DOĞUŞ ÜNİVERSİTESİ DERGİSİ

DOGUS UNIVERSITY JOURNAL

e-ISSN: 1308-6979

https://dergipark.org.tr/tr/pub/doujournal

MACHINE LEARNING FOR CROSS-SECTIONAL RETURN PREDICTABILITY: EVIDENCE FROM GLOBAL STOCK MARKETS [*]

MAKİNE ÖĞRENMESİ İLE YATAY KESİT HİSSE SENEDİ GETİRİLERİNİN TAHMİN EDİLEBİLİRLİĞİ: KÜRESEL PİYASALARDA AMPİRİK BİR ANALİZ

Ahmet Salih KURUCAN[1], Ali HEPŞEN[2]

Abstract: This work examines cross-sectional stock returns with machine learning models using global stock market data. By calculating 63 firm level characteristics, we find that our model outperforms linear models in terms of both economic and statistical performance. Shallow models, such as gradient boosted decision trees, provides more consistent and reliable performance compared to deeper ones in the context of asset pricing, likely due to a low signal-to-noise ratio and sensitivity to parameters. The results revealed that machine learning models can be developed into effective portfolios, complexity is welcomed when it enhances performance such as Sharpe ratios. Taken together, these results demonstrate the relative importance of machine learning for a modern financial system, and specifically, the ability to synthesize information from various characteristics that impact stock returns. This study challenges traditional notions of a preference for parsimony and, based on certain degrees of complexity, demonstrates strategic economic gains.

Keywords: Machine Learning, Asset Pricing, Equity Risk Premium, Predictive Modeling, Return Predictability, Financial Analysis

JEL: *C52, C55, C58, C61, G0, G12, G17*

Öz: *Bu çalışma, küresel hisse senedi piyasa verilerini kullanarak makine öğrenimi modelleriyle incelemektedir. 63 Firma karakteristiği hesaplayarak, makine öğrenme yöntemlerini uyguladığımızda modelimizin ekonomik kazanım ve performans açısından daha iyi sonuçlar göstermiştir. Derin öğrenme modelleriyle karşılaştırıldığında gradyan güçlendirmeli regresyon ağaçları, derin öğrenme modellerine kıyasla, büyük olasılıkla düşük sinyal-gürültü oranı, derin modellerinin hiper-parametrelere karşı yüksek hassasiyet göstermesi daha tutarlı ve güvenilir sonuçlar vermektedir. Sonuçlar makine öğrenmesi yöntemlerinin başarılı portföyler oluşturmak için de kullanılabileceğini göstermektedir. Ayrıca, model karmaşıklığını artırmanın Sharpe oranlarında iyileşmeler gibi ekonomik faydalar sağladığı gösterilmektedir. Sonuçlar, makine öğrenimi modellerinin tutarlılığını ve genelleme yeteneğini vurgulayarak, modern finansal sistemde makine öğreniminin önemini ortaya koymaktadır. Bu çalışma, sadelik ilkesine dair geleneksel anlayışları*

[*] This study is a part of the Ph.D. Thesis of Ahmet Salih Kurucan under the supervision of Prof. Dr. Ali Hepşen.

[1] Istanbul University, Institute of Social Sciences, Department of Finance; salihkurucan@gmail.com, ORCID: 0000-0002-1223-1629

[2] Istanbul University, Institute of Social Sciences, Department of Finance; ali.hepsen@istanbul.edu.tr , ORCID: 0000-0002-3379-70903

Geliş/Received: 16-08-2024; Kabul/Accepted: 06-09-2024

Atıf bilgisi: Kurucan, A. S. and Hepşen, A. (2025). Machine learning for cross-sectional return predictability: Evidence from global stock markets. *Doğuş Üniversitesi Dergisi*, *26*(1), 315-338, DOI: 10.31671/doujournal.1534375

sorgulamakta ve belirli bir karmaşıklık derecesine dayalı olarak stratejik ekonomik kazançlar gösterdiğini ortaya koymaktadır.

Anahtar Kelimeler: *Makine Öğrenimi, Varlık Fiyatlaması, Hisse Senedi Risk Primi, Tahminsel Modelleme, Getirilerin Tahmin Edilebilirliği, Finansal Analiz*

1. Introduction

The ability to forecast returns on stocks has been one of the primary concerns in empirical asset pricing for some time. Traditional methods have conventionally focused on historical average excess returns and a restricted set of prespecified factors in generating forecasts. Seminal models, among them the Capital Asset Pricing Model and the Fama-French three-factor model, made some useful contributions, but their strong reliance on a highly constrained number of factors badly restricts their adaptability and explanatory power within dynamic market environments. The deficiencies of these methods are especially significant in their inability to adapt to changing market conditions and to capture complex, nonlinear relationships existing within financial data.

The classical principle of parsimony, arguably most articulated for in the context of time series analysis by (Tukey, 1961), is to use no more parameters than is necessary to adequately represent a model. This worldview, rooted in avoiding overfitting and promoting generalization of models, appears quite starkly contrasted with the enormous parameterizations hall-marking modern ML models. While econometricians might view such massive parameterization as excessive and prone to poor out-of-sample performance, recent developments in ML challenge this notion. It has been found that in very high-dimensional domains like natural language processing and image recognition, models of extraordinarily high complexity often yield superior out-of-sample performance, thus reverting to the traditional paradigm of simplicity. The emerging consensus is that moving away from simplification, toward properly specified nonlinear models embracing complexity and all relevant predictors, will generally lead to improved predictability and portfolio performance. (Kelly & Malamud, 2021)

The advent of ML techniques has opened a promising avenue that could leverage high-dimensional data to improve the accuracy and robustness in return predictions. This study focuses on another fast-growing area: using ML models to understand cross-sectional stock returns with a comprehensive dataset from global stock markets in the paper. Our core interest lies in assessing whether ML models that capture non-linear interactions and complex relationships between an incredibly large array of predictors improve their economic and statistical performance relative to traditional linear models. Our dataset has global coverage; therefore, the findings derived are applicable across a wide spectrum, making our results relevant for investors operating in a wide variety of markets.

With the application of machine-learning techniques, such as gradient boosted regression trees and neural networks, the possibility of unlocking some complex patterns in data often missed by linear models will arise. Such models perform well with large numbers of predictors and nonlinear interactions between them. This flexibility can let ML models adapt to the intricacies of financial markets and hence potentially lead to more accurate and most importantly reliable return forecasts.

Our study provides evidence that ML models not only bring improvements in terms of predictive accuracy for investors but also economic outcomes and reliability. It shows that there is substantial net performance and alpha (excess return) for the ML-based return forecasts, therefore indicating their potential towards generating excess returns. This is distinctly meaningful for mean-variance investors for whom even modest improvements in out-of-sample explanatory power could result in meaningful economic gains.

The paper reviews a literature of recent studies. It then describes how we conducted our study: the ML models used and the global dataset for the analysis. Next, it reveals our empirical findings by presenting evidence on the performance of the ML models vis-à-vis traditional linear models.

2. Literature Overview

There are two major strategies in stock return prediction: Namely, the one based on time-series analysis (Atsalakis & Valavanis, 2009) and the one based on cross-sectional analysis (Subrahmanyam, 2010).

One of the most significant interests in a cross-sectional analysis lies in finding factors that have strong predictive powers to the expected return in the cross-section. The Fama-French three-factor model (Fama & French, 1992) (Fama & French, 1993) is one of the nominal works in this field. They argued that the cross-sectional structure of the stock price can be explained by three factors: Namely, the beta, the size, and the value. (Fama & French, 2015) adds two additional factors-profitability and investment. The model's inability to fully capture the returns of small, high-investment, and low-profitability firms remains a significant limitation. The inconsistent results across different regions suggest that the model may not be universally applicable, indicating a need for further research. (Fama & French, 2017) extends the five-factor asset pricing model to international markets specifically examining North America, Europe, Japan, and Asia Pacific. The model's underperformance in Japan and its failure to explain the low returns of certain small stocks suggest limitations in its current form. The study focuses heavily on regression-based analysis without delving into potential economic reasons behind the observed anomalies, leaving some interpretive questions unanswered.

In recent years, ML has been increasingly applied to the field of asset pricing and portfolio construction. (Heaton et al., 2016) develop a neural network for portfolio selection.(Harvey & Liu, 2021) study the multiple comparisons problem using a bootstrap procedure. (Giglio & Xiu, 2021) use dimension reduction methods to estimate and test factor pricing models. (Moritz & Zimmermann, 2016) apply tree based models to portfolio sorting, (Kozak, 2019) use shrinkage and selection methods for nonlinear function of expected stock returns.

(Gu et al., 2020) stands out as a valuable contribution to the field by conducting a comprehensive comparative analysis of ML methods applied to empirical asset pricing, demonstrating that ML techniques, significantly enhance predictive accuracy and provide substantial economic gains over traditional regression-based approaches. (Rasekhschaffe & Jones, 2019), (Bryzgalova et al., 2020), (Drobetz & Otto, 2020), (Tobek & Hronec, 2021), (Bali et al., 2020), (Dillschneider, 2022) and (Leippold et al., 2022) also document that more complex ML models are superior to linear models.

(Swade et al., 2023) discuss the need for more stringent methodologies in the evaluation and acceptance of new factors within the "factor zoo" to ensure that they are truly robust and not the result of overfitting or other statistical anomalies.

(Kelly & Xiu, 2023) reviewed the comprehensive integration of ML techniques into financial market analysis with a view to establishing their added value for empirical models. The overall purpose of the study is to provide a general review and comparison of performance in different models related to asset pricing, oriented toward shedding some light on the relative importance of different factors and how they impact predictive power for the models.

3. Methodology

Our study predicts cross-sectional stock returns using global stock market data, applying the advanced ML methods. The main body of the project is organized by enhancing the accuracy and robustness of return predictions. This section delineates how the dataset shall be prepared, along with various ML methods used for training purposes.

