughly 3.5% higher than a regression-based portfolio. This translates into a Sharpe ratio of 1.95.

Some might suspect that in contrast to training a model once, refitting a model over time might yield better prediction results since relationships between variables do not necessarily have to be constant over time. Irrespective of the effect on the model's

[4] The convex hulls are presented in Fig. A.2 and may be found in the Appendix.

[5] Note that to avoid any systematic bias, we use default parameters, namely a forest size of 100 trees and no restrictions on the tree depth or the maximum features a decision tree may use to split a node.

Table 1
Summary statistics by country.

Country	Stocks	MV	Ret.	Country	Stocks	MV	Ret.
Global	18 973	3.74	0.64				
Argentina	30	1.51	2.52	Malaysia	203	1.96	0.61
Australia	587	2.56	0.53	Mauritius	11	0.8	0.12
Austria	60	2.81	0.77	Mexico	88	3.98	0.92
Bahrain	6	1.09	−2.83	Morocco	31	2.02	0.64
Bangladesh	40	1.01	0.33	Netherland	87	5.45	0.68
Belgium	74	5.06	0.58	New Zealand	64	1.39	0.97
Brazil	204	3.79	0.77	Nigeria	39	2.04	0.15
Bulgaria	23	0.58	−0.08	Norway	177	4.08	0.67
Canada	896	2.87	0.41	Oman	25	1.11	0.14
Chile	48	3.24	0.87	Pakistan	63	1.24	1.13
China	3 205	2.31	0.80	Peru	32	1.4	0.93
Colombia	31	4.86	1.05	Philippines	76	2.33	1.09
Croatia	31	1.43	−0.14	Poland	141	2.09	0.27
Czech Republic	20	5.23	1.24	Portugal	35	3.53	0.37
Denmark	82	2.56	0.86	Qatar	43	4.39	0.7
Egypt	82	1.39	1.05	Romania	22	2.1	0.91
Estonia	8	0.61	−0.16	Russia	148	8.65	0.94
Finland	85	4.80	0.78	Serbia	7	0.73	0.39
France	353	6.19	0.71	Singapore	264	2.08	0.28
Germany	431	7.65	0.75	Slovenia	11	0.69	−0.31
Greece	89	1.93	−0.35	South Africa	156	3.17	1.25
Hong Kong	825	5.42	0.30	Spain	144	7.56	0.33
Hungary	14	3.42	0.86	Sri Lanka	12	0.78	0.77
India	639	3.36	0.97	Sweden	252	2.91	1.1
Indonesia	192	2.64	0.93	Switzerland	163	7.13	0.64
Ireland	30	2.92	0.58	Taiwan	642	1.97	0.3
Italy	215	4.68	0.30	Thailand	203	2.01	0.61
Japan	1 237	3.18	0.58	Tunisia	19	0.58	0.81
Jordan	20	1.53	0.12	Turkey	136	2.4	1.18
Kazakhstan	14	1.83	−0.36	USA	4432	5.12	0.68
Kenya	22	1.08	0.47	Ukraine	49	1.78	−1.03
Korea	660	2.97	0.43	United Kingdom	726	4.66	0.61
Kuwait	122	2.03	−0.60	Vietnam	79	1.54	0.67
Lithuania	18	0.59	0.68				

This table summarizes our dataset of liquid stocks. We count the number of stocks (*Stocks*), the average market capitalization in billion USD (*MV*) and the average monthly return (*Ret.*) on the country level.

Table 2
Model comparison.

Portfolio	Mean	Vol	SR	MD	Skew	Kurt	MV
Panel A: No refit							
Class.	18.92	9.10	1.95	−5.54	0.29	1.35	3.37
Reg.	15.44	10.84	1.31	−11.69	0.20	2.37	2.86
Panel B: Mov. refit							
Class.	12.37	11.86	0.94	−12.68	−0.70	2.46	3.35
Reg.	14.43	11.84	1.12	−13.13	−0.10	3.75	2.68
Panel C: Early vs. Late							
Early	16.59	9.92	1.67	−5.54	0.28	1.40	3.47
Late	9.33	11.30	0.82	−12.34	−1.23	3.28	3.37
Late-FC	5.29	13.00	0.40	−14.37	−1.36	4.63	3.31
Late-No FC	13.78	11.15	1.23	−11.25	−0.54	2.11	3.42

This table reports the performance of equally-weighted long–short portfolios based on different types of models trained with default parameters. We split the dataset into deciles based on the probability of outperformance (*Class.*) and predicted returns (*Reg.*). In Panel A, we do not refit the models over time. In Panel B, we refit the models using a moving window of the most recent 15 years of data. Panel A and B are evaluated on fifteen years of monthly stock market data (2004–2018). Panel C represents three classification models that are not refitted but span different training horizons. *Early* is trained from 1990 until the end of 2003, *Late* is trained from 2004 until mid of 2011, *Late FC* is trained on the financial crisis years 2008 and 2009, and *Late-No FC* is trained on the same horizon of *Late* but excluding financial crisis years. To avoid bias, we evaluate the performance of the models from mid-2011 until the end of 2018. *Mean* represents the average yearly return, *Vol* the average yearly volatility and *SR* the Sharpe ratio of a portfolio. We further calculate the maximum drawdown *MD*, the skewness *Skew*, the kurtosis *Kurt*, and the average monthly market capitalization of a firm in billion-dollar *MV*.

accuracy, refitting the model means losing the ability to assess whether the captured patterns remain valid over time. To shed some light on the difference in accuracy, we compare the performance of our baseline model using a rolling window approach.[6]

According to Panel B, refitting the model leads to a reduction in the Sharpe ratio in the case of the classification-, as well as the regression-based portfolios. This is surprising, given that more recent data is usually interpreted as more relevant. In principle, there exist two potential explanations for these phenomena. First, it could be that the signal-to-noise ratio decreased over time. As markets get more efficient, it could be that the model is more likely to misinterpret noise as patterns. Second, it is possible that there exist different predictive patterns in crisis and non-crisis periods such that a model trained on crisis periods might yield a lower out-of-sample prediction power.

We test the first hypothesis by constructing two different classification models. The first model (*Early*) is trained on earlier data (1990–2003). The second one (*Late*) is trained on more recent data (from 2004 until mid-2011). Both portfolios are evaluated from mid-2011 until 2018 using monthly rebalancing. Indeed, we observe a substantial performance drop in the model that is solely trained on more recent data. While a long–short portfolio based on early data achieves a Sharpe ratio of 1.67, a portfolio that follows the more recent model achieves a substantially lower Sharpe ratio of 0.82. Even though the more recent model is trained on fewer periods, the number of available stocks per year is substantially larger such that differences in the number of observations may not drive these results.

While the results point towards a decrease in the signal-to-noise ratio, we cannot yet rule out the hypothesis that the performance decrease is caused by the patterns learned during the financial crisis. We therefore evaluate two other models, a model that is trained on more recent data excluding the financial crisis (*Late-No FC*) and the *Late-FC* model that is solely trained on the financial crisis.[7]

We find strong evidence in favor of the second hypothesis. The exclusion of the financial crisis years from the training set substantially increases prediction power (*Late-No FC*). We find that the Sharpe ratio increases from 0.82 to 1.23. This is counterintuitive for machine learning models, as fewer observations are usually associated with a performance reduction. Furthermore, a model solely trained on the financial crisis years generates a lower Sharpe ratio of 0.4. As a robustness check, we train and evaluate annual models and observe a lower out-of-sample prediction power for models trained on more recent or crisis years.[8] Potentially, this behavior might be mitigated if we include additional variables that represent the overall market state, such as stock market index volatility and macroeconomic variables. Future research could replicate our findings and test whether incorporating market state variables can effectively resolve this issue.

Our results suggest that more recent data is not necessarily more valuable in predictive pattern identification. We thus decide to train a non-refitted classification model on stock market data from 1990–2003 and evaluate it on stock market data from 2004 until 2018.

4.2. Parameter optimization

The models we have compared were trained with default parameters to avoid systematic bias. To optimize the performance of a random forest model, we may adapt different parameters. We randomly divide the training set into two equally sized subsets to identify the optimal parameter settings. Then a random forest model is fitted to one of them and evaluated on the other using different parameter values.[9] Among the set of possible parameters, the most relevant is the maximum number of features that may be considered to split a node and the number of decision trees in the model. Therefore, we focus on optimizing these parameters within this section.

Suppose we follow the abovementioned approach and calculate the accuracy of the evaluation set for different tree depths. In that case, we observe that a too-strong restriction of the maximum tree depth leads to a performance reduction. With an increase in depth, the model's accuracy on the training set converges to 100%, which is a sign of overfitting. Therefore, we set the maximum tree depth to 23 as it seems two maximize the model's performance while limiting the probability of overfitting.[10]

Another important parameter is the maximum number of features available per node. By default, each decision tree may consider all features to split a node. To further reduce overfitting, it is possible to restrict the number of features a decision tree may consider to find the best split. However, we find that restricting the number of features that may be assessed at each node does not lead to an increase in the performance of the model.[11]

Finally, we have to determine the size of the forest. In general, more trees should lead to better classification results even though this effect diminishes with an increasing forest size (Breiman, 2001). A common strategy is to plot the Out-of-Bag (OOB) error rate for different forest sizes to check when the generalization error converges. More precisely, we fix a specific forest size and fit the model on a subsample of the training data. Then, the model is applied to previously unseen training data to calculate the mean prediction error, which is the OOB error rate. We find that the model converges when the number of estimators is approximately 2000, which we consequently set as forest size.[12]

[6] We retrain the model by extending the training period by the most recent year and dropping the oldest.
[7] We define 2008 and 2009 as financial crisis years.
[8] The results may be found in Table A.3 in the Appendix.
[9] Note that we keep all other parameters at their default values.
[10] The results are presented in Fig. A.1 in the Appendix.
[11] The results are presented in Fig. A.3 in the Appendix.
[12] The results are presented in Fig. A.4 in the Appendix.

