JEL Classification: G11, G12, G15, C55
Keywords: Machine learning, asset pricing, horizon predictability, anomalies

Multi-Horizon Equity Returns Predictability via Machine Learning

Lenka NECHVÁTALOVÁ - Institute of Economic Studies, Faculty of Social Sciences, Charles University & Institute of Information Theory and Automation, Czech Academy of Sciences (lenka.nechvatalova@fsv.cuni.cz)

Abstract

We investigate the predictability of global expected stock returns across various forecasting horizons using machine learning techniques. We find that the predictability of returns decreases with longer forecasting horizons both in the U.S. and internationally. Despite this, we provide evidence that using firm-specific characteristics can remain profitable even after accounting for transaction costs, especially when we consider longer forecasting horizons. Studying the profitability of long-short portfolios, we highlight a trade-off between higher transaction costs connected to frequent rebalancing and greater returns on shorter horizons. Increasing the forecasting horizon while matching the rebalancing period increases risk-adjusted returns after transaction costs for the U.S. We combine predictions of expected returns at multiple horizons using double-sorting and a turnover reducing strategy, buy/hold spread. Double sorting on different horizons significantly increases profitability in the U.S. market, while buy/hold spread portfolios exhibit better risk-adjusted profitability.

1. Introduction

Asset pricing literature introduces an array of potential predictor variables, commonly referred to as anomalies. Forecasting cross-sectional stock returns using machine learning has proven successful in comparison to more traditional methods. So far, the focus has been mainly on the one-month prediction horizon.

This study explores the impact of longer forecasting horizons on the predictability of cross-sectional equity returns. We show a diminishing predictability of cross-sectional stock returns with longer forecasting horizons, both in the U.S. and globally. Shorter horizon portfolios provide greater profitability but come with higher transaction costs stemming from frequent rebalancing and high turnover. After accounting for transaction costs, the risk-adjusted profitability of long-short decile portfolios increases with longer horizons in the U.S. Combining shorter and longer horizon forecasts using double-sorting significantly improves profitability in the U.S. Further, with the use of turnover-reducing strategy and combination of forecasts, we get better risk-adjusted returns in the U.S. International portfolios offer higher risk-adjusted returns compared to U.S.-only strategies. Global portfolios using longer horizons forecasts have similar or lower risk-adjusted returns compared to the one-month portfolio.

https://doi.org/10.32065/CJEF.2024.02.01

The financial support was provided by the Grant Agency of Charles University (grant number 316521) and the Czech Science Foundation under the 19-28231X (EXPRO) project. I would also like to thank the anonymous reviewers for their valuable comments and suggestions.

Recent research has demonstrated the effectiveness of combining anomalies[2] through machine learning-based predictive regression in achieving unprecedented out-of-sample expected returns predictability, see Gu et al. (2020), Giglio and Xiu (2019), Kelly et al. (2019), Kozak et al. (2020), Chen et al. (2020), Bryzgalova et al. (2020) and Freyberger et al. (2017). This superior predictability is not a consequence of adding more predictive variables than previous literature but of allowing nonlinear interaction of predictive variables and incorporating regularization. Avramov et al. (2023) study the economic viability of these strategies and document that machine learning portfolios extract profitability from microcap and difficult-to-arbitrage stocks, and profitability is attenuated after accounting for transaction costs due to frequent rebalancing and high turnover.

However, most of the empirical results in the asset pricing literature, including the recent work using machine learning, are focused on a one-month forecasting horizon. Investors are not horizon agnostic, though. Only an investor with a logarithmic utility function would allocate his portfolio the same way for single and multiple horizons. Since stock returns are not independent and identically distributed, an investor could use time and cross-sectional variation in expected stock returns to his advantage.