There is an intensive phase of cleaning the dataset. In this study the target variable will be the 12-month return. This target variable is of importance since it expresses the financial metric that we would like our ML models to predict. This means that the data needs to be sorted, and the training data is organized by country, stock, and date. This way, we make sure to side-step the forward-looking bias by assuming monthly characteristics lagged by 1 month, quarterly at least 4 months, and annual at least 6 months.

We set up a train-test splitting strategy. We fix a start date for training, which is year 2003, and then define a number of test periods. For each country, this predefined date range will cut the data into a training set and a test set avoiding possible data leakage to avoid overfitting in models.

Feature engineering is one of the most critical steps in the process of data preparation. Handling missing data by replacing it with the column means is a strategy in which missing values are filled. This way, the models can be trained without interruptions because of missing values, and it helps to preserve the dataset's statistical properties. Following this, the focus is on firm-level characteristics only, putting aside macroeconomic attributes to research intrinsic drivers of stock returns at the level of single companies and hence to focus on return predictability in the cross-section of stocks.

Following (Gu et al., 2020), we employ a general additive prediction model to describe the excess return of a stock, which can be written as:

$$r_{i,t+l} = E_t\left(r_{i,t+l}\right) + \varepsilon_{i,t+l}, \tag{1}$$

where $(r_{i,t+l})$ is the forward 12-months excess stock return for stocks to represent ing a long-term investing perspective, which are indexed as as $(i = 1, \dots, N_t)$ in month $(t = 1, \dots, T)$.

$$E_t\left(r_{i,t+l}\right) = g^*\left(z_{i,t}\right), \tag{2}$$

$(E_t(r_{i,t+l}))$ as a function of predictor variables that maximizes the out-of-sample explanatory power for the return $(r_{i,t+l})$ where in our case the forward 12-months expected stock return. We denote those predictors as the (P) $-$dimensional vector $(z_{i,t})$, and assume the conditional expected cumulative return $(g^*(\cdot))$ is a flexible function of these predictors. Despite its flexibility, this framework imposes some important restrictions. The function $(g^*(\cdot))$ depends neither on (i) nor (t). This contrasts with standard asset pricing approaches that reestimate a cross-sectional model each time period or that independently estimate time-series models for each stock. Additionally, $(g^*(\cdot))$ depends on (z) only through $(z_{i,t})$. This means our prediction does not use information from the history prior to (t), nor from individual stocks other than the $(i)th$.

3.1.Models

The algorithms we use are partial least squares, gradient boosted regression trees, and multilayer perceptron. We perform the usual methods from the machine learning literature (Gu et al., 2020) for estimating the models, choosing the hyperparameters, and evaluation of the prediction performance. We calculate the models separately for each market. The models are configured to predict by minimizing the root mean sqaured error on out-of-sample.

$$\mathrm{RMSE}_{t+l} = \sqrt{\frac{1}{N_{t+l}} \sum_{i=1}^{N_{t+1}} (\widehat{\varepsilon_{\iota,t+l}})^2} \,, \tag{3}$$

where $\varepsilon_{i,t+l}$ is the individual prediction error for the stock i, and N_{t+l} is the number of stocks at *t+1*.

3.1.1. Linear models

Though simplistic, linear models like simple ordinary least squares regression can serve as a baseline to which other, more sophisticated approaches can be compared. Under this assumption, with respect to the predictors and their relationship considered to be linear with the stock returns, too often it misses the real complexity of data. In order to avoid possible overfitting issues intrinsic in linear models with a large number of predictors, we introduce penalized regression method of Ridge regression. It is in these method that regularization penalties are imposed to shrink coefficients of less important variables, hence leading to improved out-of-sample prediction performance.

This naive linear model is certain to fail in the presence of many predictors. If the number of predictors, P, approaches the number of observations, T, then the linear model becomes quite inefficient or even inconsistent. It starts fitting noise instead of extracting signals. The problem with return prediction—the area where this signal-to-noise ratio is notoriously low—is especially troubling. Ridge regression introduces an L2 penalty to the ordinary least square's regression. This allows techniques such as shrinking coefficients of less important predictors towards zero and handling of multicollinearity among predictors, hence improving the predictive power of the model. A model created by ridge regression will, in some cases, offer a base from which more complex models can be compared.

3.1.2. Gradient boosted regression trees

Decision trees and their ensemble variants Random Forests and Gradient Boosted Trees are used for the capture of nonlinear interactions between predictors. These methods create recursive partitionings of the data, through which complex, nonlinear relationships may emerge without the explicit specification of interactions. Regression trees have become a popular ML approach for incorporating multiway predictor interactions and account for interactions among predictors.

CatBoost is an advanced implementation of GBRT that works much more effectively with categorical variables and further reduces overfitting. It is designed as a robust, high-performance library that can capture complex patterns in data using ensembles of decision trees. CatBoost builds trees sequentially – each tree corrects the errors made by previous ones, which has improved model accuracy incrementally.

Boosting Regression Trees, or just Boosting[1]— this is a case where we combine many decision trees into one ensemble model. While the process of bagging generated several models in parallel and then combined the results, boosting trains multiple models sequentially. Each model is trained based on the errors made by its predecessor. The weak learners used here are individual decision trees. All the trees are connected in series, and each tree tries to minimize the error of the previous tree.

The following is an overview of the GBRT algorithm: First, initialization with a constant value; then, iteratively compute the residuals of the current model, fit a decision tree to the residuals, and update the model by adding the tree's predictions scaled by a learning rate. The final prediction has to be summed up from the contributions of all trees. GBRT is an extremely powerful tool, for it iteratively corrects the mistakes made by previous trees, hence progressively improving the model with higher accuracy.

3.1.3. Neural networks

We use neural networks, including deep learning architectures, to capture complex nonlinear relationships. The models have numerous layers of interconnected nodes that enable sophisticated hierarchical learning and intricate patterning of the data. Probably one of the most powerful modeling tool. This flexibility is derived from an ability to wrap within themselves many telescoping levels of nonlinear predictor interactions, earning the synonym "deep learning." All the same, their complexity places neural networks as among the least transparent, the least interpretable, and the most highly parameterized tools.

Our analysis focuses on traditional "feed-forward" networks. An MLP neural network's objective function is to find the optimal weights and biases to help in minimizing the error between predictions of the model and the real target values. This would be through the use of a loss function quantifying the difference between the predicted and actual values, Mean Squared Error in the case of regression or Cross-Entropy Loss for classification. Moreover, in the objective function, a regularization part is added to penalize the large weights and biases to avoid overfitting so that the

[1] Boosting is originally described in (Schapire, 1990) and (Freund, 1995) for classification problems to improve the performance of a set of weak learners. (Jerome H. Friedman, 2001) extend boosting to contexts beyond classification, eventually leading to the gradient boosted regression tree.

model generalizes well to new data. This regularization thus goes on balancing the fitting of training data accurate enough and keeping the model simple.

3.2.Model evaluation and validation

Model performance will, however, be tested rigorously out-of-sample with techniques pertaining to cross-validation. Main attention in the realm of key performance metrics is accorded to R-squared, Mean Absolute Error, and Sharpe Ratio, along with both predictive accuracy and economic significance. What's more, we do robustness checks so that the predictions from our models are not sensitive to any particular assumptions or data partitioning. To assess predictive performance for individual excess stock return forecasts, we calculate the out-of-sample (R^2) as

$$R^2_{\text{oos}} = 1 - \frac{\sum_{(i,t)\in\mathcal{T}_3}\left(r_{i,t+l} - \widehat{r_{i,t+l}}\right)^2}{\sum_{(i,t)\in\mathcal{T}_3} r^2_{i,t+l}},$$

Our approach will be seeking predictive models that are guaranteed to be not only accurate but also generalizable across different market conditions and datasets by combining ML techniques and statistical methodologies. In the final step, evaluation of the trained models will be done with test data. Several of the more key performance metrics computed would be MAE and R-squared, through which one can evaluate model accuracy and the power of explanation. The MAE provides an average of the size of the errors in the predictions, while the R-squared measures how well a model explains the variability of the target variables.

One-way pairwise comparison of methods is afforded by the (Diebold & Mariano, 1995) test for differences in out-of-sample predictive accuracy between two models.

The Diebold-Mariano test is a statistical test used to compare the predictive accuracy of two competing forecasts. The test's main advantage lies in the model-free property and does not ask for the generated forecasts from nested models. Specifically, this test statistic can be simply written as the difference between forecast errors from two models, adjusted for possible autocorrelation and heteroskedasticity in the differences between forecast errors. In particular, if e1,t and e2,t are the forecast errors from the two models at time t, then the test statistic is computed as:

$$DM = \frac{\bar{d}}{\sqrt{\widehat{\sigma^2}/T}} \tag{4}$$

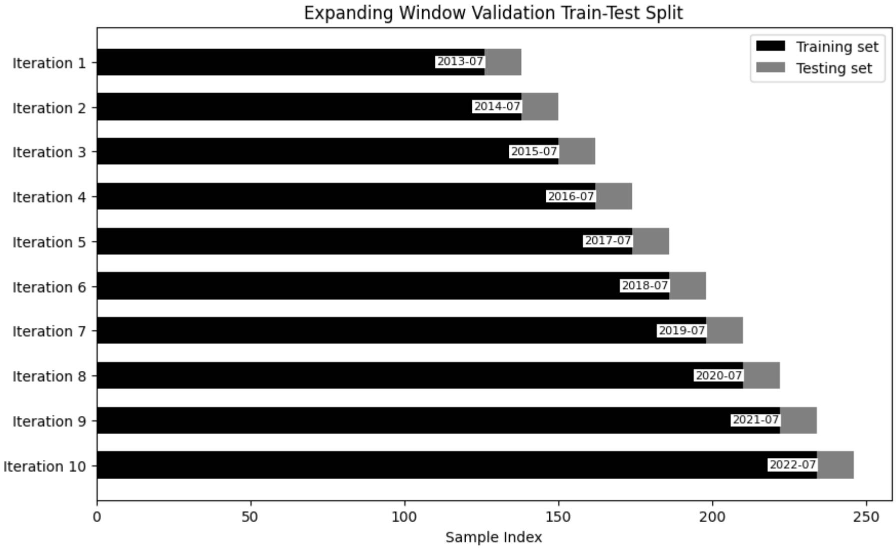
where ($\bar{d}$) is the mean of the loss differential ($d_t = e_{1,t} - e_{2,t}$), and ($\widehat{\sigma^2}$) is an estimator of the variance of (d_t), which can be consistently estimated using the Newey-West method to account for autocorrelation. Under the null hypothesis of equal predictive accuracy, the DM statistic is asymptotically standard normal. This methodology has been appropriately modified in a number of empirical contexts—including stock-level prediction errors, where strong cross-sectional dependence may be present—by comparing the cross-sectional average of prediction errors rather than individual stock returns. It is important to emphasize that the model-free nature of the

Diebold-Mariano test means it should be interpreted as a comparison of forecasts and not as a comparison of fully articulated econometric models.