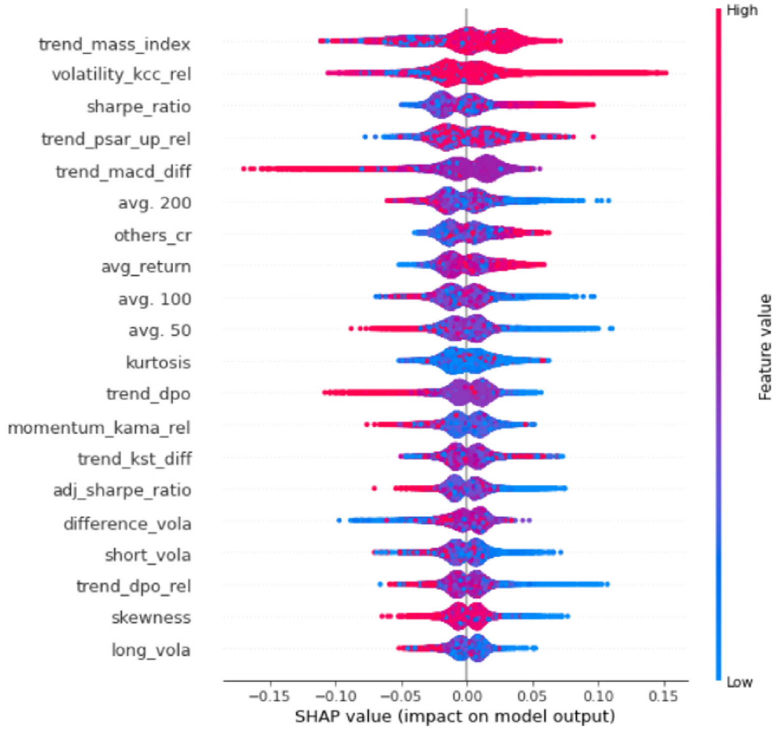

Fig. 1. Shapley value of outperformers as identified by a global model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5. Model evaluation

5.1. Feature evaluation

In order to assess which of the features are most important, we extract feature importances from the model. Feature importances are obtained by counting the nodes where a specific feature was used for splitting and relating it to the total amount of nodes. Since focusing on 96 different features is not feasible, we calculate the feature importance for each category. Note that simply comparing feature importances might be misleading if some categories contain many more features than others. Therefore, we also relate the number of features per category with the total number of features.

We further calculate the feature importances of random forest models trained on stocks from certain regions only. This enables us to test whether feature importances vary across regions. Trend-based features seem most relevant for our international model. In around 36.55% of all cases, the random forest model chooses a trend-related feature to split a node. However, given that 38.67% Instead, momentum and volume factors seem to be favored more often than expected. In total, we obtain a relatively similar image for regionally restricted models. Even though they are trained on different subsamples, the obtained feature importances are mainly in line with the ones from the international model. We observe the largest difference for volume-based features in Europe. Here, volume-based features seem to be less important than in other regions.[13]

A more sophisticated approach to identifying relevant features is by calculating Shapley values, a concept borrowed from the area of cooperative Game Theory (Shapley, 1951). Essentially a Shapley value determines the *fair* payoff to the players based on their contribution to the outcome. Lundberg and Lee (2017) translate this approach to the area of Machine Learning by suggesting SHAP (SHapley Additive exPlanations). Here, features may be seen analogously to the participants in a cooperative game. Each feature contributes to the final prediction of a model. The Shapley value thus represents a feature's individual contribution to a model's final predictions. The higher the value, the higher the contribution of a feature to the final prediction and thus, its importance.

To reduce the problem of colinearity, we drop features that show a higher correlation than 0.9, which reduces our feature space from 181 to 122 dimensions.[14] Given that the computation time increases exponentially in the number of features (Molnar, 2020), we further reduce the amount of features by keeping only those twenty features that are most often considered to split a node.

Fig. 1 shows the Shapley values for a large number of outperformers in our training set. Observations with a blue (red) color represent low (high) feature values. Due to the limits of computational power, we calculate Shapley values on 10,000 randomly

[13] The feature importances are presented in Table A.4 in the Appendix.

[14] Note that we consider two versions, an absolute and a relative one, for most of the 96 features such that we consider 181 features in total.

drawn outperformers rather than the entire dataset. There are multiple observations to be made. First, we observe positive Shapley values for stocks with a lower average daily return in the last fifty (hundred, two hundred) days. The same holds for the average daily volatility of a stock in the short term and, to a lesser extent, in the long term. More importantly, the difference between a stock's short-term and long-term volatility seems relevant. If the stock was more volatile in the short than the long term, the model is more likely to classify the stock as an outperformer. All these characteristics increase the likelihood that our model will classify a stock as an outperformer. We further observe a strong impact of a trend indicator called *Moving Average Convergence Divergence* (MACD). It is calculated by subtracting a long-term moving average from a short-term one. We find that stocks with a more positive trend in the short- than long-term are less likely to be classified as outperformers.

We additionally investigate whether there exist regional differences concerning feature importance and present the corresponding figures Fig. A.5 and Fig. A.6 in the Appendix. First, we find that the short-term volatility of stocks has a high impact in the US and the emerging market model. In both models, low volatility increases the probability that a stock is classified as an outperformer. Second, higher values for the negative volume index, a measure that only considers percentage changes of days where the trading volume decreases, lead to increased profitability that a stock is classified as an outperformer in both regions. The assumption behind this is that the percentage of *smart* money should be higher in periods of lower trading volume.[15] Third, trend variables seem to be more influential in emerging markets. Here, we observe at least five trend indicators among the most influential features. Stocks with below-average returns are less likely to be classified as outperformers. In the US, we do not observe a similarly clear pattern.

Overall, despite differences in the ranking of the variables, regionally restricted models seem to learn similar patterns, which suggests that an internationally trained model should perform well irrespective of the region it is applied to. These results strongly indicate that the model identified many abnormal return patterns present in different markets across the globe.

5.2. Portfolio evaluation

Researchers usually evaluate the performance of a classification model by considering its accuracy, precision, and recall. However, considering that the amount of correct classifications is less critical in our setting, we focus on the actual return difference for correct and wrong classifications. A high accuracy could still indicate that ranking-based investments are not profitable if the underperformance of wrong classifications is sufficiently larger than the outperformance of correct classifications. Therefore, we calculate the *return spread*, which is the deviation between the average stock return of the dataset and the mean return of the stocks that share similar class probabilities.

According to Fig. A.7, which may be found in the Appendix, the *return spread* is positive for stocks with a high probability of outperformance. In contrast, it is strongly negative for stocks with only low probabilities. We also observe that stocks, which have a probability of at least 50%

To study the model's performance more extensively, we construct model-based long, short, and long–short portfolios and evaluate their performance.

According to Table 3, the average returns of the top portfolios are substantially larger than the average returns of the bottom portfolios. The difference is around 41 percentage points annually for a portfolio of hundred equally-weighted stocks. While the mean return of a portfolio that is long in the best ten (twenty) percent of stocks is lower, it remains substantially higher than the 11.02% annualized average return we observe for a portfolio that is equally invested in all stocks that match our liquidity criteria (*Total*). We further find that the volatility of the portfolios with the highest outperformance probabilities is between three to five percentage points lower than the equally-weighted benchmark. A similar pattern may be observed for the maximum drawdown, suggesting that stocks that perform well according to our model are also more resilient during crises. Consequently, we observe Sharpe ratios of up to 1.81, which is almost three times as high compared to the *Total* portfolio.

However, considering the negative skewness and the high kurtosis, it seems that the portfolio returns do not follow a normal distribution. Since the Sharpe ratio does not consider the tail risk of these portfolios, we follow the suggestion of Pezier (2004) and consider an adjusted version of the Sharpe ratio:

$$ASR = SR(1 + \frac{S}{6}SR - \frac{(K-3)}{24}SR^2) \tag{1}$$

where *ASR* is the *Adjusted Sharpe ratio*, *SR* is the Sharpe ratio, *S* is the skewness and *K* is the kurtosis. Controlling for tail risk, we find that the outperformance in the top portfolios disappears. These results align with the idea that investors charge a premium for being exposed to tail risk.

Considering investments into the lowest ranked stocks, we observe an annualized average return of −19.90% for the *B-100* portfolio. This portfolio's volatility and maximum drawdown is also higher, leading to a negative Sharpe ratio of −0.97, significantly smaller than the Sharpe ratio of the *Total* portfolio. If we construct larger portfolios, the performance difference decreases but remains negative. What is striking is that these portfolios have a lower negative skewness and kurtosis of around 3, indicating that investors are not exceptionally exposed to tail risk here.

Given the performance of the top and bottom portfolios, it is not surprising that we observe extraordinarily high profits if we combine long and short investments. For example, an equally-weighted long–short portfolio with a hundred long and short positions yields an annualized average return of 40.88%. Since the yearly volatility of such a portfolio is not too high (15.97%), this translates

[15] For more information, see https://school.stockcharts.com/doku.php?id=technical_indicators:negative_volume_inde.

Table 3
Top, bottom and top-bottom portfolios metrics.

Portfolio	Mean	Vol	SR	MD	Skew.	Kurt.	ASR	MV
Panel A: EW:								
Total	11.02	16.08	0.61	−22.55	−1.16	4.21	0.53	3.66
T-100	21.00	10.96	1.81	−18.59	−1.68	11.50	−1.21	3.86
T-dec	18.71	12.09	1.45	−19.33	−1.46	8.72	0.21	5.38
T-quint	18.04	12.96	1.30	−20.26	−1.44	7.69	0.47	5.35
B-100	−19.89	21.84	−0.97	−30.48	−0.64	3.14	−1.06	1.24
B-dec	−5.16	19.18	−0.33	−26.61	−0.92	3.18	−0.35	1.76
B-quint	−0.57	18.75	−0.09	−25.30	−0.94	2.98	−0.09	2.07
TB-100	40.88	15.97	2.49	−9.42	0.07	0.62	4.10	2.55
TB-dec	23.87	11.64	1.95	−8.29	0.19	0.62	2.81	3.57
TB-quint	18.62	10.47	1.66	−6.95	0.26	0.86	2.19	3.71
Panel B: VW:								
Total	8.35	14.52	0.49	−20.19	−1.17	3.94	0.44	3.66
T-100	15.79	11.85	1.23	−15.18	−1.08	3.92	0.89	3.86
T-dec	13.72	12.08	1.04	−15.40	−0.92	3.72	0.84	5.38
T-quint	12.83	12.92	0.90	−16.98	−1.01	3.92	0.74	5.35
B-100	−16.74	24.28	−0.74	−27.49	0.15	3.13	−0.72	1.24
B-dec	−4.09	19.37	−0.27	−24.77	−0.62	2.59	−0.28	1.76
B-quint	−1.08	18.01	−0.13	−23.17	−0.87	2.63	−0.13	2.07
TB-100	32.53	20.13	1.56	−18.71	−0.69	1.94	1.45	2.55
TB-dec	17.81	13.03	1.28	−8.95	0.07	0.53	1.51	3.57
TB-quint	13.91	10.63	1.20	−6.72	0.48	1.71	1.41	3.71

We present performance metrics for equally- and value-weighted investments into the highest (T) and lowest (B) ranked stocks. We re-balance the portfolios monthly and evaluate their performance from 2004 until the end of 2018. The portfolios contain, on average, 100, 600 (decile), and 1200 (quintile) stocks and are re-balanced monthly. The *TB* portfolio comprises long investments into the top and short investments into the bottom portfolio. *Mean* represents the average yearly return, *Vol* the average yearly volatility and *SR* the Sharpe ratio of a portfolio. We further calculate the maximum drawdown *MD*, the skewness *Skew*, the kurtosis *Kurt*, the adjusted Sharpe ratio *ASR* and the market capitalization in billion-dollar *MV*.

into a SR of 2.49. More importantly, the outperformance may not be explained with exposure to tail risk. We observe a positive skewness and an exceptionally low kurtosis of 0.86. Consequently, the *ASR* is as high as 4.1.[16]

We also construct value-weighted portfolios to ensure that the results persist for larger stocks. As expected, the average return of the top portfolio decreases to 15.79% per year. While the volatility is slightly higher, the exposure to tail risk is substantially reduced. We observe an *ASR* of 0.89 (0.84, 0.74) for investments into the hundred (ten percent, twenty percent) highest-ranked portfolios, which is more than twice as high as the value-weighted benchmark. We thus conclude that tail risk exposure is mainly induced by the returns of smaller stocks rather than being related to the overall investment strategy. Again, long–short portfolios yield the highest average returns, Sharpe ratios, and adjusted Sharpe ratios.