We examine the predictability of expected stock returns across multiple horizons using machine learning. 153 anomalies from Tobek and Hronec (2021) are used as variables in predictive regressions of expected cumulative returns from one month to two years ahead. We find decreasing predictability with longer horizons for both the U.S. and internationally. This is consistent with Baba Yara (2020), who proposes an economically restricted machine learning model and documents the decreasing predictability of cross-sectional returns on longer horizons in the U.S. On the other hand, it differs from the conclusion of Gu et al. (2020), who document higher predictability on a one-year horizon compared to a shorter, one-month horizon. This could be a result of different information sets being used, e.g. inclusion of macroeconomic variables and different universe definitions.

So far, literature looking at longer forecasting horizons of cross-sectional stock returns is mainly limited to the U.S. stock market. Blitz et al. (2023) and Cakici et al. (2023) compare different machine learning approaches and look at the profitability of portfolios. Blitz et al. (2023) show that one-month portfolios can be outperformed by longer horizons portfolios by using a turnover-reducing portfolio construction. Cakici et al. (2023) document decreasing profitability with a longer forecasting horizon, larger firm size and in more recent times. Leung et al. (2021) use stochastic gradient trees to contrast one and six-month forecasts, concluding that profitability depends on the ability of investors to execute trades efficiently. These studies do not incorporate transaction costs into their portfolios but rather use fixed costs or calculate break-even transaction costs.

[2] This predictability is typically based on individual firm characteristics, and it is common to use the terms anomaly and firm characteristic interchangeably. Examples of firm characteristics used as stock return predictors are momentum (Jegadeesh and Titman, 1993), accruals (Sloan, 1996), size and book-to-market ratio (Fama and French, 1992a). For the comprehensive list of anomalies documented in the literature, please see the large replication study of Hou et al. (2020).

Avramov et al. (2023) study the economic viability of machine learning strategies and document that they extract profitability from the microcap and difficult-to-arbitrage stocks, and profitability is attenuated after accounting for transaction costs due to the high turnover of these portfolios. Other papers examining transaction costs when combining multiple anomalies are, for example, Frazzini et al. (2012) and DeMiguel et al. (2020).

We study how the performance of portfolios using longer horizon forecasts compares to one one-month portfolio benchmark and provide evidence from the U.S. market and internationally. We construct multiple portfolios with varying turnover levels, enabling us to use either a single horizon forecast or a combination of forecasts from different horizons. We look at portfolios' performance after accounting for transaction costs. Using estimated transaction costs allows us to better compare different portfolios and allows us to get a more accurate picture as transaction costs decrease over time. Gu et al. (2020) and Tobek and Hronec (2021) focus on the monthly horizon and document strong predictability in the cross-section of returns using several machine learning methods. We replicate their results on a highly liquid universe of stocks and use it as our benchmark.

First, we construct long-short decile portfolios, keeping the rebalancing frequency equal to the forecasting horizon. There is decreasing profitability with longer horizons. However, after accounting for transaction costs, the risk-adjusted returns are higher for longer-horizon portfolios in the U.S. We also investigate the performance of portfolios over time. There is a decrease in profitability for all horizons, but shorter horizons are more affected. After 2005, using longer horizon forecasts provides better Sharpe ratios compared to the benchmark both in the U.S. and internationally.

Second, we combine predictions for two different horizons via double sorting. We independently sort stocks based on predicted cumulative returns from two different horizons. Each month, we buy stocks in the top 15% for both horizons and sell stocks in the bottom 15% for both horizons. In the U.S., this leads to large performance gains over our benchmark. Internationally, this also leads to higher risk-adjusted returns; however, the difference is not that stark.

Further, we employ a buy/hold spread strategy, proposed by Novy-Marx and Velikov (2019). It is a turnover-reducing strategy where the hurdle is higher to buy into a position than to hold a position once it is in a portfolio. We adjust it to allow us to combine two forecasting horizons. Each month, we buy (sell) firms in the top (bottom) 10% of longer-horizon forecast. However, firms in the portfolio from the previous months will be sold (bought) only if they are not in the top (bottom) 20% for the one-month forecast. Buy/hold spread portfolios in the U.S. have better risk-adjusted profitability. This holds for the international sample as well; however, the difference is mild.