3.2.1. Expanding-window validation

To ensure meaningful evaluation, we adopted the expanding window approach for assessing our model's performance. This alone entailed fitting our model on data from a start date up to some point in the series, using the next year for prediction and calculation of an R2 score. We repeated this for all other consecutive years while also increasing the window of data used in training and making predictions with the model. This gave us a full insight into the model's capability to make correct predictions throughout our analysis period. We can in this way emulate use of the models in real scenarios where the investors add new data collected over time.

Figure 1: Expanding-window validation train-test split



Each test period is subsequent in time to the related training period. There is no futures information because the training data stops just before each test period begins. Running the performance of the model over multiple test periods can proxy an investor's learning curve. In this way, experimenters understand how an investor's predictive accuracy improves over time during many iterations when more data becomes available and the model is re-trained periodically.

3.3.Factor importances

The method we followed in order to use variables of major importance to our model was based on the panel predictive R^2. For each predictor j, we computed the reduction in panel predictive R^2. obtained by setting all its values to zero while keeping all other model estimates fixed. If this score deteriorated appreciably, we decided that this factor was important for the predictive power of the model.

Global performance across the multiple test periods confirms these will not be factors that the model uses that lose their predictive power over time. More importantly, this is integral to building a robust model that doesn't rely on transient patterns or anomalies. We are to ascertain, looking at feature importance and consistency across test periods, that our model is leveraging stable and reliable predictors.

This expanding window cross-validation approach will ensure that our models are evaluated very close to real-world applications; hence, it gives us a very good and reliable assessment of their predictive capabilities in financial markets. This evaluation will assess not only the generalization ability of the model on future data but also track improvement in investor learning and verify if the predicting factors are stable.

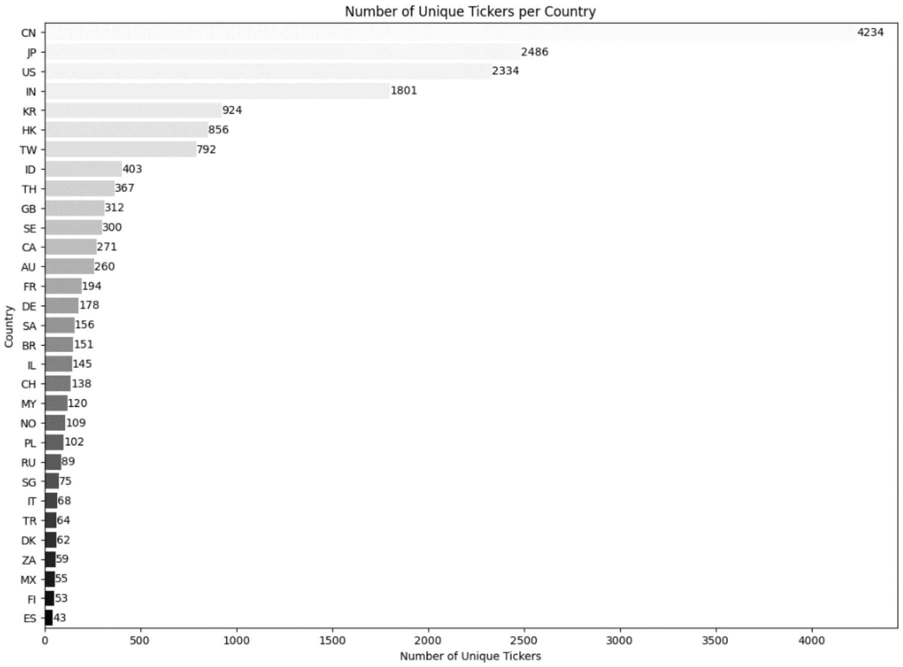
In this study, the detailed analysis of the predictability of stock returns using proper cleaning of data, feature engineering, and implementation of ML techniques are duly attempted.

4. An Empirical Study of Global Equities

This analysis is well-grounded on a diverse dataset with global stock market data. Our sample comprises around 15 thousand companies, spanning 31 countries and 12 sectors. The dataset begins in the year 2003, deliberately chosen despite the availability of earlier data for developed markets. This choice reflects our focus on capturing the nuances of emerging and frontier markets that have become more significant in the global perspective from this period onward with a dataset of balanced stocks. Such broad coverage is indispensable for tapping into the unique economic conditions, varied market structures, and diverse investor behaviors that characterize the global equity markets.

The study uses firm-level and price-related features as predictor variables, including 28 price-related features with rolling and other transformations, volume, momentum, volatility, along with sector-based features like industry dummies. It also includes 63 pertinent financial ratios drawn from financial statements and 151 features engineered from the core dataset. Stocks are part of this dataset representing a corresponding uniquely existing publicly traded company.

Figure 2. Number of stocks per country



Overall, the dataset's diversity in terms of geography, market size, and sectoral representation forms a robust basis for applying ML models to forecast stock returns.

5. Emprical Results

The empirical findings of this study provide several important insights into the predictability of cross-sectional stock returns across different global markets. In particular, we notice anomalies in the evolution of R^2 over time. These crises have caused huge market volatility and disruptions, which could affect the relationships between variables in our models and the cross-sectional equity returns. During periods of economic uncertainty and financial distress, there can be huge changes in investor behavior, sentiments of the market, and risk aversion.

5.1.Model performance and comparison

Traditional linear models gave a baseline and showed something like limited predictive power against which to compare other methods in this high-dimensional setting. Correspondingly, the values of out-of-sample R-squared were relatively low for linear models, which explains that much variability in stock returns could not be explained by these models alone.

Table 1. Monthly out-of-sample stock-level prediction performance of penalized linear model (percentage R^2)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Country	2013-07 2014-06	2014-07 2015-06	2015-07 2016-06	2016-07 2017-06	2017-07 2018-06	2018-07 2019-06	2019-07 2020-06	2020-07 2021-06	2021-07 2022-06	2022-07 2023-04
CN	**0,279**	**0,265**	**0,039**	-0,486	-0,361	-0,09	-0,034	-0,048	-0,006	-0,02
JP	**0,239**	-0,035	**0,059**	**0,277**	-0,248	-0,09	**0,017**	-0,036	-0,013	**0,202**
US	**0,159**	-0,209	**0,208**	**0,212**	**0,001**	**0,056**	**0,157**	**0,087**	-0,246	**0,118**
IN	**0,035**	-0,019	**0,044**	**0,004**	-0,109	-0,033	-0,001	**0,003**	-0,008	**0,007**
EU	**0,131**	**0,082**	**0,157**	**0,176**	-0,212	**0,012**	**0,141**	**0,079**	-0,251	**0,068**
KR	**0,026**	-0,027	-0,032	**0**	-0,021	-0,015	-0,001	-0,003	**0,001**	**0,002**
HK	-0,007	-0,035	-0,002	-0,003	**0,003**	**0,047**	**0,063**	**0,055**	**0,118**	**0,119**
TW	**0,005**	-0,106	**0,008**	**0,028**	-0,014	**0,028**	**0,055**	**0,018**	-0,028	**0,091**
ID	-0,024	-0,14	**0,006**	-0,034	-0,04	-0,011	**0,006**	**0,007**	**0,022**	**0,027**
TH	**0,219**	-0,396	**0,072**	-0,132	-0,263	-0,196	-0,055	**0,077**	-0,131	-0,153
GB	**0,164**	**0,103**	**0,266**	**0,366**	-0,097	**0,023**	**0,2**	**0,237**	-0,565	-0,084
CA	-0,009	-0,26	**0,039**	-0,06	-0,16	-0,053	**0,034**	**0,021**	-0,096	**0,008**
AU	-0,022	-0,069	**0,086**	**0,1**	-0,019	-0,004	**0,047**	**0,039**	-0,144	**0,079**
BR	-0,193	-0,233	**0,062**	**0,139**	**0,001**	**0,149**	**0,092**	-0,173	-0,096	**0,12**
CH	**0,25**	**0,195**	**0,347**	**0,453**	-0,469	-0,064	**0,257**	**0,234**	-0,482	-0,047
MY	**0,065**	-0,161	**0,08**	**0,068**	-0,131	-0,06	**0,001**	**0,007**	-0,082	**0,081**
NO	-0,005	-0,159	**0,074**	**0,065**	-0,008	-0,002	**0,111**	**0,016**	-0,086	**0,042**
RU	-0,035	-0,11	-0,07	**0,015**	**0,029**	-0,082	-0,08	-0,022	**0,003**	-0,053
SG	**0,081**	-0,609	**0,052**	**0,233**	-0,348	**0,005**	**0,105**	**0,136**	-0,304	-0,29
TR	**0,371**	-0,51	**0,144**	**0,4**	-0,552	**0,352**	**0,305**	**0,215**	**0,37**	**0,395**
MX	**0,079**	-0,202	-0,096	-0,221	-0,425	-0,315	-0,075	**0,083**	-0,133	**0,136**

The neural network MLP model had better predictive power than a simple linear model because some of the nonlinearities or interactions among predictors were captured, which the linear model failed to do. However, while performing better than the linear approach, improvements were still modest and inadequate to handle the complexities of the data fully. Although MLP models are more powerful in capturing nonlinear relationships, they are not generally well-suited for tabular data, very common in financial datasets. Neural networks do a good reverse engineering in unstructured data like images or text but sometimes fail on structured data and relations inside of it that are present in tabular data.