We also include a more conservative test to control for limits to arbitrage by repeating the above calculations using stocks with a large market capitalization only. We use two different size benchmarks here. On the one hand, we increase the minimum market capitalization to $10 billion. On the other hand, following the argumentation that an absolute market capitalization threshold is not ideal given that the value of the Dollar varies over time, we also apply a relative size threshold. For the US stock market, Chen et al. (2020) roughly obtain the S&P500 constituents if they consider stocks that represent more than 0.01% of the total market capitalization. We therefore apply the same threshold in international markets.

According to Panel A in Table 4, an equally (a value) weighted long–short investment into the 10% highest ranked stocks with a large market capitalization yields an average return of 14.67% (12.41%), which is larger than the average stock return in the dataset. At the same time, we observe lower volatility, leading to superior Sharpe ratios of up to 1.11. The adjusted Sharpe ratios are similar since investors are not particularly exposed to tail risk here. Investments in large underperformers yield a below-average stock return, irrespective of the stock weighting. We also observe higher volatility and a stronger maximum drawdown. Consequently, we again observe substantially higher values for SR and ASR compared to the respective benchmarks.

We obtain similar results if we consider the relative size threshold in Panel B. More specifically, we obtain a Sharpe ratio of 1.21 for equally-weighted investments into the 10% highest ranked large firms and a Sharpe ratio of −0.14 for a portfolio comprising the lowest ranked stocks. These results strongly indicate that the identified patterns in the data are also present in highly liquid firms.

To rule out the possibility that our dataset suffers from a selection bias, we compare our equally- and value-weighted portfolios with randomly constructed portfolios (bootstrapping).[17] We do not find any results pointing in this direction.

As a next step, we test whether some well known pricing factors may explain the observed outperformance. We therefore run different factor regression models using developed, global, and country-specific factor data. We obtain the developed factor data

[16] We observe similar effects for larger portfolios.

[17] The results of this bootstrapping exercise may be found in Fig. A.8 in the Appendix.

Table 4
Large capitalized stocks.

Portfolio	Mean	Vol	SR	MD	Skew.	Kurt.	ASR	MV
Panel A: Abs.								
Total EW	8.23	14.25	0.49	−19.38	−1.13	3.66	0.44	31.43
T-EW	14.67	12.15	1.11	−13.50	−0.44	2.40	1.05	30.95
B-EW	1.28	19.26	0.00	−23.04	−0.37	3.16	0.00	27.93
TB-EW	13.38	13.59	0.90	−12.54	−0.12	1.79	0.92	29.44
Total VW	6.88	13.95	0.41	−18.88	−1.10	3.49	0.38	31.43
T-VW	12.41	12.42	0.90	−11.83	−0.49	1.54	0.88	30.95
B-VW	1.73	17.91	0.03	−22.50	−0.49	2.62	0.03	27.93
TB-VW	10.68	13.28	0.71	−10.31	0.01	1.65	0.73	29.44
Panel B: Rel.								
Total EW	9.18	14.90	0.54	−20.77	−1.18	4.05	0.48	12.33
T-EW	15.96	12.17	1.21	−18.16	−1.18	6.34	0.68	12.91
B-EW	−1.37	18.89	−0.14	−23.05	−0.68	2.12	−0.14	9.18
TB-EW	17.33	12.33	1.31	−6.55	0.38	0.86	1.62	11.04
Total VW	7.75	14.25	0.46	−19.57	−1.13	3.74	0.42	12.33
T-VW	13.35	12.19	1.00	−14.31	−0.81	2.96	0.87	12.91
B-VW	−1.90	18.42	−0.17	−22.72	−0.60	2.37	−0.17	9.18
TB-VW	15.25	12.48	1.13	−8.30	0.13	0.98	1.28	11.04

We present performance metrics for the equally- and value-weighted top (T) and bottom (B) portfolios that consist of long investments into the highest (lowest) ranked stocks, according to our random forest classifier. We re-balance the portfolios monthly and evaluate their performance from 2004 until the end of 2018. The *TB* portfolio comprises long investments into the top and short investments into the bottom portfolio. Panel A shows the results for investments into stocks with a market capitalization of at least \$10 billion. Panel B shows the results for investments into stocks that represent at least 0.01% of the total market capitalization. *Mean* represents the average yearly return, *Vol* the average yearly volatility and *SR* the Sharpe ratio of a portfolio. We further calculate the maximum drawdown *MD*, the skewness *Skew*, the kurtosis *Kurt*, the adjusted Sharpe ratio *ASR* and the market capitalization in billion-dollar *MV*.

from the website of Kenneth R. French. The global factor data is constructed by applying the methodologies described on Kenneth R. French's website to our dataset of international stocks as much as possible. We also create portfolio-specific factor data by grouping factor data on the country level concerning the country exposure of a portfolio. By doing so, we control for a potential overweighting of stocks from countries that generally showed a strong performance.

Panel A in Table 5 shows the exposure of the different portfolios (decile split) to the developed factors available on the homepage of Kenneth R. French. Next to the expected high correlation with the market, we find a strong correlation with the SMB factor in the equally-weighted portfolio, suggesting that the portfolio contains smaller stocks that drive the portfolio's profitability. We also observe a negative exposure to the CMA factor in the bottom portfolio. More importantly, these factors may not fully explain the out- and underperformance of any of the portfolios. We observe a highly significant yearly positive five-factor alpha of 11.79% (6.79%) for an equally- (value-) weighted portfolio that is long in the 10% of stocks with the highest probability of outperformance. On the contrary, investments into the 10% of stocks with the lowest probability of outperformance yield a significant negative alpha of −11.25% and −9.84%, respectively. If we construct long-short portfolios, we obtain even higher alphas of 21.79% (equally-weighted) and 15.38% (value-weighted) with a t-statistic of 6.56 and 4.36, respectively. Note that we so far did not control for the momentum effect. Moreover, since the portfolios may also contain stocks from emerging markets, using factors for developed markets might not be ideal. We therefore repeat the calculations by running a six-factor model using global factor data and provide the results in Panel B. We find that the bottom portfolio loads negatively on the momentum factor, leading to a roughly one percentage point lower alpha. We obtain similar results if we weight country-specific factors according to the country exposure of a portfolio in Panel C. Thus, an overweighting of individual countries might not explain the observed outperformance.

5.3. Portfolio optimization

Next to stock picking, investors also strive to optimize the return-to-volatility ratio of their stocks in the portfolio. One way to achieve this is by optimizing portfolio weights, for instance, via a mean–variance approach (Markowitz, 1952). The idea is that portfolio weights are chosen such that the return-to-volatility ratio is maximized. However, the performance of this approach depends heavily on the accuracy of the return forecasts. Investors often use past returns to measure future returns, even though this might lead to portfolios where past winners are strongly overweighted. To circumvent this issue, we suggest *SOP*, a model-based future return estimate. In principle, we use the outperformance probabilities calculated by the model and shift them down by 0.5. Since this measure is naturally bounded between −0.5 and 0.5, we expect to see less extreme weights in the portfolio compared to a mean–variance portfolio that uses past returns. We hypothesize that this should translate into a lower portfolio concentration and volatility.

To test whether this is the case, we compare the mean and volatility of mean–variance portfolios using either *past return* or *SOP* as a measure of future return. Note that we select stocks randomly rather than forming portfolios based on the stocks with the highest probability. The reason is that the expected returns for high-ranked stocks would be very similar by construction. In

Table 5
Five- and six-factor model regressions using different factor data.

Factor	EW top	VW top	EW bottom	VW bottom	EW TB	VW TB
Panel A: Dev. factors						
MKTRF	0.72***	0.77***	0.85***	0.82***	−0.13*	−0.05
	(14.85)	(17.94)	(11.72)	(10.88)	(−1.74)	(−0.68)
SMB	0.3***	−0.11	0.61***	0.38**	−0.29*	−0.47**
	(4.25)	(−1.43)	(3.66)	(2.08)	(−1.72)	(−2.31)
HML	−0.07	−0.15	0.28	0.49***	−0.36*	−0.65***
	(−0.62)	(−1.57)	(1.61)	(2.55)	(−1.88)	(−3.39)
RMW	0.21	0.28**	−0.27	−0.25	0.48**	0.53**
	(1.42)	(2.12)	(−1.13)	(−0.93)	(2.0)	(1.95)
CMA	−0.17	−0.05	−0.91***	−1.08***	0.76***	1.03***
	(−0.97)	(−0.4)	(−4.03)	(−4.62)	(3.14)	(4.26)
Alpha	11.79***	6.79***	−11.25***	−9.84***	21.79***	15.38***
	(6.59)	(4.01)	(−3.52)	(−2.92)	(6.56)	(4.36)
Panel B: Global factors						
MKTRF	0.69***	0.74***	0.87***	0.81***	−0.18**	−0.07
	(14.67)	(19.76)	(12.61)	(11.1)	(−2.43)	(−0.93)
SMB	0.59***	0.18**	0.88***	0.5**	−0.26	−0.29
	(5.45)	(2.0)	(4.6)	(2.35)	(−1.28)	(−1.26)
HML	−0.06	−0.13	0.12	0.25	−0.2	−0.4*
	(−0.47)	(−0.99)	(0.68)	(1.25)	(−0.94)	(−1.73)
WML	−0.02	0.06	−0.17**	−0.26***	0.15	0.31***
	(−0.18)	(0.68)	(−2.19)	(−3.47)	(1.07)	(2.5)
RMW	0.08	0.08	−0.13	−0.1	0.22	0.19
	(0.92)	(0.96)	(−0.95)	(−0.56)	(1.49)	(1.03)
CMA	−0.14	−0.03	−0.36	−0.51*	0.24	0.51
	(−1.04)	(−0.17)	(−1.44)	(−1.7)	(0.88)	(1.5)
Alpha	12.44***	7.0***	−10.7***	−8.73***	21.9***	14.49***
	(5.69)	(3.27)	(−3.67)	(−2.88)	(6.23)	(4.2)
Panel C: Country factors						
MKTRF_usd	0.64***	0.69***	0.91***	0.8***	−0.26***	−0.16**
	(14.77)	(17.32)	(14.55)	(12.09)	(−3.17)	(−2.03)
SMB_usd	0.57***	0.06	1.05***	0.38**	−0.67***	−0.42*
	(5.04)	(0.49)	(6.99)	(2.01)	(−3.07)	(−1.72)
HML_usd	0.14	0.01	−0.19	0.19	0.09	−0.27
	(1.28)	(0.09)	(−1.44)	(0.93)	(0.46)	(−1.2)
WML_usd	0.0	0.06	−0.24***	−0.26***	0.22	0.33**
	(0.03)	(0.75)	(−4.43)	(−4.7)	(1.43)	(2.33)
RMW_usd	0.02	0.03	−0.0	−0.37*	0.21	0.38
	(0.34)	(0.44)	(−0.02)	(−1.66)	(1.13)	(1.61)
CMA_usd	−0.26**	−0.15	0.14	−0.21	−0.17	0.13
	(−2.0)	(−1.24)	(0.63)	(−0.91)	(−0.59)	(0.4)
Alpha	11.27***	6.26***	−12.17***	−8.74***	22.58***	15.48***
	(5.21)	(3.12)	(−4.33)	(−2.55)	(5.92)	(4.08)

This table shows the factor exposures of equally- and value-weighted, model-driven portfolios. In Panel A, we use the five factors for developed markets that are available on the homepage of Kenneth R. French. In Panel B, we construct global factor data by following the approach of Kenneth R. French as closely as possible. In Panel C, we construct country factors and weigh them by the country weights of the portfolio to control for country-specific effects. We employ heteroskedasticity adjusted robust standard errors according to MacKinnon and White (1985). * indicates significance at the 10% level, ** indicates significance at the 5% level and *** indicates significance at the 1% level.

the extreme case, if all considered stocks in the long-leg (short-leg) share the same probabilities of outperformance, we would have zero variance within expected returns. Since including a variable without variation is the same as excluding it from the objective function, we would simply get the weights of a minimum variance portfolio, where expected stock returns do not enter the objective function.