The rest of this paper is organized as follows: Section 2 describes the data and methodology used in our analysis. Section 3 contains multi-horizon prediction results and decile, double sorted, and buy/hold spread long-short portfolios performances, with evidence from the U.S. and international datasets. Section 4 summarizes and concludes our work.

2. Data and Methodology

2.1 Data

Our analysis is done firstly on the U.S. and then on the international dataset. The international dataset contains stocks from the U.S. and 22 other developed countries. For the United States equity data, we use the CRSP/Compustat Merged Database from the Center for Research in Security Prices. For international equity data, we use Datastream from Refinitiv. Additionally, I/B/E/S Estimates are used to calculate several anomalies. We use the U.S. consumer price index to estimate transaction costs and the monthly U.S. T-bill rate, which will be used for anomaly calculation, both from Datastream. We also use the Market minus risk-free rate for the U.S. and developed markets from the data library provided by French (2020).

Table 1 Descriptive Statistics

	U.S.			***International (excl. U.S.)***		
	r	***MC***	***Number of firms***	***r***	***MC***	***Number of firms***
Mean	0.94	6107.76	1100.11	0.42	6414.29	1871.55
Std	11.38	22937.00	249.45	11.58	16081.87	281.34
25%	-4.81	348.35	947.00	-5.51	828.54	1661.00
50%	0.75	1156.27	1042.50	0.16	1936.31	1911.50
75%	6.47	3763.79	1250.50	5.96	5205.55	2061.75

Notes: Column *r* corresponds to monthly returns and is in percentages, *MC* stands for market capitalization (in millions of dollars), and the number of firms in the cross-section each month are reported for the U.S. and international sample (with the U.S. excluded). The period from 1963 to 2018 for the U.S. and 1980 to 2018 internationally is covered.

The dataset is filtered and preprocessed, including a strict liquidity filter to exclude thinly traded stocks. Details on this process are provided in Appendix C. Table 1 presents descriptive statistics for preprocessed and filtered universe. Summary statistics for monthly returns, market capitalization, and the number of firms at the end of the month are presented separately for U.S. and international (excluding U.S.) datasets. The average monthly return in the U.S. is two times higher than in the international sample. At the end of the month, the U.S. dataset has 1100 firms on average, and the international dataset (with the U.S. excluded) has 1870 firms on average.

We calculate 153 anomalies that were published in the academic literature. We follow the list of anomalies and their construction from Tobek and Hronec (2021). These anomalies fall into three main categories: 93 fundamental anomalies, which can be further classified into accruals, intangibles, profitability, and value factors[3]. Additionally, there are 49 market friction anomalies[4] and 11 I/B/E/S

[3] Fundamental anomalies include accruals (Sloan, 1996), asset liquidity (Ortiz-Molina and Phillips, 2014), investment (Titman et al., 2004), leverage (Bhandari, 1988) and assets-to-market (Fama and French, 1992b).

[4] Market friction anomalies include, for example, seasonality (Heston and Sadka, 2008), short-term reversal (Jegadeesh, 1990), industry momentum (Moskowitz and Grinblatt, 1999) and momentum (Jegadeesh and Titman, 1993).

anomalies[5]. All anomalies are calculated at the firm-specific level on a monthly basis. Fundamental data are available at a yearly frequency, but the anomalies are updated at a monthly frequency using financial statement data from financial years ending at least 6 months prior. This approach accommodates varying fiscal year-end dates and prevents the use of outdated information. Missing observations in anomalies are imputed with a cross-sectional median for the firm's region. Further, to address outliers and stabilize training, all anomalies are normalized based on cross-sectional quantiles within the firm's region, following standard procedures used e.g. in Gu et al. (2020) or Kozak et al. (2020).

Transaction costs are estimated at a monthly frequency using closing quoted spread proxy (Chung and Zhang, 2014) and volatility over volume proxy (Fong et al., 2018). More details on the estimation of transaction costs can be found in subsection B.3.