Furthermore, neural network models have hyperparameters that need tuning, such as the number of layers, number of neurons, and learning rate; this could turn out to be very complex and resource-intensive. This complexity often comes at the cost of interpretability paid in full. The lack of transparency possibly becomes a major limitation in areas like finance, where it is foremost to understand what factors drive model decisions.

Table 2. Monthly out-of-sample stock-level prediction performance of neural network model (percentage R^2)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Country	2013-07 2014-06	2014-07 2015-06	2015-07 2016-06	2016-07 2017-06	2017-07 2018-06	2018-07 2019-06	2019-07 2020-06	2020-07 2021-06	2021-07 2022-06	2022-07 2023-04
CN	**0,324**	**0,236**	0,031	-0,46	**-0,364**	-0,069	-0,048	-0,06	**-0,013**	**0,001**
JP	**0,239**	**-0,028**	**0,03**	**0,369**	-0,181	-0,069	**0,016**	**-0,025**	**-0,009**	**0,231**
US	**0,154**	-0,218	**0,206**	**0,207**	**-0,024**	**0,049**	**0,152**	**0,088**	-0,246	**0,12**
IN	**0,024**	**-0,01**	**0,063**	**-0,001**	-0,117	**-0,004**	**-0,016**	**0,002**	**0,014**	**-0,013**
EU	**0,101**	**0,089**	**0,153**	**0,153**	-0,218	**0,024**	**0,159**	**0,077**	-0,174	**-0,001**
KR	**0,015**	**-0,036**	-0,056	**-0,001**	-0,022	**-0,046**	**0**	**0,008**	**-0,001**	**0,008**
HK	**-0,001**	**0,054**	**0**	**-0,021**	**0,024**	**0,004**	**0,094**	**-0,305**	**0,08**	**0,122**
TW	**-0,015**	-0,34	**0,057**	**0,058**	**-0,013**	**0,046**	**0,142**	**0,005**	**-0,024**	**0,061**
ID	**-0,039**	-0,097	**0,008**	**0**	**-0,004**	**-0,029**	**0,029**	**0,013**	**0,026**	**0,017**
TH	**0,216**	-0,677	**0,076**	**-0,075**	**-0,13**	-0,284	-0,066	**0,089**	**-0,187**	**-0,388**
GB	**0,132**	**0,109**	**0,254**	**0,371**	**-0,172**	**-0,029**	**0,195**	**0,228**	-0,518	-0,13
CA	**0,003**	-0,273	**0,036**	**-0,075**	**-0,154**	**-0,045**	**0,034**	**0,02**	**-0,124**	**0,011**
AU	**-0,301**	**-0,065**	**0,006**	**0,074**	**-0,074**	**-0,146**	**0,048**	**0,039**	**-0,012**	**0,038**
BR	-0,167	**-0,182**	**0,025**	**0,063**	**0,039**	**0,219**	**0,046**	-0,081	**-0,041**	**0,094**
CH	**0,247**	**0,183**	**0,347**	**0,434**	-0,397	**-0,077**	**0,269**	**0,129**	-0,547	**-0,001**
MY	**-3,136**	**-0,144**	**0,071**	**0,067**	-3,797	**-0,305**	**-0,099**	**-0,102**	**-0,191**	**0,097**
NO	**-0,016**	**-0,131**	**0,074**	**0,089**	**-0,208**	**0,01**	**0,094**	**0,021**	**-0,053**	**0,048**
RU	**-0,382**	**-0,177**	**-0,209**	**0,014**	**-0,009**	**-0,005**	**-0,219**	**-0,04**	-0,014	**0,139**
SG	**-0,35**	-0,894	**0,027**	**0,161**	-0,233	**-0,001**	**-0,075**	**0,106**	-0,402	**-0,563**
TR	**0,354**	-0,49	**0,145**	**0,407**	-0,478	**0,357**	**0,292**	**0,227**	**0,288**	**0,404**
MX	**0,08**	-0,38	**-0,051**	-0,055	-0,357	-0,471	**-0,089**	**0,083**	**-0,205**	**0,131**

Among the primary advantages of GBRT is handling complex interactions between predictors through averaging across multiple decision trees. This ensemble approach greatly reduces variance, hence improving the stability of the model and making it more robust to overfitting. The shallow models outperform the deep ones, likely due to the fact that the data is sparse in nature and contains a low signal-to-noise ratio within asset pricing. Each tree in an ensemble adds incrementally to the final prediction—both additive and transparent—hence allowing clear understanding of how the model arrives at the conclusion.

Compared to neural networks and other complex models, GBRT models are relatively light in terms of computational resources for training and do not require extensive hyperparameter tuning. One of the definite advantages of their transparency is that users can literally see how different predictors differ with respect to importance and how they contribute towards different outputs from the model. This type of interpretability is highly valued, particularly in the finance domain, where much is gained through interpretability in order to build trust and achieve regulatory compliance.

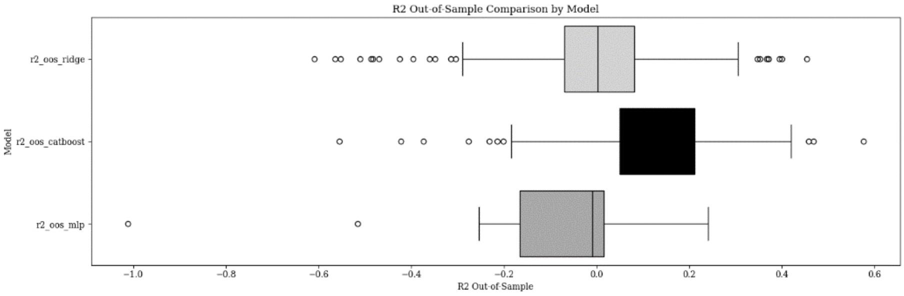
Although GBRT models did bring slight improvement in out-of-sample R-squared values over the linear models, they do show a level of predictive accuracy. In general, the GBRT model represents that sweet spot among interpretability, resource efficiency, and predictive power, which makes it competitive in many predictive modeling tasks—most centrally, those in which clarity into the logic of decision-making comes nearly as important as the accuracy of these predictions.

Table 3. Monthly out-of-sample stock-level prediction performance of GBRT model (percentage R^2)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Country	2013-07 2014-06	2014-07 2015-06	2015-07 2016-06	2016-07 2017-06	2017-07 2018-06	2018-07 2019-06	2019-07 2020-06	2020-07 2021-06	2021-07 2022-06	2022-07 2023-04
CN	**0,19**	**0,217**	-0,076	-0,201	**0,149**	-0,038	-0,014	-0,006	**0,001**	**0,051**
JP	**0,373**	**0,004**	**0,085**	**0,327**	-0,11	-0,088	**0,07**	**0,154**	**0,081**	**0,259**
US	**0,222**	-0,098	**0,26**	**0,267**	**0,108**	**0,146**	**0,229**	**0,213**	-0,163	**0,159**
IN	**0,072**	**0,058**	**0,11**	**0,129**	-0,025	**0,114**	**0,058**	**0,05**	**0,039**	**0,088**
EU	**0,221**	**0,199**	**0,248**	**0,281**	-0,072	**0,159**	**0,184**	**0,207**	-0,096	**0,2**
KR	**0,09**	**0,022**	-0,008	**0,071**	-0,005	**0,048**	**0,09**	**0,057**	**0,023**	**0,111**
HK	**0,104**	**0,139**	**0,169**	**0,243**	**0,264**	**0,266**	**0,248**	**0,209**	**0,201**	**0,189**
TW	**0,061**	-0,013	**0,142**	**0,132**	**0,133**	**0,115**	**0,108**	**0,149**	**0,058**	**0,205**
ID	**0,032**	-0,044	**0,01**	**0,108**	**0,125**	**0,117**	**0,123**	**0,07**	**0,177**	**0,144**
TH	**0,269**	-0,185	**0,145**	**0,063**	**0,035**	-0,02	-0,03	**0,09**	**0,191**	**0,149**
GB	**0,212**	**0,227**	**0,308**	**0,419**	**0,033**	**0,131**	**0,285**	**0,35**	-0,555	-0,027
CA	**0,212**	-0,115	**0,207**	**0,146**	**0,169**	**0,089**	**0,203**	**0,276**	**0,09**	**0,114**
AU	**0,143**	**0,206**	**0,209**	**0,233**	**0,246**	**0,144**	**0,202**	**0,288**	**0,012**	**0,202**
BR	-0,109	**0,027**	**0,172**	**0,245**	**0,079**	**0,223**	**0,158**	-0,149	**0,081**	**0,207**
CH	**0,262**	**0,232**	**0,458**	**0,468**	-0,423	**0,204**	**0,288**	**0,356**	-0,373	**0,066**
MY	**0,089**	**0,075**	**0,259**	**0,178**	-0,032	**0,037**	**0,109**	**0,065**	**0,088**	**0,139**
NO	**0,074**	**0,145**	**0,239**	**0,312**	**0,29**	**0,059**	**0,269**	**0,224**	**0,045**	**0,116**
RU	**0,095**	**0,084**	**0,09**	**0,218**	**0,105**	**0,049**	**0,113**	**0,123**	-0,032	**0,115**
SG	**0,255**	-0,276	**0,179**	**0,225**	-0,03	**0,122**	**0,214**	**0,282**	-0,06	**0,096**
TR	**0,346**	-0,214	**0,142**	**0,377**	-0,164	**0,406**	**0,419**	**0,26**	**0,395**	**0,575**
MX	**0,113**	-0,129	**0,065**	-0,117	-0,232	-0,062	**0,095**	**0,179**	**0,139**	**0,087**

Below is the boxplot comparing the R^2 out-of-sample for three models: Ridge Regression, CatBoost which is a GBRT model and MLP. Of these, the CatBoost model represented in black performed best. It had a higher median of R^2 values and a tighter interquartile range. This means it gives relatively more consistent predictions as compared with both the MLP and Ridge Regression models. That means that the rather simple and partially interpretable GBRT model was at least as good as the more complex MLP and the penalized linear model, which testifies to its efficiency in the grasping of underlying patterns in data.