We determine the weights of the minimum variance portfolios as follows. For a given month and portfolio, we obtain the daily stock return data for all portfolio components within the previous 12 months. We then calculate the covariance matrix of the stock components using past return data. Doing so ensures that no future information is considered during the optimization. We calculate the mean–variance portfolio weights by using either *past return* or *SOP* as a measure of future return. We then construct portfolios based on these weights at the beginning and evaluate their performance at the end of the month. We repeat this process every month in the evaluation period from 2004 until 2018 to obtain a series of portfolio returns that we can compare against equally- and value-weighted portfolios.

Table 6
Optimized randomly drawn portfolios.

Short	Portfolio	Mean	Vol	SR	MD	Skew.	Kurt.	ASR	Alpha
No	EW	8.0	14.0	0.49	−16.65	−0.82	2.14	0.46	2.24
No	VW	6.73	13.61	0.41	−15.33	−0.71	1.81	0.39	1.25
No	MeanVP_{RET}	6.96	18.7	0.31	−22.21	−0.92	4.23	0.29	−1.21
No	MeanVP_{SOP}	11.48	10.06	1.02	−14.63	−1.03	4.69	0.77	4.21*
Yes	MeanVP_{RET}	13.92	74.43	0.17	−96.78	−0.42	3.47	0.17	18.52
Yes	MeanVP_{SOP}	22.95	29.1	0.75	−40.86	−0.21	7.14	0.66	15.13

We randomly select 100 stocks with a market capitalization of more than $10 billion and evaluate different portfolio weightings. Next to equally- and value-weighted portfolios, we report the performance of MeanVP_{RET}, a mean–variance portfolio that uses past returns as a measure of future return. MeanVP_{SOP} is a mean–variance portfolio using shifted outperformance probabilities to measure future returns. *Mean* represents the average yearly return, *Vol* the average yearly volatility and *SR* the Sharpe ratio of a portfolio. We further calculate the maximum drawdown *MD*, the skewness *Skew*, the kurtosis *Kurt*, the market capitalization in billion-dollar *MV*, as well as the five-factor alpha (*Alpha*) using factor data for developed markets provided by Kenneth R. French. * indicates significance at the 10% level, ** indicates significance at the 5% level and *** indicates significance at the 1% level.

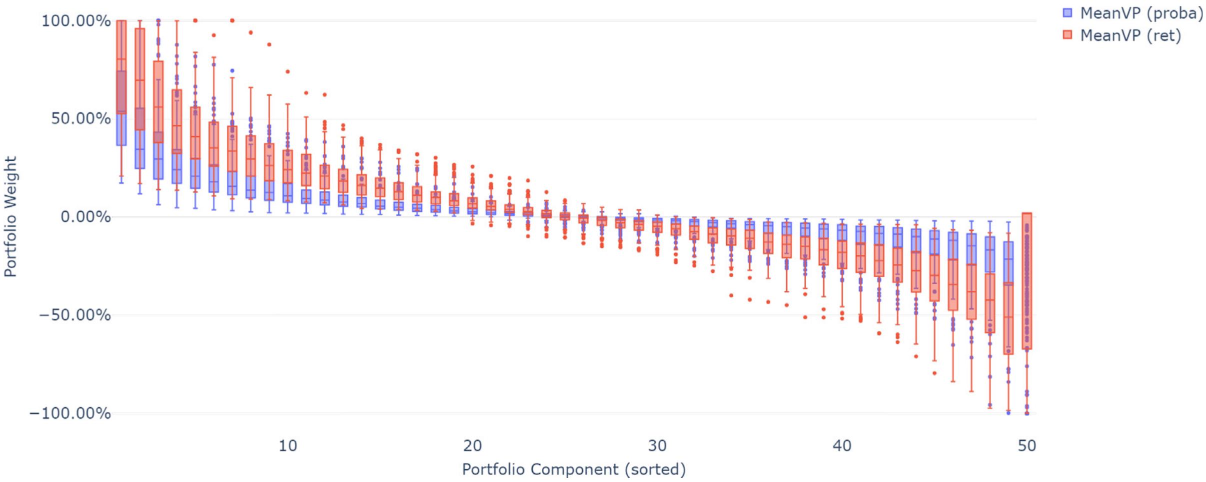

Fig. 2. Distribution of portfolio weights of mean–variance portfolios using different return measures.

As Avramov et al. (2021) point out, ML-driven investment strategies tend to overweight stocks that face strong limits to arbitrage and thus the reportedly high profitability of ML-driven investments is often hard to realize in practice. To ensure that our results do not suffer from this effect, we restrict the investment universe to large stocks (above $10 billion market capitalization).

Table 6 compares different portfolio strategies based on hundred randomly selected stocks with a large market capitalization (monthly rebalancing). If we disallow negative portfolio weights (long positions only), we find evidence that the $MeanVP_{RET}$ seems more concentrated than an equally-weighted portfolio. We observe a roughly four to five percentage point higher volatility and a substantially higher maximum drawdown of 22.21%. At the same time, the average return is lower than the average return of an equally- (value-) weighted investment into the entire universe of stocks with a market capitalization above $10 billion.

In line with our initial hypothesis, we observe a substantially lower volatility and maximum drawdown if we construct the MeanVP_{SOP}. The volatility is as low as 10.06% per year, and the maximum drawdown is less than 15%. At the same time, we observe a substantially higher annual return (11.48%). This translates into a Sharpe ratio of 1.02 which is substantially higher than the MeanVP_{RET} and the equally- and value-weighted portfolios. We even observe a slightly significant yearly five-factor alpha of 4.21%. Even if we control for tail risk, these results persist. The ASR is 0.77 compared to 0.46 (0.39) for the equally- and value-weighted investments.

A similar image may be obtained if we allow short selling. We observe a substantially higher Sharpe ratio for MeanVP_{SOP} (0.75) compared to the MeanVP_{RET} (0.17). This is driven by both a larger mean and a smaller volatility. While the mean increases from roughly 14% to 23%, the yearly volatility is reduced from 74.43% to less than 30%. Also, the maximum drawdown is reduced from almost 97% to around 41%. Again, the results persist after controlling for tail risk.

To further elaborate on the differences in portfolio weights for optimized portfolios with no shorting restriction, we provide a visual representation of the portfolio weights of fifty large stocks over time in Fig. 2. As we rank the component weights monthly and present both the mean and the standard deviation in a boxplot for each ranking position over the whole investment horizon

Table 7
Regional differences.

Portfolio	Mean	Vol	SR	MD	Skew	Kurt	Alpha
Panel A: Regional (VW)							
Emerging Markets	19.99	15.27	1.23	−24.12	−1.69	7.73	12.66***
Europe	13.93	12.60	1.01	−11.87	−0.61	1.59	11.63***
North America	15.05	13.57	1.02	−19.11	−0.96	4.65	6.91***
Pacific	14.31	14.47	0.91	−19.74	−1.22	4.99	7.39*
Panel B: Country (VW)							
Australia	14.14	14.52	0.89	−21.90	−1.18	5.49	6.38*
Canada	16.82	11.94	1.31	−14.29	−0.82	4.20	12.79***
China	21.51	28.77	0.71	−25.08	−0.05	1.18	8.42
France	17.60	14.45	1.14	−10.32	0.13	1.04	14.45***
Germany	12.16	15.23	0.72	−19.32	−0.73	4.00	4.95
Hong Kong	15.25	21.37	0.66	−33.51	−1.03	6.52	6.7
India	24.16	24.73	0.93	−17.03	1.29	7.80	12.02***
Italy	13.14	15.34	0.78	−11.91	0.33	1.99	8.88**
Japan	13.44	18.82	0.65	−16.35	−0.00	0.96	9.5**
South Africa	32.67	18.01	1.75	−28.00	−1.29	7.44	24.59***
South Korea	19.43	23.29	0.78	−22.75	0.39	2.64	9.99
Spain	15.26	17.26	0.82	−11.52	−0.11	0.09	10.32***
Sweden	20.17	18.31	1.04	−23.92	−0.98	4.20	13.07***
Switzerland	10.76	13.92	0.69	−17.50	−0.62	2.86	9.97***
Taiwan	19.41	16.29	1.12	−18.51	−0.67	3.63	13.41***
United Kingdom	16.11	12.86	1.16	−12.16	−0.53	1.16	10.72***
USA	14.75	14.01	0.97	−18.87	−0.83	3.90	6.4***

In Panel A, we present the performance metrics of value-weighted long portfolios (decile split) that are restricted to individual regions. Panel B shows the performance of country-specific value-weighted long portfolios. We evaluate the performance over a fifteen-year time horizon starting in 2004. *Mean* represents the average yearly return, *Vol* the average yearly volatility and *SR* the Sharpe ratio of a portfolio. We further calculate the maximum drawdown *MD*, the skewness *Skew*, the kurtosis *Kurt*, and the market capitalization in billion-dollar *MV*. We also calculate a six-factor alpha using regional (country-specific) factor data. We employ heteroskedasticity adjusted robust standard errors according to MacKinnon and White (1985). * indicates significance at the 10% level, ** indicates significance at the 5% level and *** indicates significance at the 1% level.

(180 months), we need a static portfolio size as otherwise, the weights are not comparable. We thus randomly select fifty large stocks before we apply the mean–variance optimization. According to Fig. 2, the portfolio weights of MeanVP$_{SOP}$ are less extreme in comparison to the MeanVP$_{RET}$. For example, the largest long component of the MeanVP$_{RET}$ receives a median weight of 80.42%, whereas it is only 53.74% in the case of MeanVP$_{SOP}$. A similar message is obtained for the largest short positions. Here, we observe that the average portfolio weight is less negative for the MeanVP$_{SOP}$ than MeanVP$_{RET}$.