2.2 Expected Stock Returns

Following the approach of Lewellen (2015) or Gu et al. (2020), we use predictive regressions for excess stock returns, which are concerned with the conditional mean estimation:

$$R_{i,t+1} = \mathrm{E}_t\left(R_{i,t+1}\right) + \epsilon_{i,t+1}$$
$$\mathrm{E}_t\left(R_{i,t+1}\right) = g\left(Z_{i,t}\right)$$

where stocks are indexed as $i = 1, \dots, N$, months by $t = 1, \dots, T$, $Z_{i,t}$ are stock characteristics or predictive signals and g is a general function of these predictive signals estimated to optimize the out-of-sample predictability of $\mathrm{E}_t\left(R_{i,t+1}\right)$.

This setting encompasses the setup of Lewellen (2015), who uses Fama-MacBeth regressions, and therefore, the functional form of g is a simple linear combination of ordinary least squares. Directly addressing shortcomings of the ordinary least squares approach, Gu et al. (2020) use a variety of machine learning methods, such as Elastic net, Random Forests, Gradient Boosted Trees and Neural Networks, to represent the function g. Machine learning methods aim to explicitly allow non-linearity, the interaction of predictive variables, and regularization. The goal is to aggregate all of the available input and anomalies and condense them into one real-valued output.

We extend the forecasting horizon from one month to multiple horizons. We also explicitly allow heterogeneity in predictability across horizons:

$$\mathrm{E}_t\left(R_{i,t+h}\right) = g_h\left(R_{i,t+h}\right) + \epsilon_{i,t+h}$$

where h is the forecasting horizon. We consider horizons from monthly, $h = 1$, to two years, $h = 24$.

[5] I/B/E/S anomalies studied include analyst value (Frankel and Lee, 1998), changes in analyst earnings forecasts (Hawkins et al., 1984), and long-term growth forecasts (La Porta, 1996).

2.3 Model Estimation

For the model estimation, we split the dataset into training, validation and testing sets that keep the time ordering, following Gu et al. (2020) and Tobek and Hronec (2021). In our modelling process, hyperparameter search involves systematically exploring different combinations of model parameters to optimize out-of-sample performance. This search is conducted using the training and validation samples from our dataset. The training sample is used to train the model with various hyperparameter configurations, while the validation sample is employed to evaluate and compare the performance of these configurations. By iterating through different parameter settings and assessing their impact on model outcomes using the validation set, we aim to identify the optimal set of hyperparameters that maximize the predictive accuracy and generalizability of the model. This iterative search process ensures that our model is finely tuned and robust for making accurate predictions on unseen data. The best set of hyperparameters is selected using mean square error calculated on the validation sample. We utilize the model selected through hyperparameter search to generate out-of-sample predictions by evaluating its performance on the test dataset. This step ensures that our model's predictions are validated against unseen data, offering a robust assessment of its real-world predictive capability.

For our analysis, we adopt an expanding window approach, where we sequentially train and evaluate our predictive models across multiple time periods. To illustrate, our first model uses all available data up to 1994, splitting it into a 70% training set and a 30% validation set. Subsequently, we assess the model's performance on our testing sample, data from 1995. This procedure is repeated annually from 1995 to 2018, with each iteration involving a hyperparameter search with the selected model used to generate out-of-sample predictions.

As our model, we use a feedforward neural network. As a form of robustness, we also include results using gradient-boosted trees. These methods, optimization and regularization techniques are described in subsection B.1.

In our model estimation process, we use cumulative returns at horizon h that have been cross-sectionally demeaned as our target labels. Each horizon is independently estimated, and normalized anomalies serve as the input features for our model. In case the firm is delisted during the period for which we calculate cumulative returns, we use returns that are available and disregard months when stock is delisted.