Figure 3. Summary of R^2 out-of-sample performance by model



5.2. Diebold-Mariano tests

A granular comparison of predictive accuracy between the GBRT model and penalized linear regression model for different countries and time periods is presented. Based on these results, evidence indicates that the GBRT model outperforms a penalized linear regression model. Moreover, the p-values corresponding to these results further confirm that the GBRT model has very high statistical significance superiority, hence contributing to its reliability in various contexts. There are periods and regions, specifically in China during certain time frames, when the GBRT model fails to perform better than the penalized linear regression model; this is characterized by negative DM values. Notwithstanding these instances, the GBRT model attains an overall winning streak natured from the data. This trend underlines how important it is to have more advanced models, like GBRT, which are heavier on computation in exchange for better predictive accuracy and stability across different markets and time frames.

Table 4. Diebold-Mariano tests per year and country

Country	1	2	3	4	5	6	7	8	9	10
	2013-07 2014-06	2014-07 2015-06	2015-07 2016-06	2016-07 2017-06	2017-07 2018-06	2018-07 2019-06	2019-07 2020-06	2020-07 2021-06	2021-07 2022-06	2022-07 2023-04
CN	-63,556	-30,189	-31,118	**83,628**	**124,912**	**16,866**	**13,754**	**16,506**	**3,593**	**16,068**
JP	**35,137**	**6,764**	**5,532**	**26,289**	**25,629**	**0,482**	**12,752**	**39,175**	**22,011**	**29,328**
US	**9,537**	**13,614**	**10,201**	**11,584**	**19,249**	**19,389**	**21,823**	**30,188**	**16,775**	**9,729**
IN	**12,755**	**10,131**	**11,999**	**22,64**	**13,814**	**32,164**	**15,18**	**8,066**	**8,73**	**19,155**
EU	**10,294**	**13,102**	**12,683**	**15,585**	**12,525**	**16,371**	**7,176**	**16,433**	**16,712**	**15,169**
KR	**10,458**	**7,065**	**3,24**	**10,764**	**2,624**	**9,662**	**14,994**	**9,786**	**5,404**	**14,41**
HK	**12,349**	**16,54**	**17,973**	**28,423**	**31,175**	**27,049**	**22,127**	**18,362**	**9,06**	**6,468**
TW	**6,644**	**12,258**	**17,17**	**13,024**	**17,867**	**11,425**	**9,465**	**18,182**	**9,101**	**14,893**
ID	**4,349**	**6,692**	**0,472**	**10,942**	**11,623**	**11,48**	**9,511**	**7,356**	**11,591**	**8,985**
TH	**5,39**	**9,231**	**8,583**	**11,464**	**16,215**	**13,919**	**1,938**	**2,411**	**14,386**	**14,992**
GB	**2,537**	**8,007**	**4,72**	**6,046**	**5,883**	**7,048**	**10,209**	**10,705**	**0,604**	**5,575**
CA	**11,019**	**6,099**	**7,952**	**10,877**	**14,033**	**8,581**	**10,788**	**18,565**	**8,922**	**5,105**
AU	**9,918**	**11,488**	**8,088**	**5,82**	**11,886**	**9,094**	**9,148**	**14,224**	**6,204**	**5,036**
BR	**3,249**	**8,903**	**4,022**	**6,263**	**2,968**	**5,499**	**4,56**	**3,257**	**8,06**	**3,658**
CH	**0,495**	**1,338**	**3,864**	**0,864**	**1,31**	**7,183**	**2,179**	**7,708**	**3,503**	**3,453**
MY	**1,696**	**9,501**	**5,927**	**3,53**	**4,682**	**4,634**	**4,721**	**3,275**	**5,537**	**3,165**
NO	**3,067**	**7,618**	**6,186**	**6,69**	**6,871**	**1,767**	**5,189**	**9,89**	**4,336**	**1,706**
RU	**4,868**	**4,485**	**5,836**	**6,488**	**4,13**	**5,39**	**6,548**	**4,355**	-0,268	**6,175**
SG	**4,076**	**10,198**	**3,4**	-0,667	**6,211**	**5,054**	**4,86**	**5,296**	**5,032**	**6,784**
TR	-1,665	**6,929**	-0,101	-6,669	**11,237**	**2,887**	**6,697**	**2,13**	**3,415**	**5,43**
MX	**1,579**	**2,709**	**3,985**	**2,162**	**5,553**	**9,59**	**5,207**	**3,813**	**6,455**	-1,306

The below table shows the *p-values* of DM test results. The statistically-significant results emphasize that the GBRT model not only performs better overall but also adapts more effectively to the complexities.

Table 5. P-values of Diebold-Mariano test per test period and country

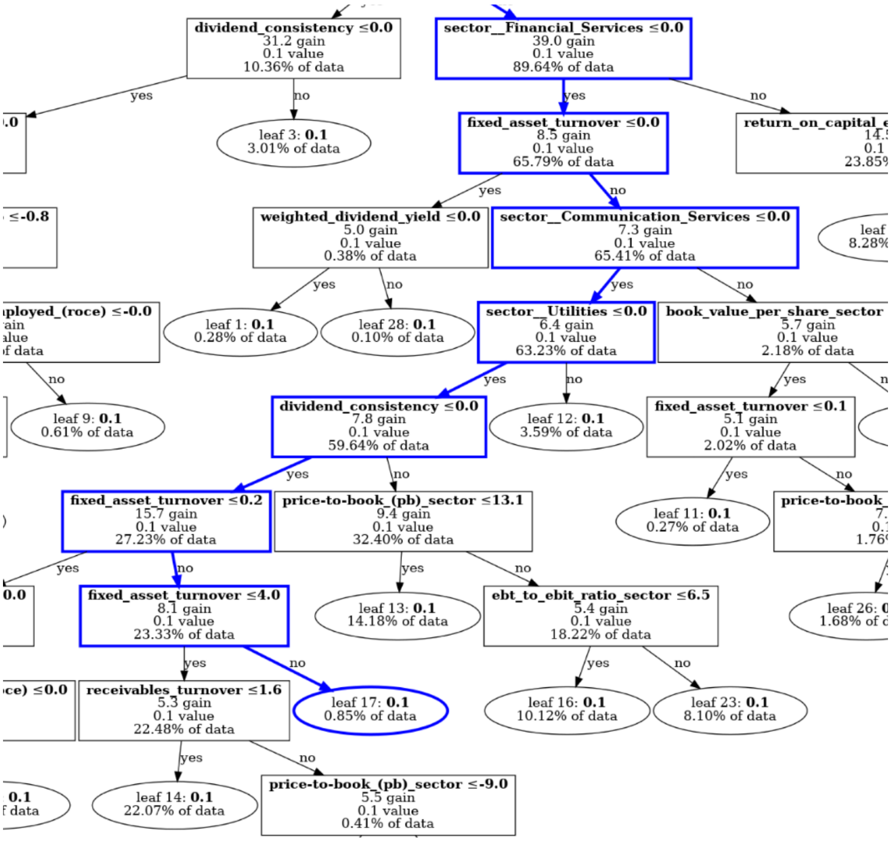
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Country	2013-07 2014-06	2014-07 2015-06	2015-07 2016-06	2016-07 2017-06	2017-07 2018-06	2018-07 2019-06	2019-07 2020-06	2020-07 2021-06	2021-07 2022-06	2022-07 2023-04
CN	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
JP	**0**	**0**	**0**	**0**	**0**	0,63	**0**	**0**	**0**	**0**
US	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
IN	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
EU	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
KR	**0**	**0**	0,001	**0**	0,009	**0**	**0**	**0**	**0**	**0**
HK	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
TW	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
ID	**0**	**0**	0,637	**0**	**0**	**0**	**0**	**0**	**0**	**0**
TH	**0**	**0**	**0**	**0**	**0**	**0**	0,053	0,016	**0**	**0**
GB	0,011	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0,546	**0**
CA	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
AU	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
BR	0,001	**0**	**0**	**0**	0,003	**0**	**0**	0,001	**0**	**0**
CH	0,621	0,181	**0**	0,388	0,19	**0**	0,029	**0**	**0**	0,001
MY	0,09	**0**	**0**	**0**	**0**	**0**	**0**	0,001	**0**	0,002
NO	0,002	**0**	**0**	**0**	**0**	0,077	**0**	**0**	**0**	0,088
RU	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0,789	**0**
SG	**0**	**0**	0,001	0,505	**0**	**0**	**0**	**0**	**0**	**0**
TR	0,096	**0**	0,92	**0**	**0**	0,004	**0**	0,033	0,001	**0**
MX	0,114	0,007	**0**	0,031	**0**	**0**	**0**	**0**	**0**	0,192

5.3. Factor Importances

Among other things, feature importance is an important aspect of ML that concerns model interpretability and decision-making. It is the method to identify and quantify the contribution of each feature in a predictive model, which allows us to interpret the variables most influential to model predictions. The literature uses terms such as "predictor," "feature,", "characteristics,", and "factor" synonymously to refer to these variables. Understanding feature importance lots not only refines the models by concentrating on the factors that most influence the outcome but also increases transparency and trust in model predictions, hereby making this component inseparable from model development and evaluation.