Based on these findings, we argue that investors can maximize a portfolio's return-to-volatility ratio while avoiding a too-high concentration by constructing a mean–variance portfolio using *SOP* to measure the future return.

6. Robustness

6.1. Regional restriction

Intuitively, one may hypothesize that the performance of the global, model-based portfolios could be driven by stocks that are traded in less efficient stock markets. One of the promoting factors could be potential trade restrictions for foreign investors. For example, foreign investors could not trade Chinese or Indian stocks for quite some time during our evaluation period. We construct portfolios on the region and country-level to control for these factors. Moreover, we only focus on the long leg since we lack country-specific information on short-selling possibilities (see Table 7).

According to Panel A, we observe significant alphas for all the regionally-restricted, value-weighted portfolios.[18] While we observe a highly significant alpha of more than 12% in emerging markets, investments into North American stocks yield a high five-factor alpha of more than 7% annually. If we consider country-specific investments, we obtain a similar image in Panel B.[19] Except for China, Germany, and South Korea, we observe significant five-factor alphas, indicating that politically induced investment restrictions are not likely to explain the market's superior returns.

[18] We report value-weighted rather than equally-weighted portfolios to mitigate the effects of limits to arbitrage.
[19] Note that we only display countries with a sufficiently large amount of available stocks during our investment period.

Table 8
Portfolio performance within different time periods.

Group	EW top	VW top	EW bottom	VW bottom	EW TB	VW TB
VStoxx						
Low	7.8***	3.86*	−3.37	−0.45	9.98*	4.4
	(3.56)	(1.71)	(−0.92)	(−0.09)	(1.77)	(0.7)
High	13.0***	8.14***	−19.81***	−13.78**	31.32***	20.61***
	(4.05)	(2.44)	(−4.31)	(−2.37)	(6.07)	(3.69)
Spread						
Low	10.02***	3.87*	0.95	5.72	10.89**	3.22
	(3.85)	(1.91)	(0.21)	(0.96)	(2.0)	(0.54)
High	11.92***	7.95***	−21.41***	−18.37***	32.06***	25.47***
	(4.1)	(2.47)	(−6.3)	(−4.68)	(6.92)	(4.71)
Sentiment						
Low	11.68***	6.13**	−13.34***	−8.43	26.0***	18.68***
	(4.11)	(2.37)	(−3.08)	(−1.51)	(5.0)	(3.34)
High	11.68***	6.85**	−9.5**	−5.9	17.06***	11.74*
	(3.32)	(2.02)	(−2.23)	(−0.98)	(2.58)	(1.68)
Time						
2004–2011	9.26***	5.24	−18.65***	−17.09***	25.73***	19.58***
	(3.04)	(1.5)	(−6.37)	(−4.69)	(6.0)	(3.85)
2011–2018	11.53***	5.12**	0.44	7.03	17.41***	8.35
	(5.02)	(2.06)	(0.08)	(1.0)	(3.57)	(1.36)

We conduct median splits based on variables like *volatility*, *sentiment*, *bid–ask spread*, and *time* to assess the performance of different portfolios in various market environments. We consider investments into the 10% of stocks with the largest (lowest) outperformance probabilities and provide yearly five-factor alphas using developed factor data from Kenneth R. French. We employ heteroskedasticity adjusted robust standard errors according to MacKinnon and White (1985). * indicates significance at the 10% level, ** indicates significance at the 5% level and *** indicates significance at the 1% level.

6.2. Profitability over time

The superior returns of the model-based portfolios could also be related to specific periods. Among others, Avramov et al. (2021) find evidence that ML-driven investment strategies perform substantially better within periods of economic distress. We conduct multiple median splits to test whether model-based investments generate positive alpha in non-crisis periods.

Table 8 shows the performance of the previously studied portfolios for different periods. We conduct a median split based on the VStoxx volatility index and find that the alphas obtained from long–short investments are substantially higher in periods of high volatility. While the equally-weighted long–short portfolio generates a yearly alpha of 31.32% in more volatile periods, the obtained yearly five-factor alpha is only marginally significant in low volatile periods. We obtain a highly similar image in the case of the average bid–ask spread. What is striking is that within low volatile and low bid–ask spread periods, short investments into the lowest-ranked stocks do not generate significant alpha. In contrast, we observe high economically and statistically significant alphas in periods of economic distress.

If we conduct a median split based on investor sentiment as measured by Baker and Wurgler (2006), we obtain highly significant alphas in low and high sentiment periods. Finally, we also observe lower alphas in more recent periods. However, we cannot infer that markets got more efficient over time. The reason is that the performance decrease in the long–short portfolio is mainly driven by the short leg. Since we have seen that shorting low-ranked stocks is only profitable in crisis periods, the performance reduction could also be explained by the lack of a severe crisis in the more recent half of the evaluation period.

6.3. Transaction costs

We have yet to consider transaction costs when evaluating the different portfolio strategies. One may hypothesize that the suggested investment strategy requires frequent trading and thus induces high transaction costs, ultimately cutting profitability. To control for transaction costs, we repeat our main calculations for stocks with a market capitalization of above $10 billion using net portfolio returns. We calculate net portfolio returns as follows. First, we define the average bid–ask spread in the previous month and use it as a proxy for the transaction costs induced by trading a stock in a particular month. In principle, we model transaction costs as the money an investor loses if he buys a security and instantaneously sells it again.

For those stocks where we miss the bid–ask spread, we assume a median bid–ask spread of the dataset in that particular month. Second, we subtract the respective bid–ask spreads from those stocks that newly entered the portfolio. By doing so, we ensure that transaction costs only occur once. Note that our measure of transaction costs is not suitable for short investments. We would need international short-interest data to model the transaction costs of the short-leg effectively. Since we lack access to this data, we focus on long investments only.

Table 9 shows the effect of transaction costs on the performance of the different portfolios. We observe a decrease in the average return of the portfolios of roughly two percentage points. If we control for transaction costs, the observed Sharpe ratios remain

Table 9
Portfolio strategies vs. Benchmark index, 100 stocks max.

Portfolio	Mean	Vol	SR	MD	Skew	Kurt	ASR	Alpha
Panel A: Gross returns								
Tot	8.23	14.25	0.49	−19.38	−1.13	3.66	0.44	2.44*
T-100	13.74	12.64	0.99	−16.14	−0.74	3.48	0.85	6.89***
T-dec	14.67	12.15	1.11	−13.50	−0.44	2.40	1.05	7.67***
T-quint	14.33	12.50	1.05	−14.65	−0.55	2.99	0.95	7.58***
Panel B: Net returns								
Tot	5.88	14.31	0.33	−19.38	−1.12	3.60	0.31	0.51
T-100	11.42	12.69	0.81	−16.14	−0.72	3.44	0.72	4.95***
T-dec	12.34	12.18	0.92	−13.50	−0.43	2.39	0.88	5.65***
T-quint	12.12	12.55	0.87	−14.65	−0.54	2.95	0.80	5.66***

In this table, we evaluate the impact of transaction costs on the performance of portfolios of stocks with a market capitalization above $10 billion using an equally-weighted approach. We proxy transaction costs by the firm-specific average bid–ask spread in the previous month. *Mean* represents the average yearly return, *Vol* the average yearly volatility and *SR* the Sharpe ratio of a portfolio. We further calculate the maximum drawdown *MD*, the skewness *Skew*, the kurtosis *Kurt*, and the market capitalization in billion-dollar *MV*. We also calculate a five-factor alpha using developed factor data from Kenneth R. French. We employ heteroskedasticity adjusted robust standard errors according to MacKinnon and White (1985). * indicates significance at the 10% level, ** indicates significance at the 5% level and *** indicates significance at the 1% level.

Table 10
RP-PCA factor loadings.

Coefficient	(1)	(2)	(3)	(4)	(5)
Panel A: Top portfolio					
Constant	1.13***	0.27**	0.51***	0.21***	0.21***
	(8.77)	(2.28)	(6.1)	(3.37)	(4.92)
Factor 1	12.17***	16.01***	16.76***	18.27***	18.97***
	(11.57)	(20.78)	(31.34)	(48.1)	(85.02)
Factor 2		26.51***	26.0***	30.56***	31.69***
		(13.72)	(20.16)	(29.83)	(39.18)
Factor 3			−19.73***	−19.64***	−21.14***
			(−7.88)	(−11.37)	(−16.74)
Factor 4				37.57***	40.31***
				(10.36)	(14.78)
Factor 5					35.59***
					(10.7)
R^2	0.75	0.91	0.94	0.97	0.99
R^2_{adj}	0.75	0.9	0.94	0.97	0.98
Panel B: Bottom portfolio					
Constant	−1.88***	−1.01***	−0.3***	0.07	0.07*
	(−10.83)	(−4.28)	(−3.21)	(1.24)	(1.67)
Factor 1	22.74***	18.85***	21.02***	19.15***	19.84***
	(27.36)	(16.49)	(52.72)	(49.52)	(95.43)
Factor 2		−26.9***	−28.4***	−34.03***	−32.93***
		(−6.09)	(−18.26)	(−31.33)	(−43.47)
Factor 3			−57.23***	−57.34***	−58.8***
			(−23.03)	(−25.69)	(−44.99)
Factor 4				−46.39***	−43.73***
				(−15.26)	(−19.78)
Factor 5					34.62***
					(11.04)
R^2	0.83	0.88	0.98	0.99	1.0
R^2_{adj}	0.83	0.88	0.98	0.99	1.0

This table illustrates the exposure of two portfolios to the five most prominent factors identified through RP-PCA analysis: one that is long in the top 4% of stocks (Panel A) and another that is short in the bottom 4% of stocks (Panel B). We further document the R-Squared and Adjusted R-Squared. We employ heteroskedasticity adjusted robust standard errors according to MacKinnon and White (1985). * indicates significance at the 10% level, ** indicates significance at the 5% level and *** indicates significance at the 1% level.

Table 11
RP-PCA extended factor model regressions.