To optimize model performance, we conduct an extensive hyperparameter search. The hyperparameter search follows Tobek and Hronec (2021) and extends to cover more degrees of model complexity as it could vary across horizons. This search spans various aspects of model complexity, such as network architecture and training parameters. We explore six distinct network architectures, ranging from single-layer to three-layer configurations, with options for wide networks featuring 150 nodes per layer or narrower networks with fewer nodes in each hidden layer (e.g., 32 nodes in the first layer, 16 in the second, and 8 in the third, if present). Within our hyperparameter search, we investigate different batch sizes (256 or 1024), dropout rates (0.1, 0.01, and 0.001), and learning rates (0.1, 0.01, and 0.001) to optimize model performance. Our training process involves fixed epochs (25), Adam optimization with specific betas (0.9 and 0.999), and early stopping with a patience of 5 epochs. To enhance model robustness, we initialize and train five models with different random seeds and use an ensemble of them. Reducing the learning rate on plateau patience is applied after each epoch, with the learning rate halved if there is no improvement.

2.4 Portfolio Formation

To assess the economic significance of our forecasts, we construct multiple portfolios. As we are interested in various forecasting horizons and the effects of transaction costs we consider three different portfolio construction methods. They offer different portfolio turnover levels and allow us to combine multiple forecasts together. All portfolios are based on portfolio sorting, a frequently used method in asset pricing[6]. We also use equal-weighting, as we only focus on the universe of most liquid stocks[7].

Decile Sorting

If we have forecasts for one-month returns, we commonly see long-short decile portfolios that are rebalanced monthly. For longer forecasting horizons, a quite straightforward extension of this is prolonging the holding period b to match the forecasting horizon h.

Because we have predictions at a monthly frequency, we have to decide how to use this information for a holding period longer than one month. One approach would be to simply invest using the most recent predictions, wait for the holding period, and then rebalance. This approach discards predictions from the months when the portfolio is not rebalanced. But more importantly, we would be exposed to rebalance timing risk affecting our portfolio (Hoffstein et al., 2020).

To counter rebalance timing risk, we use multiple sub-portfolios to create overlapping portfolios with staggered rebalancing schedule (Jegadeesh and Titman, 1993). To create sub-portfolios, we divide available capital at the start of investing into b equal parts, where b equals the holding period. Each sub-portfolio will then function as a separate portfolio. The sub-portfolios will be rebalanced in a staggered manner - each month, one of the subportfolios, the one that was rebalanced b months ago, is rebalanced to reflect the current information. This way, we are able to use all the information from the forecasts.

To create a long-short decile portfolio, each month, we cross-sectionally sort stocks based on the returns predictions. We buy firms that are in the highest predicted return decile and short firms from the lowest decile for each month. This trade is executed in the corresponding sub-portfolio, and the positions are held for the holding period b of months. The final portfolio value is obtained as a sum of the values of individual sub-portfolios. Portfolio returns are obtained by weighting returns from sub-portfolios.

[6] See, a survey of Green et al. (2013).

[7] See Section C.3.

Double Sorting

A way to use two forecasts together, e.g. using different forecasting horizons or combining different models, is double sorting. In our empirical analysis, we explore the effectiveness of double sorting by combining forecasts with different horizons. Specifically, we create portfolios by pairing a one-month horizon forecast with a longer-term one. This will hopefully allow us to capture immediate market dynamics as well as broader trends. We construct long-short portfolios by buying firms that rank in the top 15% according to both short-term and long-term forecasts. Conversely, we short-sell firms ranking in the bottom 15% for both horizons. By doing so, we aim to capitalize on stocks exhibiting consistent performance across different timeframes while mitigating risks associated with those demonstrating poor performance. Portfolios are rebalanced each month, avoiding the use of stale signals.

Buy/Hold Spread Strategy

Buy/hold spread, also referred to as banding, is a transaction cost mitigation technique applied during portfolio selection. The strategy's aim is to reduce turnover. It works by having a stricter rule to trade into position than to trade out of it. For example, the 10%/20% strategy means that we buy stocks that belong to the top 10% of the stocks and hold them as long as they are in the top 20%. At the same time, we sell the lowest 10% of the stocks and hold them until they are no longer in the bottom 20%.