In terms of model explainability, Tree 0 in figure below acts as an initialization step for this GBRT model, following an iterative process to minimize prediction errors using the most informative feature, as decided in a loss function/measure by which it will split data to generate its first set of predictions. These predictions are then added to the initial model—usually just the mean of the target variable—and scaled by a learning rate parameter, which determines to what extent Tree 0 will influence the final prediction. This process lays the base for the next trees, which further improve the predictions by learning from residual errors of previous rounds. The role of Tree 0 is very important, as it gives an initial correction that guides further developments of the model, making sure each new tree contributes towards a more accurate and robust prediction. Moreover, ordered boosting, as in models like CatBoost, makes use of categorical factors both very effective and keeps off target leakage, improving predictive accuracy throughout the iterative process.

Figure 4. Section of the tree plot 0 of us country GBRT model



Our analysis identified that volatility metrics are one of the most important factors which emerged as the top factors across all markets underpinning its critical role in capturing market uncertainty and risk. Maximum drawdown for the same period ranked highly as well, given its central role in assessing possible downside risks by measuring the largest loss from peak to trough within a year.

Another central variable was enterprise value, even if it took on different meaningful levels across markets. It becomes, therefore, a very important metric that includes both debt and equity to represent a company's total value, particularly in regions with varied economic activities. At the sector level, the price-to-book ratio never moved from the top factors, indicating that relative valuation was very important in determining undervalued or overvalued sectors. This is further boosted by the high rating accorded to momentum, which captures trends in stock prices and reflects the momentum effect of past good performers continuing to be so in the short run; corresponding with rich empirical evidence (Asness et al., 2013) that picks value and momentum as the two most conspicuous and prevalent patterns in asset returns.

Another very highly prominent factor that came into the picture is an equity multiplier, a financial leverage ratio describing the overall assessment of its financial structure and risk profile of the company as a whole.

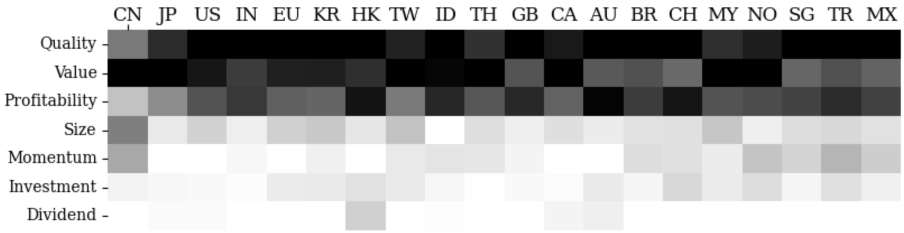
Net current asset value measures a company's difference between the current assets and liabilities as an indicator of liquidity; it showed variable importance in different markets, most rampant in regions where this aspect was really important, like the emerging markets. Dividend consistency was rather insignificant when looked at universally, but featured centrally in markets characterized by mature companies because it relates to factors for stability and attraction of income-focused investors. Tangible asset value and large capitalization were also influential, especially in heavy industries and developed markets where large-cap firms dominate.

Large-cap stocks, had greater influence in developed markets, where these companies often dominate indices and attract the most significant investor interest. While for small-cap stocks, against the backdrop of higher volatility and a risk level, local economic conditions and investor sentiment would be very critical to their reaction, hence of imperative consideration in explaining growth-oriented or more speculative investment dynamics. Not precisely sector-based, the large-cap/small-cap stock distinction is key to investment portfolio construction and risk management, as it does have an effect on the overall risk-return profile of investments within different market segments.

The figure presents the rankings considered in the study regarding their average total model contribution aggregated within categories for each country. The factor importance is calculated as the reduction of the R^2_{oos}. In developed markets, quality, value and profitability factors rank highly. In developed markets, quality, value, and profitability factors rank highest, with quality being the most significant across the markets. Value remains significant, but its prominence has declined in developed markets. Profitability factors continue to play an important role in financial environments, while investment and dividend factors have the lowest importance for return predictability.

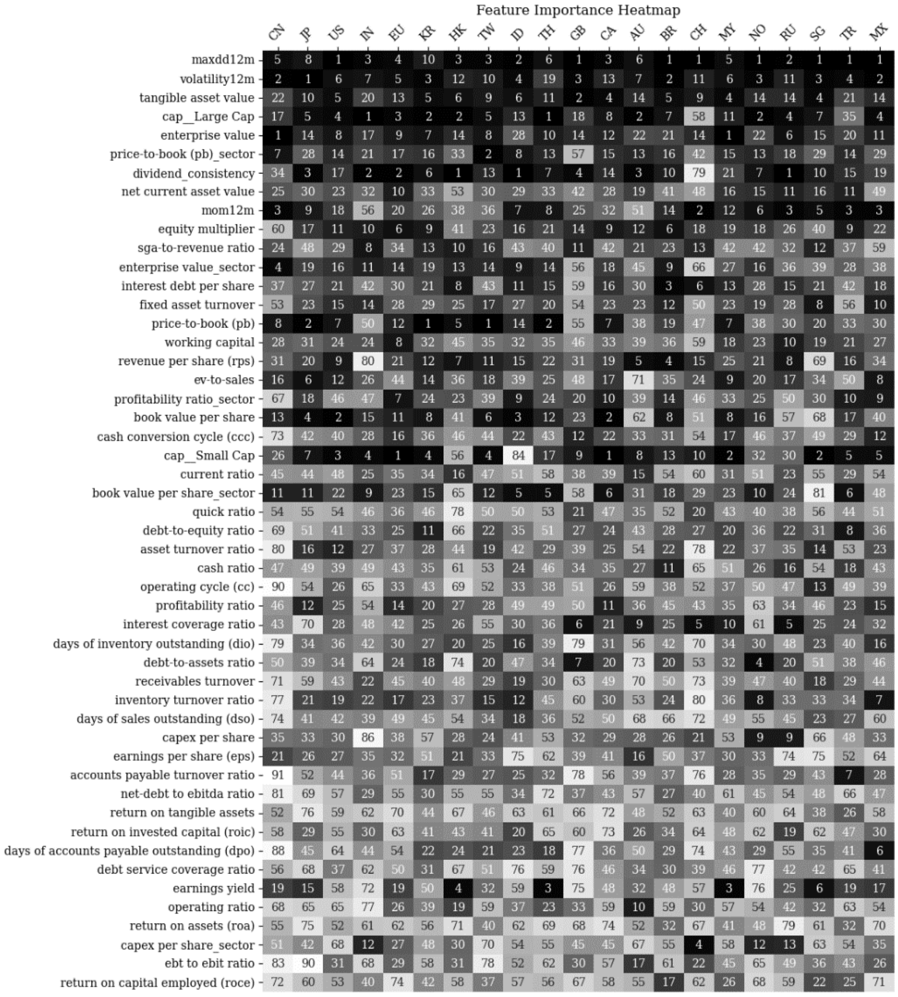
The importance ranking for each feature is color-coded so that dark colors are assigned high importance and light colors represent low importance.

Figure 5. Factor ranking by category for each country



The figure below visualizes the relative importance of different financial metrics or factors in detail for countries about predicting stock returns.