Factor	EW top	VW top	EW bottom	VW bottom	EW TB	VW TB
Panel A: T-100						
MKT-RF	−0.08	0.19**	0.06	−0.01	−0.15	0.19
	(−1.6)	(2.34)	(0.68)	(−0.03)	(−1.32)	(0.93)
SMB	−0.06	−0.16**	−0.14	−0.4**	0.09	0.25
	(−1.1)	(−2.03)	(−1.54)	(−1.93)	(0.96)	(1.12)
HML	−0.02	−0.17*	0.14*	0.31*	−0.18*	−0.5**
	(−0.41)	(−1.84)	(1.78)	(1.65)	(−1.79)	(−2.08)
RMW	−0.23***	0.19	−0.22	0.15	−0.0	0.05
	(−2.59)	(1.2)	(−1.5)	(0.53)	(−0.03)	(0.14)
CMA	−0.02	0.26**	−0.29**	−0.13	0.29*	0.4
	(−0.15)	(2.04)	(−2.19)	(−0.39)	(1.77)	(1.06)
Factor 1	0.2***	0.08***	0.06***	0.05***	0.0	−0.01
	(14.52)	(7.69)	(8.93)	(4.63)	(0.49)	(−0.79)
Factor 2	0.33***	0.15***	−0.12***	−0.1***	0.14***	0.12***
	(13.02)	(8.1)	(−8.87)	(−5.07)	(13.39)	(7.21)
Factor 3	−0.26***	−0.08***	−0.24***	−0.2***	0.09***	0.1***
	(−8.24)	(−3.25)	(−12.01)	(−7.98)	(5.31)	(4.81)
Factor 4	0.46***	0.2***	−0.19***	−0.21***	0.21***	0.2***
	(8.09)	(4.37)	(−6.45)	(−3.53)	(9.97)	(4.77)
Factor 5	0.38***	0.14**	0.2***	0.16**	−0.04	−0.06
	(5.58)	(2.16)	(4.49)	(2.38)	(−1.13)	(−0.98)
Alpha	3.87***	−3.12	−3.86*	1.61	6.57**	−5.88
	(2.53)	(−1.36)	(−1.72)	(0.31)	(2.32)	(−0.98)
R^2	0.94	0.85	0.95	0.82	0.87	0.67
R^2_{adj}	0.93	0.84	0.95	0.81	0.86	0.65

This table shows the factor exposure for equally (EW) and value-weighted portfolios (VW) that are long (short) in the 100 highest (lowest) ranked stocks. Portfolios are re-balanced monthly. Next to the five factors suggested by Fama and French (2015), we include five additional factors identified via RP-PCA. We employ heteroskedasticity adjusted robust standard errors according to MacKinnon and White (1985). * indicates significance at the 10% level, ** indicates significance at the 5% level and *** indicates significance at the 1% level.

larger than the Sharpe ratio an investor would obtain by investing in the entire dataset. Even if we control for tail exposure, we observe an ASR of up to 0.88, which is twice as high as an equal investment into all stocks, assuming zero transaction costs. The alpha we obtain remains highly significant and ranges from 5% to 6%. We thus argue that transaction costs alone may not explain the outperformance of the portfolios.

6.4. RP-PCA

We have shown earlier that none of the factor models may fully explain the outperformance of our model-guided portfolios. However, there could be unobserved factors that these portfolios load on. If this is the case, the abnormal return might be interpreted as a risk premium for these unobserved factors rather than as a sign of market inefficiency. To test this hypothesis, we run a RP-PCA analysis as suggested by Lettau and Pelger (2020) to identify the most essential factors within the cross-section of stocks. More specifically, we subdivide the dataset into 25 portfolios with monthly rebalancing using the outperformance probabilities calculated by our model as a sorting variable. Using excess returns, we then identify the five factors that best describe the 25 portfolios and test how much of the variance may be explained by these factors.

Panel A in Table 10 shows the exposure of the portfolio constructed based on the 4% of stocks with the highest outperformance probabilities. The most important factor already explains 81% We find similar results for the portfolio invested in the stocks with the lowest outperformance probabilities in Panel B.

As a next step, we test whether the outperformance reported so far might be explained by the five most important RP-PCA factors. We therefore extend the baseline five-factor model by the five most relevant RP-PCA factors.

According to Table 11, we find that all ten factors may explain up to 94% of the returns of a portfolio that is equally invested in the hundred stocks with the highest outperformance probabilities. Nevertheless, we observe a yearly ten-factor alpha of 3.87% which is significant at the 1% level. At the same time, a portfolio that is invested in the stocks with the lowest outperformance probability yields a negative alpha of −3.86%. An investor that buys the first and shorts the latter thus ends up with a yearly alpha of 6.57% which is significant at the 5% level. If we value-weight the portfolios, we observe no significant alphas, suggesting that smaller stocks might drive unexplained outperformance.

7. Conclusion

This paper proposes a new approach to identify future outperformers in international stock markets. While most researchers focus on return prediction, we propose a model that predicts the outperformance probability of a stock during the subsequent month. We train a random forest model based on technical indicators and test it on a fifteen-year horizon with monthly investments into sufficiently liquid stocks. We find that out- and underperformers indeed share common technical attributes that machine learning models may discover. The highly non-linear combination of many features seems to drive our model's performance. We observe Sharpe Ratios of up to 2.49 for equally-weighted and 1.56 for value-weighted portfolios, which is substantially higher than their equally- and value-weighted benchmarks. Limits to arbitrage may not explain our findings, given that we observe a similar pattern for stocks with a market capitalization above $10 billion.

The observed outperformance may not be explained by any of the most prominent pricing factors in the literature. We find that model-guided equally- (value-) weighted long–short portfolios yield highly significant alphas. For instance, an equally-weighted long (short) portfolio in the stocks with the highest (lowest) ranked ten percent of stocks yields a highly significant five-factor alpha of 21.79%. We obtain qualitatively similar results using value weights.

We further investigate whether outperformance probabilities may be used to model future stock returns in a mean–variance optimization context. We find that using our measure of future stock returns *SOP* rather than *Mean* leads to less concentrated portfolios with a superior return-to-volatility ratio. In the absence of short selling, we observe a Sharpe ratio that is more than twice as high as the Sharpe ratios of equally- and value-weighted portfolios.

We employ additional robustness tests to ensure that some trading friction might not drive our findings. Overall, neither regional restrictions, transaction costs, nor exposure to unobserved factors may fully explain the observed alpha. While the long–short investment approach works particularly well in crisis periods, its long leg remains profitable in non-crisis periods. Given that the technical indicators applied in this paper have been largely overlooked within academia, we hypothesize that our model might have discovered return patterns that investors might have been unaware of in the past.

By nature, our findings are highly relevant for fund managers pursuing an active stock selection strategy. Not only can our stock ranking approach serve as a guideline for future investments, but it may also be enhanced with other stock rankings derived from fundamental or textual data. Future research may investigate whether adding fundamental or textual data to the model may lead to a more valuable stock ranking.

CRediT authorship contribution statement

Christian Breitung: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing, Visualization.

Appendix

A.1. Calculating outperformance probabilities

To facilitate the replication of the results presented in this paper, we provide a detailed description of the steps to calculate the outperformance probabilities mentioned in our paper. First, stock return data has to be downloaded. Within this paper, we use Refinitiv as a data provider. Alternative providers should also work if they offer *open*, *high*, *low*, *close* and *volume* data on a daily basis. Second, we calculate the technical indicators using the python package *ta*. We construct additional indicators by dividing the raw indicators by the corresponding stock prices to get rid of the stock price level that is still contained in some of the indicators. We further add custom-defined indicators. Note that the indicators we construct are calculated on a daily level. Since we aim at rebalancing portfolios at the beginning of each month, we use the latest available indicators from the previous month to avoid including any information that was not yet available at the beginning of a month.

Second, we apply our liquidity thresholds to remove all stocks from the dataset that are not sufficiently liquid (less than $300 million market capitalization and less than 180 trading days in the previous year). For the remaining stocks, we assign labels based on their next month performance. More specifically, we determine monthly median returns and assign a value of one (zero), if the monthly stock return is above (below) the median return in that month. Using the indicators and labels mentioned above, we train a random forest classification on the concatenation of all monthly return information from 1990 until 2003.

To evaluate the out-of-sample prediction power of our model, we construct portfolios based on predicted outperformance probability from 2004 until 2018 in a third step. We count the number of trees that predict an outperformance and divide it by the total number of decision trees in the model to obtain an outperformance probability ranking. Finally, we use this ranking to construct portfolios at the beginning of a month and hold these positions for one month.

Table A.1
List of features.

	Acronym	Description/Source
1	avg. 50	Return of the last 50 trading days
2	avg. 100	Return of the last 100 trading days
3	avg. 200	Return of the last 200 trading days
4	sharpe_ratio	Yearly return divided by yearly volatility, assuming risk free rate of 0.
5	adj_sharpe_ratio	Yearly return divided by yearly volatility corrected for skewness.
6	skewness	Skewness of daily returns
7	kurtosis	Kurtosis of daily returns
8	max_spread	Maximum difference between largest and smallest daily return
9	avg_return	Average daily return
10	short_vola	Daily volatility during last 50 trading days.
11	long_vola	Daily volatility during last year.
12	difference_vola	Difference between short- and long-term daily volatility.
13	trade_days	Number of trade days within the last year
14	volume_adi	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volume-indicators
15	volume_obv	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volume-indicators
16	volume_cmf	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volume-indicators
17	volume_fi	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volume-indicators
18	volume_mfi	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volume-indicators
19	volume_em	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volume-indicators
20	volume_sma_em	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volume-indicators
21	volume_vpt	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volume-indicators
22	volume_nvi	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volume-indicators
23	volume_vwap	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volume-indicators
24	volatility_atr	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
25	volatility_bbm	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
26	volatility_bbh	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
27	volatility_bbl	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
28	volatility_bbw	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
29	volatility_bbp	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
30	volatility_bbhi	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
31	volatility_bbli	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
32	volatility_kcc	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
33	volatility_kch	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
34	volatility_kcl	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
35	volatility_kcw	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
36	volatility_kcp	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
37	volatility_kchi	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
38	volatility_kcli	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
39	volatility_dcl	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
40	volatility_dch	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
41	volatility_dcm	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
42	volatility_dcw	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
43	volatility_dcp	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
44	volatility_ui	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#volatility-indicators
45	trend_macd	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
46	trend_macd_signal	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
47	trend_macd_diff	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
48	trend_sma_fast	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
49	trend_sma_slow	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
50	trend_ema_fast	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators

Table A.2
List of features (continued).

	Acronym	Description/Source
50	trend_ema_fast	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
51	trend_ema_slow	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
52	trend_adx	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
53	trend_adx_pos	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
54	trend_adx_neg	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
55	trend_vortex_ind_pos	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
56	trend_vortex_ind_neg	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
57	trend_vortex_ind_diff	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
58	trend_trix	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
59	trend_mass_index	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
60	trend_cci	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
61	trend_dpo	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
62	trend_kst	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
63	trend_kst_sig	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
64	trend_kst_diff	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
65	trend_ichimoku_conv	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
66	trend_ichimoku_base	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
67	trend_ichimoku_a	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
68	trend_ichimoku_b	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
69	trend_visual_ichimoku_a	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
70	trend_visual_ichimoku_b	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
71	trend_aroon_up	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
72	trend_aroon_down	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
73	trend_aroon_ind	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
74	trend_psar_up	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
75	trend_psar_down	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
76	trend_psar_up_indicator	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
77	trend_psar_down_indicator	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
78	trend_stc	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#trend-indicators
79	momentum_rsi	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#momentum-indicators
80	momentum_stoch_rsi	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#momentum-indicators
81	momentum_stoch_rsi_k	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#momentum-indicators
82	momentum_stoch_rsi_d	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#momentum-indicators
83	momentum_tsi	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#momentum-indicators
84	momentum_uo	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#momentum-indicators
85	momentum_stoch	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#momentum-indicators
86	momentum_stoch_signal	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#momentum-indicators
87	momentum_wr	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#momentum-indicators
88	momentum_ao	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#momentum-indicators
89	momentum_kama	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#momentum-indicators
90	momentum_roc	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#momentum-indicators
91	momentum_ppo	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#momentum-indicators
92	momentum_ppo_signal	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#momentum-indicators
93	momentum_ppo_hist	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#momentum-indicators
94	others_dr	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#others-indicators
95	others_dlr	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#others-indicators
96	others_cr	https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html#others-indicators

Table A.3
Performance of yearly trained models.