While Novy-Marx and Velikov (2019) use this technique on only a single forecast or characteristic, we extend this to combine two different forecasts. We use forecasts from two models with different predicting horizons and use the one with the longer horizon as a buy signal and the shorter, one-month forecast, as a hold signal. The reasoning behind this is that we will buy (sell) a longer-term position, and then each month, we check whether the new, additional information from the shorter horizon supports holding the position or not. With our approach, we do not have consistency between buy and hold signals. For a single forecast buy/hold spread portfolio, it always holds that if a firm is in the buy category, then it is also in the hold category at time t. However, with two different forecasts, conflicting recommendations can occur. For instance, the longer-horizon buy signal might suggest buying, while the shorter-horizon hold signal recommends continuing to short the firm. To address these conflicts, we introduce a rule to remove firms from the portfolio when the buy and hold signals suggest opposing actions (the buy signal would buy while the hold signal would sell and vice versa).

In practice, for a firm to be included in our portfolio, it must meet specific criteria based on both the longer forecast horizon and the one-month horizon predictions. We enter a long (short) position in a firm if it ranks within the top (bottom) 10% according to the longer forecast horizon and is above the 20th (below 80th) percentile for the one-month horizon predictions.

Once a firm is in our portfolio, we check each month whether to keep the long (short) position or remove the firm from our portfolio. The long (short) position is kept if it is again recommended by the longer-horizon forecast (buy condition above) or if the firm is in the top (bottom) 20% for the one-month forecast and above the 10th (below 90th) percentile for the longer-horizon forecast.

In summary, the buy/hold spread technique, when extended to incorporate multiple horizon forecasts, offers a systematic approach to portfolio management, balancing long-term positioning with short-term adjustments based on forecasted signals. It also contributes to reducing turnover and, consequently, transaction costs.

Returns Calculation

Independent of the type of portfolio we are constructing, we need to calculate returns from our portfolio (or each sub-portfolio in case of a longer holding period) while accurately accounting for transaction costs. This is calculated iteratively, as trading needs to reflect the current weights of the portfolio. At a given month, we have target actions for each firm - buy, sell, hold, nothing/remove from the portfolio. Based on these actions, we assign target weights w_{it}^{*} to each firm each month. When we rebalance the portfolio, we divide available capital using equal weighting between the firms we intend to buy (sell). For decile sorting and double sorting, we fully reflect the current target actions. In a buy/hold spread portfolio, some firms are kept in a portfolio, and the rest of the capital is divided between firms we aim to buy or sell. For capital of one unit, we aim to have long positions sum to one and short sum to minus one.

When transaction costs are present, we need to account for that so that we do not overbuy and maintain our total weights within limits. The actual weight that is bought is:

$$\begin{aligned} w_{it} &= w_{it}^{*} - ts_{it} \cdot tc_{it} \\ ts_{it} &= w_{it} - w_{i(t-1)}^{end,norm} \end{aligned}$$

where ts_{it} is trade size, tc_{it} are transaction costs, and $w_{i(t-1)}^{end,norm}$ is the normalized weight at the end of the previous month for firm i.

Weight of a firm at the end of a month is $w_{it}^{end} = w_{it} \cdot (1 + R_{it})$, in case we hold a position, and zero otherwise. The normalized weight is calculated as

$$w_{it}^{norm} = \frac{2w_{it}}{\sum_i |w_{it}|}$$

In case we remove a given stock from our portfolio during month t, we will reflect the trading costs incurred in the returns of that month. The return from holding a firm i during the month t is calculated as $w_{it}R_{it}$.