Figure 6. Factor rankings by country



	CN	JP	US	IN	EU	KR	HK	TW	ID	TH	GB	CA	AU	BR	CH	MY	NO	RU	SG	TR	MX
maxdd12m	5	8	1	3	4	10	3	3	2	6	1	3	6	1	1	5	1	2	1	1	1
volatility12m	2	1	6	7	5	3	12	10	4	19	3	13	7	2	11	6	3	11	3	4	2
tangible asset value	22	10	5	20	13	5	6	9	6	11	2	4	14	5	9	4	14	14	4	21	14
cap__Large Cap	17	5	4	1	3	2	2	5	13	1	18	8	2	7	58	11	2	4	7	35	4
enterprise value	1	14	8	17	9	7	14	8	28	10	14	12	22	21	14	1	22	6	15	20	11
price-to-book (pb)_sector	7	28	14	21	17	16	33	2	8	13	57	15	13	16	42	15	13	18	29	14	29
dividend_consistency	34	3	17	2	2	6	1	13	1	7	4	14	3	10	79	21	7	1	10	15	19
net current asset value	25	30	23	32	10	33	53	30	29	33	42	28	19	41	48	16	15	11	16	11	49
mom12m	3	9	18	56	20	26	38	36	7	8	25	32	51	14	2	12	6	3	5	3	3
equity multiplier	60	17	11	10	6	9	41	23	16	21	14	9	12	6	18	19	18	26	40	9	22
sga-to-revenue ratio	24	48	29	8	34	13	10	16	43	40	11	42	21	23	13	42	42	32	12	37	59
enterprise value_sector	4	19	16	11	14	19	13	14	9	14	56	18	45	9	66	27	16	36	39	28	38
interest debt per share	37	27	21	42	30	21	8	43	11	15	59	16	30	3	6	13	28	15	21	42	18
fixed asset turnover	53	23	15	14	28	29	25	17	27	20	54	23	23	12	50	23	19	28	8	56	10
price-to-book (pb)	8	2	7	50	12	1	5	1	14	2	55	7	38	19	47	7	38	30	20	33	30
working capital	28	31	24	24	8	32	45	35	32	35	46	33	39	36	59	18	23	10	19	21	27
revenue per share (rps)	31	20	9	80	21	12	7	11	15	22	31	19	5	4	15	25	21	8	69	16	34
ev-to-sales	16	6	12	26	44	14	36	18	39	25	48	17	71	35	24	9	20	17	34	50	8
profitability ratio_sector	67	18	46	47	7	24	23	39	9	24	20	10	39	14	46	33	25	50	30	10	9
book value per share	13	4	2	15	11	8	41	6	3	12	23	2	62	8	51	8	16	57	68	17	40
cash conversion cycle (ccc)	73	42	40	28	16	36	46	44	22	43	12	22	33	31	54	17	46	37	49	29	12
cap__Small Cap	26	7	3	4	1	4	56	4	84	17	9	1	8	13	10	2	32	30	2	5	5
current ratio	45	44	48	25	35	34	16	47	51	58	38	39	15	54	60	31	51	23	55	29	54
book value per share_sector	11	11	22	9	23	15	65	12	5	5	58	6	31	18	29	23	10	24	81	6	48
quick ratio	54	55	54	46	36	46	78	50	50	53	21	47	35	52	20	43	40	38	56	44	51
debt-to-equity ratio	69	51	41	33	25	11	66	22	35	51	27	24	43	28	27	20	36	22	31	8	36
asset turnover ratio	80	16	12	27	37	28	44	19	42	29	39	25	54	22	78	22	37	35	14	53	23
cash ratio	47	49	39	49	43	35	61	53	24	46	34	35	27	11	65	51	26	16	54	18	43
operating cycle (cc)	90	54	26	65	33	43	69	52	33	38	51	26	59	38	52	37	50	47	13	49	39
profitability ratio	46	12	25	54	14	20	27	28	49	49	50	11	36	45	43	35	63	34	46	23	15
interest coverage ratio	43	70	28	48	42	25	26	55	30	36	6	21	9	25	5	10	61	5	25	24	32
days of inventory outstanding (dio)	79	34	36	42	30	27	20	25	16	39	79	31	56	42	70	34	30	48	23	40	16
debt-to-assets ratio	50	39	34	64	24	18	74	20	47	34	7	20	73	20	53	32	4	20	51	38	46
receivables turnover	71	59	43	22	45	40	48	29	19	30	63	49	70	50	73	39	47	40	18	29	44
inventory turnover ratio	77	21	19	22	17	23	37	15	12	45	60	30	53	24	80	36	8	33	33	34	7
days of sales outstanding (dso)	74	41	42	39	49	45	54	34	18	36	52	50	68	66	72	49	55	45	23	27	60
capex per share	35	33	30	86	38	57	28	24	41	53	32	29	28	26	21	53	9	9	66	48	33
earnings per share (eps)	21	26	27	35	32	51	21	33	75	62	39	41	16	50	37	30	33	74	75	52	64
accounts payable turnover ratio	91	52	44	36	51	17	29	27	25	32	78	56	39	37	76	28	35	29	43	7	28
net-debt to ebitda ratio	81	69	57	29	55	30	55	55	34	72	37	43	57	27	40	61	45	54	48	66	47
return on tangible assets	52	76	59	62	70	44	67	46	63	61	66	72	48	52	63	40	60	64	38	26	58
return on invested capital (roic)	58	29	55	30	63	41	43	41	20	65	60	73	26	34	64	48	62	19	62	47	30
days of accounts payable outstanding (dpo)	88	45	64	44	54	22	24	21	23	18	77	36	50	29	74	43	29	55	35	41	6
debt service coverage ratio	56	68	37	62	50	31	67	51	76	59	76	46	34	30	39	46	77	42	42	65	41
earnings yield	19	15	58	72	19	50	4	32	59	3	75	48	32	48	57	3	76	25	6	19	17
operating ratio	68	65	65	77	26	39	19	59	37	23	33	59	10	59	30	57	54	42	32	63	54
return on assets (roa)	55	75	52	61	62	56	71	40	62	69	68	74	52	32	67	41	48	79	61	32	70
capex per share_sector	51	42	68	12	27	48	30	70	54	55	45	45	67	55	4	58	12	13	63	54	35
ebt to ebit ratio	83	90	31	68	29	58	31	78	52	62	30	57	17	61	22	45	65	49	36	43	26
return on capital employed (roce)	72	60	53	40	74	42	58	37	57	56	67	58	55	17	62	26	68	59	22	25	71

5.4. Machine learning portfolios

In this chapter, we will describe the construction and performance of portfolios that take advantage of the predictive power of a host of ML models. In particular, we will assess whether improved prediction capabilities of such models translate to better alpha. Construction of ML-driven portfolios begins with applying predictive models that estimate expected returns for a broad universe returns of individual stocks.. We take this comprehensive dataset to use the predicted returns as a key input for portfolio optimization. These steps are involved in the construction of these portfolios:

It ranks stocks by their expected returns and then selects top-ranked stocks to include in a portfolio subject to some criteria. The basic strategy is to pick up the top 10% quantile with the highest predicted returns. The long-only constraint restricts the portfolio to only long positions, avoiding the additional risks and complexities associated with short selling.

The portfolios are built yearly, rebalanced each July for the entire test period. This timing is based on the model's prediction of 12-month returns, allowing a portfolio to be held for an entire year before being reevaluated and rebalanced based on the latest predictions. This 12-month horizon typifies usual cycles of investments and gives sufficient time for the predictive models to pick up underlying market dynamics and trends. A view into the performance of ML-driven portfolios is gauged against traditional metrics such as cumulative returns, Sharpe ratios, and maximum drawdowns. The below table are the cumulative returns of long-only portfolios across different countries, indicating the relative efficacy of the linear models in matching ML approaches in portfolio construction.

Table 6. Cumulative returns for portfolios by country

Country	Linear Model Portfolio	Machine Learning Portfolio	Cumulative Return Gain
CN	0,784	1,767	0,983
EU	3,941	9,032	5,091
IN	2,314	5,595	3,281
JP	1,812	5,652	3,840
KR	0,554	3,330	2,776
MX	1,044	3,477	2,433
SG	0,648	2,772	2,124
TH	0,735	3,200	2,465
TR	1,862	8,518	6,656
TW	2,010	3,675	1,665
US	3,513	9,947	6,434

The empirical study results indicate that the Sharpe ratio generally increases with model complexity. The Sharpe ratios of linear models are not robust and vary between developed and emerging countries. In contrast, the gain in Sharpe ratios with ML models is significant in most countries.

Table 7. Sharpe ratios for portfolios by country

Country	Linear Model Portfolio	Machine Learning Portfolio	Sharpe Ratio Gain
CN	0,561	0,844	0,283
EU	1,441	1,889	0,448
IN	1,014	1,501	0,487
JP	0,938	1,630	0,692
KR	0,439	1,098	0,659
MX	0,773	1,002	0,229
SG	0,532	1,125	0,593
TH	0,591	1,172	0,581
TR	0,716	1,176	0,460
TW	1,031	1,194	0,163
US	1,407	1,815	0,408

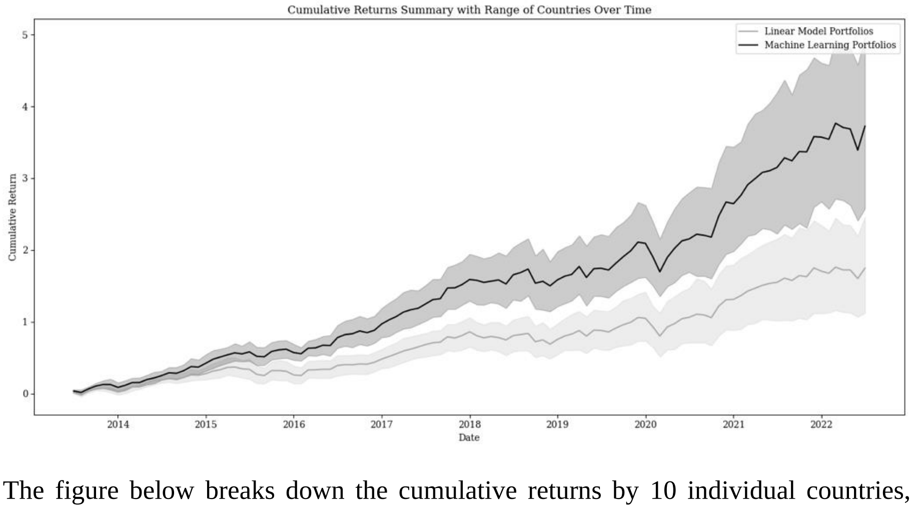
The empirical study's max drawdown results below highlight the worst peak-to-trough declines for Linear and ML Portfolios across various countries. The data shows that maximum drawdowns differ by country, with the ML Portfolio generally reducing max drawdowns compared to the Linear Model.

Table 8. Max drawdowns for portfolios by country

Country	Linear Model Portfolio	Machine Learning Portfolio	Max DD Gain
CN	(0,772)	(0,774)	(0,002)
EU	(0,201)	(0,232)	(0,031)
IN	(0,835)	(0,535)	0,300
JP	(1,892)	(0,722)	1,170
KR	(1,375)	(1,110)	0,265
MX	(2,193)	(0,485)	1,708
SG	(1,650)	(1,338)	0,312
TH	(3,666)	(1,965)	1,701
TR	(4,205)	(2,436)	1,769
TW	(0,426)	(0,285)	0,141
US	(0,327)	(0,321)	0,006