Portfolio	Mean	Vol	SR	MD	Skew	Kurt	MV
1990	12.88	10.99	1.17	−10.39	−0.49	2.85	3.28
1991	11.43	7.17	1.59	−4.60	0.05	0.99	3.30
1992	2.50	5.97	0.41	−4.99	−0.40	0.65	3.27
1993	12.62	8.09	1.55	−8.68	0.32	7.11	3.44
1994	12.63	6.78	1.85	−4.78	−0.15	0.41	4.01
1995	12.57	8.87	1.41	−11.16	−0.63	6.35	3.58
1996	10.47	9.46	1.10	−5.73	−0.42	0.31	3.82
1997	13.30	10.43	1.27	−8.18	−0.04	2.17	3.78
1998	10.33	8.94	1.15	−6.41	0.20	1.18	3.20
1999	12.67	13.44	0.94	−10.45	0.27	2.38	3.16
2000	9.87	14.35	0.68	−14.38	−0.19	2.62	3.13
2001	8.46	14.42	0.58	−11.30	0.08	1.85	2.89
2002	11.06	10.11	1.09	−6.57	−0.02	1.40	3.27
2003	10.72	12.71	0.84	−7.28	0.28	0.53	3.34
2004	11.85	13.82	0.85	−8.25	0.68	2.29	3.41
2005	11.37	8.45	1.34	−6.00	0.85	4.31	3.72
2006	14.64	13.11	1.11	−9.49	0.23	1.18	3.02
2007	6.57	21.08	0.31	−16.36	0.38	2.86	2.87
2008	8.11	15.74	0.51	−17.10	−0.37	3.88	3.02
2009	30.80	20.70	1.48	−14.40	0.64	2.06	3.23
2010	21.57	24.92	0.86	−13.91	0.90	1.14	3.74

This table reports the performance of equally-weighted long–short portfolios based on different yearly trained classification models. We split the dataset into deciles based on the probability of outperformance. We evaluate the performance of these models from mid 2011 until end of 2018. *Mean* represents the average yearly return, *Vol* the average yearly volatility and *SR* the Sharpe ratio of a portfolio. We further calculate the maximum drawdown *MD*, the skewness *Skew*, the kurtosis *Kurt* and the market capitalization in billion dollar *MV*.

Table A.4
Feature importances.

Feature	Total	North America	Europe	Pacific	Emerging markets	Relative share
Momentum	18.21	18.51	18.10	18.34	18.16	17.13
Trend	36.55	0.36.40′	37.47	36.51	37.15	38.67
Volatility	21.58	22.28	22.02	22.13	21.12	25.41
Volume	13.45	13.51	12.01	13.66	13.77	11.05
Other	10.20	9.30	10.41	9.36	9.79	7.73

By counting how often a feature is used to split a node and relating it to the total number of nodes, we calculate the feature importances for our baseline model, as well as regionally trained models. Due to the large amount of features we employ, we follow the classifications of the Python package we apply and split all features into the following groups: momentum, trend, volatility, volume and other. We present all numbers in percentage points.

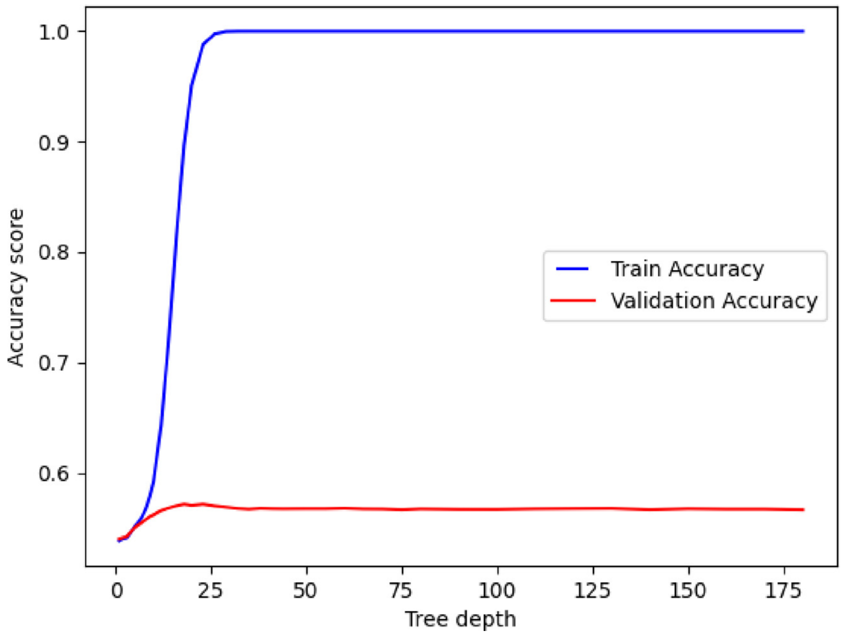

Fig. A.1. Accuracy on the training and validation sets in relation to tree depth.

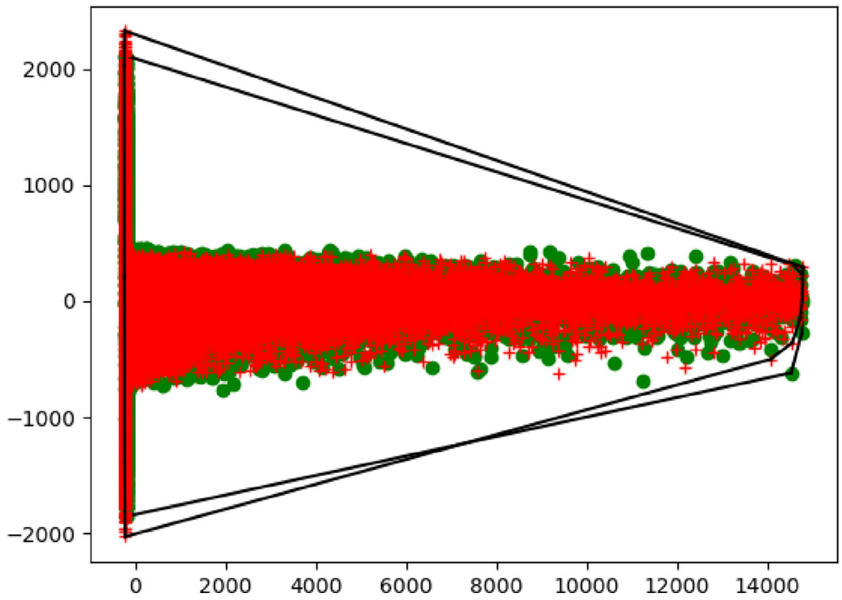

Fig. A.2. Convex hulls of good and bad stocks. Principal component analysis of outperformers (green) and underperformers (red). We reduce the dimensionality of the feature space to visualize whether a linear separation of both classes is possible. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

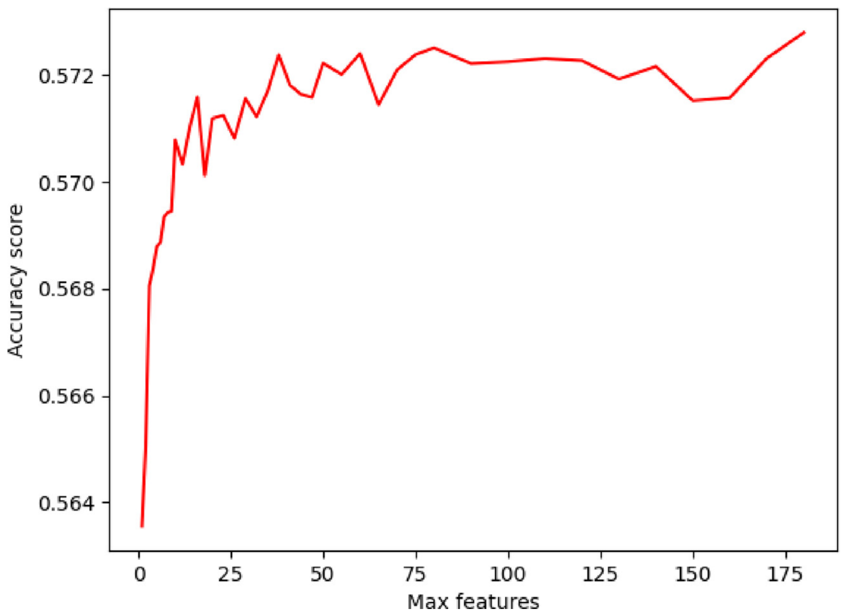

Fig. A.3. Accuracy for different values of maximum features considered at each node.

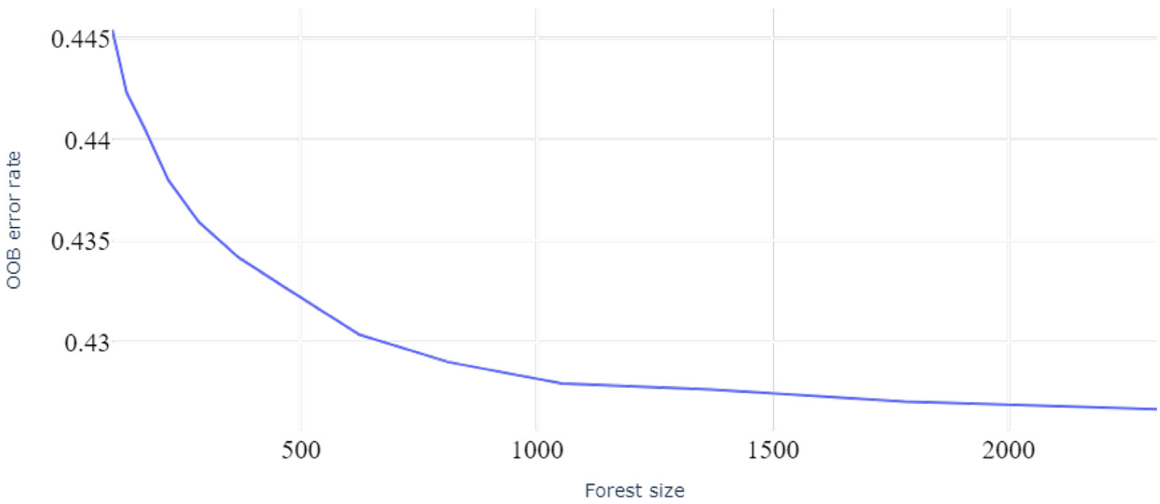

Fig. A.4. Out-of-Bag error rate for different numbers of estimators.