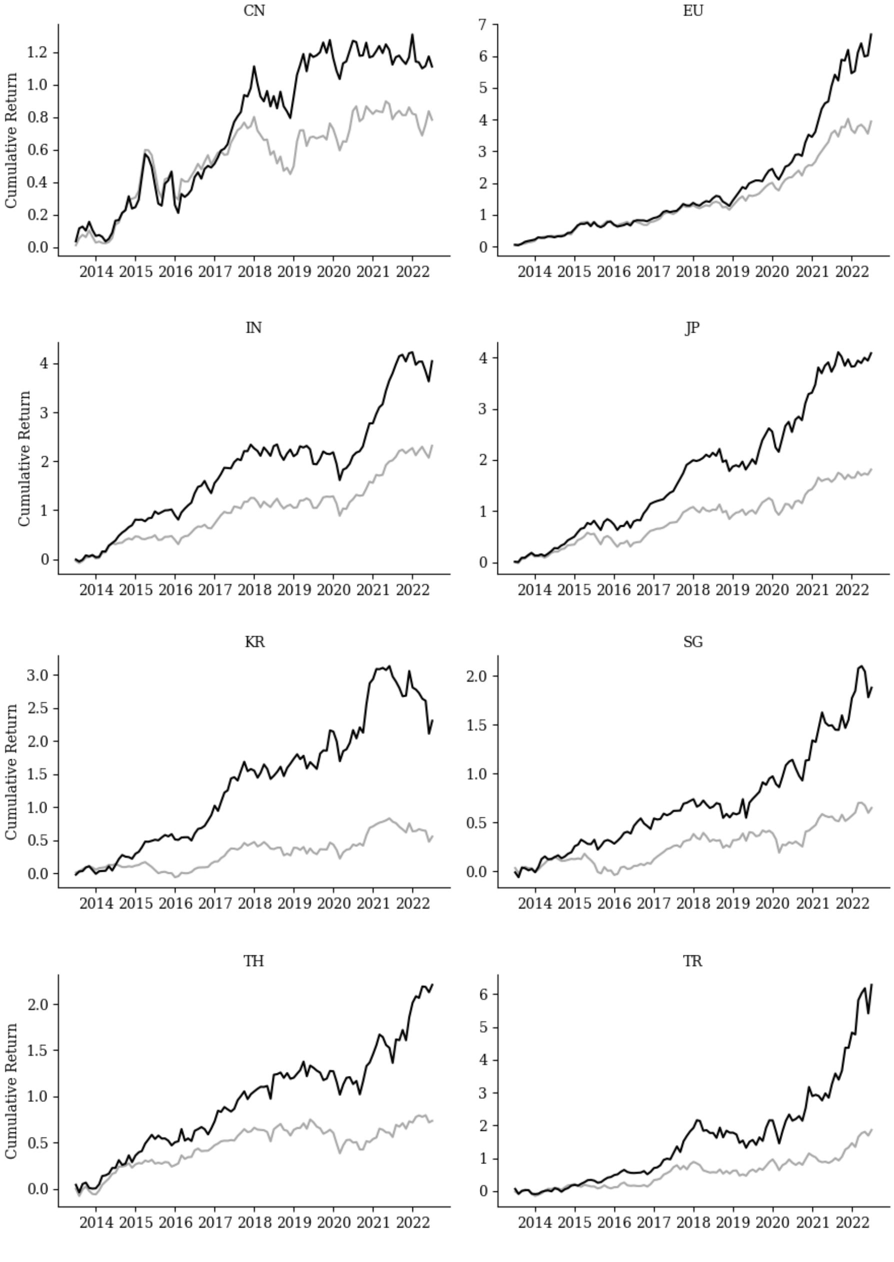
The first chart shows the cumulative returns for two portfolios likely the Linear and the ML Portfolio along with the average line and range of outcomes across different countries. The ML portfolio (indicated by the darker line) generally shows higher cumulative returns over time compared to the linear model portfolio. The shaded areas represent the variability or range of returns across different countries, indicating that while the ML Portfolio tends to outperform, there is significant variation in outcomes depending on the specific market.

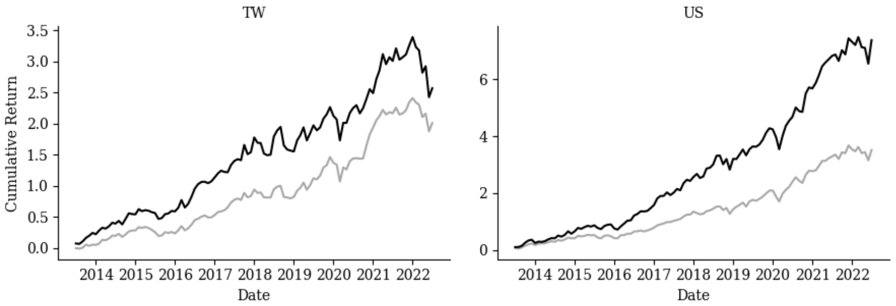
Figure 7. Cumulative returns of portfolios summary with minimum and maximum ranges of countries (2013 to 2023)



The figure below breaks down the cumulative returns by 10 individual countries, providing a more granular view. In all countries, the ML Portfolio outperforms the linear model portfolio, often with a significant margin. However, the degree of outperformance varies by country, highlighting that while ML generally offers better returns, its relative advantage can differ depending on local market conditions.

Figure 8. Cumulative returns of portfolios from 2013 to 2023 by country





6. Conclusion

Using 63 stock characteristics in 31 countries, we conclude that this is an encouraging result: all models provide positive returns and underline the potential of ML for portfolio optimization. However, it is important to optimize hyperparameters in the next level of performance. Our analysis indicates that machine learning, especially neural networks and regression trees, enhances our understanding of asset prices by picking up a nonlinearity in the interactions, which more traditional models tend not to explain. The most crucial stock characteristics feeding the models belong to the traditionally popular factors such as value, size, momentum, and reversal. However, the models cannot be limited to a few parameters as the contribution of factors is not uniform. In developed markets, quality, value, and profitability factors rank highest. Value remains significant, although its prominence has declined in these regions. Profitability is more important than value in emerging markets, likely due to inefficiencies in pricing mechanisms. Investment and dividend factors have the lowest importance for return predictability. It generally turns out that shallower models are superior to deeper ones, possibly as a result of the low signal-to-noise ratio, which makes it difficult to distinguish meaningful patterns from random noise. The second concern is that deep learning is computationally intensive, requiring careful hyperparameter tuning due to the sensitivity to hyperparameters when it is implemented in the general framework of machine learning. Not to mention, the old parsimony principle—use as few parameters as possible to avoid overfitting—is in contrast with the parameterized nature of machine learning. The evidence from our study indicates that such machine learning models can be developed into effective portfolios. It is only for increasing complexity that complex models hold the potential to greatly improve predictions and Sharpe ratio, and thus complexity is positively adopted for portfolio results. The fact that the algorithms have been successful return predictors indicates a place of increasing importance in the fintech industry and its modern financial systems. The gains in Sharpe ratios coming with more complex models, therefore, underscore the benefits of model complexity, directly challenging the traditional simplicity paradigm in financial modeling.

References

Asness, C. S., Moskowitz, T. J., & Pedersen, L. H. (2013). Value and Momentum Everywhere. *The Journal of Finance*, *68*(3), 929–985.

Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying stock market forecasting techniques – Part II: Soft computing methods. *Expert Systems with Applications*, *36*(3), 5932–5941. https://doi.org/10.1016/j.eswa.2008.07.006

Bali, T. G., Goyal, A., Huang, D., Jiang, F., & Wen, Q. (2020). The Cross-Sectional Pricing of Corporate Bonds Using Big Data and Machine Learning. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3686164

Bryzgalova, S., Pelger, M., & Zhu, J. (2020). Forest Through the Trees: Building Cross-Sections of Stock Returns. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3493458

Diebold, F., & Mariano, R. (1995). Comparing Predictive Accuracy. *Journal of Business &amp; Economic Statistics*, *13*(3), 253–263.

Dillschneider, Y. (2022). *A Machine Learning Framework for Asset Pricing* (SSRN Scholarly Paper 4097100). https://doi.org/10.2139/ssrn.4097100

Drobetz, W., & Otto, T. (2020). Empirical Asset Pricing via Machine Learning: Evidence from the European Stock Market. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3640631

Fama, E. F., & French, K. R. (1992). The Cross-Section of Expected Stock Returns. *The Journal of Finance*, *47*(2), 427–465. https://doi.org/10.1111/j.1540-6261.1992.tb04398.x

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3–56. https://doi.org/10.1016/0304-405X(93)90023-5

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, *116*(1), 1–22. https://doi.org/10.1016/j.jfineco.2014.10.010

Fama, E. F., & French, K. R. (2017). International tests of a five-factor asset pricing model. *Journal of Financial Economics*, *123*(3), 441–463. https://doi.org/10.1016/j.jfineco.2016.11.004

Freund, Y. (1995). Boosting a Weak Learning Algorithm by Majority. *Information and Computation*, *121*(2), 256–285. https://doi.org/10.1006/inco.1995.1136

Giglio, S., & Xiu, D. (2021). Asset Pricing with Omitted Factors. *Journal of Political Economy*, *129*(7), 1947–1990. https://doi.org/10.1086/714090

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *Review of Financial Studies*, *33*(5). https://doi.org/10.1093/rfs/hhaa009

Harvey, C. R., & Liu, Y. (2021). Lucky factors. *Journal of Financial Economics*, *141*(2), 413–435. https://doi.org/10.1016/j.jfineco.2021.04.014

Heaton, J. B., Polson, N. G., & Witte, J. H. (2016). *Deep Learning in Finance* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1602.06561

Jerome H. Friedman. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Kelly, B. T., & Malamud, S. (2021). The Virtue of Complexity in Machine Learning Portfolios. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3984925

Kelly, B. T., & Xiu, D. (2023). Financial Machine Learning. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4501707

Kozak, S. (2019). Kernel Trick for the Cross Section. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3307895

Leippold, M., Wang, Q., & Zhou, W. (2022). Machine learning in the Chinese stock market. *Journal of Financial Economics*, *145*(2). https://doi.org/10.1016/j.jfineco.2021.08.017

Moritz, B., & Zimmermann, T. (2016). Tree-Based Conditional Portfolio Sorts: The Relation between Past and Future Stock Returns. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2740751

Rasekhschaffe, K. C., & Jones, R. C. (2019). Machine Learning for Stock Selection. *Financial Analysts Journal*, *75*(3). https://doi.org/10.1080/0015198X.2019.1596678

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227. https://doi.org/10.1007/BF00116037

Subrahmanyam, A. (2010). The Cross-Section of Expected Stock Returns: What Have We Learnt from the Past Twenty-Five Years of Research? *European Financial Management*, *16*(1), 27–42. https://doi.org/10.1111/j.1468-036X.2009.00520.x

Swade, A., Hanauer, M. X., Lohre, H., & Blitz, D. (2023). Factor Zoo. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4605976

Tobek, O., & Hronec, M. (2021). Does it pay to follow anomalies research? Machine learning approach with international evidence. *Journal of Financial Markets*, *56*, 100588. https://doi.org/10.1016/j.finmar.2020.100588

Tukey, J. W. (1961). Box and Jenkins. In *Time Series Analysis: Forecasting and Control*.