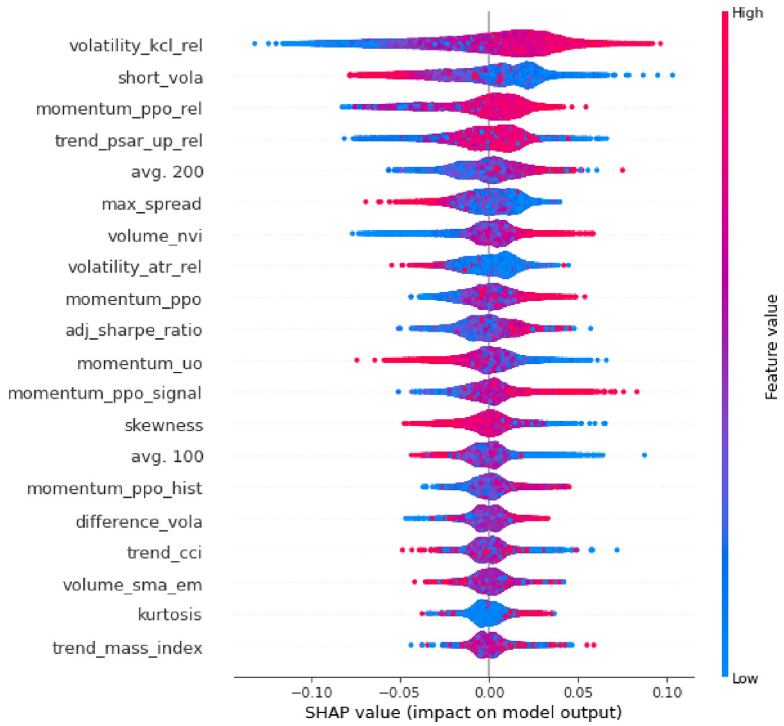

Fig. A.5. Shapley value of outperformers as identified by an US market model.

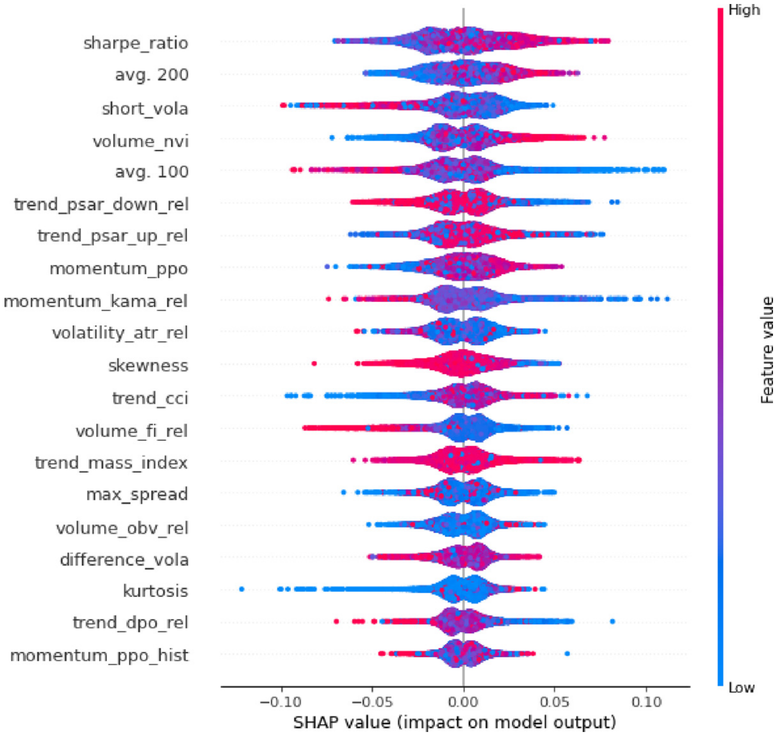

Fig. A.6. Shapley value of outperformers as identified by an emerging market model.

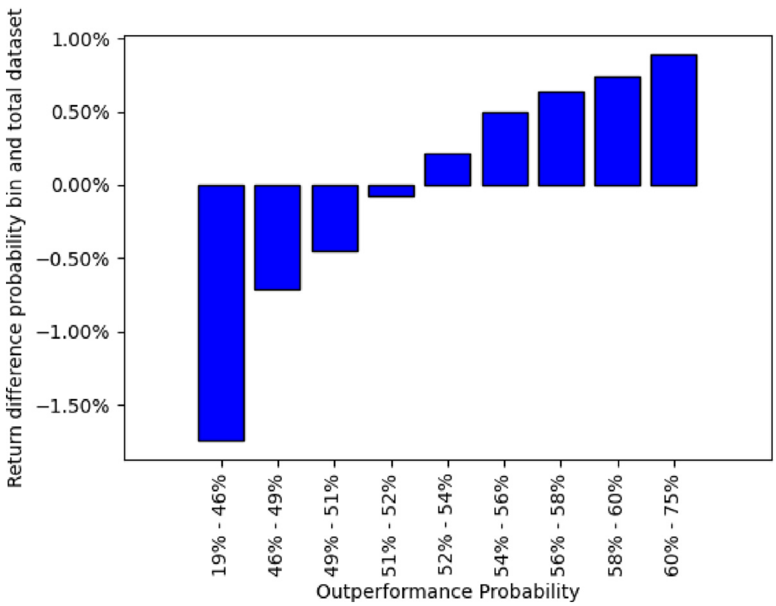

Fig. A.7. Spread between median return and median return of stocks within a certain outperformance probability range.

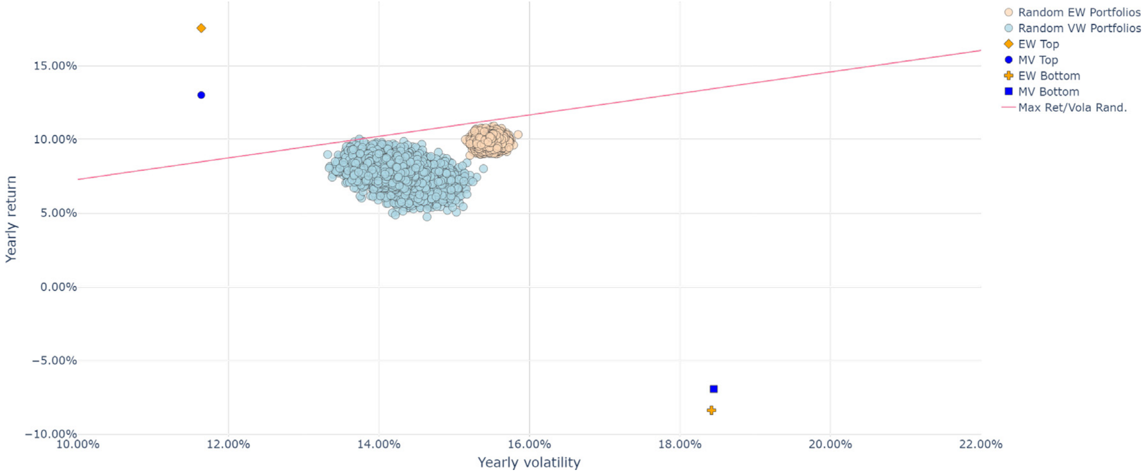

Fig. A.8. Model-based portfolios vs. random portfolios.

References

Abarbanell, J.S., Bushee, B.J., 1998. Abnormal returns to a fundamental analysis strategy. Account. Rev. 19–45.

Avramov, D., Cheng, S., Metzker, L., 2021. Machine learning versus economic restrictions: Evidence from stock return predictability. Available At SSRN 3450322 URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3450322.

Azevedo, V., Kaiser, S., Müller, S., 2022. Stock market anomalies and machine learning across the globe. Available At SSRN 4071852 URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4071852.

Baker, M., Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. J. Finance 61 (4), 1645–1680.

Ban, G.-Y., El Karoui, N., Lim, A.E., 2018. Machine learning and portfolio optimization. Manage. Sci. 64 (3), 1136–1154.

Basak, S., Kar, S., Saha, S., Khaidem, L., Dey, S.R., 2019. Predicting the direction of stock market prices using tree-based classifiers. N. Am. J. Econ. Financ. 47, 552–567.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Breitung, C., Müller, S., 2022. When firms open up: Identifying value relevant textual disclosure using simBERT. Available At SSRN URL: https://ssrn.com/abstract=4215290.

Cakici, N., Fieberg, C., Metko, D., Zaremba, A., 2022. Machine learning goes global: Cross-sectional return predictability in international stock markets. Available At SSRN URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4141663.

Chen, L., Pelger, M., Zhu, J., 2020. Deep learning in asset pricing. Available At SSRN 3350138 URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3159577.

Cohen, L., Malloy, C., Nguyen, Q., 2020. Lazy prices. J. Finance 75 (3), 1371–1415.

De la Fuente, A., 2000. Mathematical Methods and Models for Economists. Cambridge University Press.

DeMiguel, V., Garlappi, L., Nogales, F.J., Uppal, R., 2009. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. Manage. Sci. 55 (5), 798–812.

Dos Santos Pinheiro, L., Dras, M., 2017. Stock market prediction with deep learning: A character-based neural language model for event-based trading. In: Proceedings of the Australasian Language Technology Association Workshop 2017. pp. 6–15.

Emerson, S., Kennedy, R., O'Shea, L., O'Brien, J., 2019. Trends and applications of machine learning in quantitative finance. In: 8th International Conference on Economics and Finance Research (ICEFR 2019).

Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. J. Financ. Econ. 116 (1), 1–22.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. Rev. Financ. Stud. 33 (5), 2223–2273.

Hanauer, M.X., Kalsbach, T., 2023. Machine learning and the cross-section of emerging market stock returns. Emerging Markets Review 55, 101022.

Horning, N., et al., 2010. Random forests: An algorithm for image classification and generation of continuous fields data sets. In: Proceedings of the International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences, Osaka, Japan, Vol. 911.

Jagannathan, R., Ma, T., 2003. Risk reduction in large portfolios: Why imposing the wrong constraints helps. J. Finance 58 (4), 1651–1683.

Lettau, M., Pelger, M., 2020. Factors that fit the time series and cross-section of stock returns. Rev. Financ. Stud. 33 (5), 2274–2325.

Liu, C., Wang, J., Xiao, D., Liang, Q., 2016. Forecasting s&p 500 stock index using statistical learning models. Open J. Stat. 6 (06), 1067.

Lo, A.W., Mamaysky, H., Wang, J., 2000. Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. J. Finance 55 (4), 1705–1765.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.

MacKinnon, J.G., White, H., 1985. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. J. Econometrics 29 (3), 305–325.

Markowitz, H., 1952. Portfolio selection. J. Finance 7 (1), 77–91, URL: http://www.jstor.org/stable/2975974.

Molnar, C., 2020. Interpretable Machine Learning. Lulu. com.

Obthong, M., Tantisantiwong, N., Jeamwatthanachai, W., Wills, G., 2020. A survey on machine learning for stock price prediction: algorithms and techniques.

Pezier, J., 2004. Risk and risk aversion,[w:] alexander c., shedy e. In: The Professional Risk Managers' Handbook. A Comprehensive Guide To Current Theory and Best Practices, Vol. 1. p. 19.

Shapley, L.S., 1951. Notes on the N-Person Game–I: Characteristic-Point Solutions of the Four-Person Game. Rand Corporation.

Welch, I., Goyal, A., 2008. A comprehensive look at the empirical performance of equity premium prediction. Rev. Financ. Stud. 21 (4), 1455–1508.