For performance evaluation of portfolios, we use several metrics, with their definitions in subsection B.2. Monthly mean and standard deviation, annualized Sharpe ratio and maximum drawdown are presented in the main text. Additionally, the Sortino ratio, conditional value at risk at 99%, Alpha and Beta for our portfolios are reported in Appendix A. The selection of these metrics was based on related literature for easier comparison.

3. Empirical Results

We obtain predictions of cumulative returns at multiple horizons using feedforward neural networks, separately for the U.S. and the international dataset (with the U.S. included). The out-of-sample forecasts are from 1995 to 2018, 277 months in total. We investigate the predictive ability of those forecasts at different horizons. We then provide results of portfolios constructed from multi-horizon returns forecasts in the U.S. and internationally. As a robustness check, gradient-boosted regression trees were also used to obtain the forecasts, with results presented in subsection D.2.

The U.S. has approximately 316,000 firm-month out-of-sample forecasts, averaging 1172 firms each month. Internationally, we have 817,000 observations with an average of 3,000 firms monthly. The models trained on international data also have more training data, which could help with learning. We follow the approach of Lewellen (2015), who assess the predictive ability of forecasts using regression of realized returns on these out-of-sample predictions. Table 2 shows the predictive ability of the forecasts at different horizons. The t-statistics are calculated using Newey-West correction with $h + 4$ lags as a way to account for the overlap in regressions. The predictive slope is from regressing demeaned cumulative returns on predictions that were made for the corresponding horizon. The slopes are positive, significant for most horizons, and decrease with longer horizons. This implies that the predictions contain too much variation, and we would need to shrink the predictions to obtain a more precise estimate of the expected return. R^2 decreases with longer horizons for both the U.S. and the international sample, suggesting that the predictability decreases with longer horizons. R^2 is higher internationally, which could be due to a larger training sample.

Table 2 Predictive Ability of Return Forecasts

	U.S.			*International*		
	Slope	***t-stat***	***R^2***	***Slope***	***t-stat***	***R^2***
1	0.460	24.414	0.292	0.507	37.253	0.351
2	0.258	7.138	0.239	0.334	20.803	0.329
3	0.117	2.021	0.121	0.157	3.085	0.177
4	0.026	1.649	0.038	0.064	3.254	0.071
5	0.047	2.707	0.052	0.016	2.107	0.024
6	0.005	1.221	0.004	0.017	3.015	0.025
9	0.007	2.009	0.009	0.022	1.903	0.037
12	0.003	1.005	0.003	0.001	1.079	0.003
24	0.003	1.873	0.004	0.003	4.300	0.007

Notes: The table reports the predictive ability of return forecasts at various horizons. The slope, t-statistics and R^2 for horizon *h* are from a regression of the demeaned cumulative return on return prediction at the corresponding horizon. Results are for the period between 1995 and 2018 and are either for U.S. or international sample. Newey-West correction with *h* + 4 lags is applied. R^2 is reported in percentages.

The decreasing predictability with longer horizons matches the conclusion of Baba Yara (2020). Gu et al. (2020), have the opposite conclusion, they report higher $R^2$for yearly predictions than for monthly ones. This could be because of different anomalies being used or differing datasets. We have a more liquid universe of stocks and do not include macroeconomic predictors.

3.1 Evidence from the United States

Decile Portfolios

We construct long-short decile portfolios. For simplicity, when presenting the results, we keep the holding period *b* equal to the forecasting horizon *h*. We use the staggered rebalancing schedule with *h* sub-portfolios to fully utilize the monthly-frequency predictions. This way, the turnover and transaction costs are lowered significantly for longer horizons. There are, on average, 235 firms in a sub-portfolio.

Table 3 presents the mean, standard deviation, annualized Sharpe ratio and maximum drawdown both for portfolios without and with transaction costs included. For comparison, if we invested in buy-and-hold of S&P500 for the same period, we would have mean monthly returns of 0.64% with a Sharpe ratio of 0.53. In our analysis, a one-month long-short decile portfolio will serve as our benchmark. The results for our one-month decile portfolio without transaction costs are consistent with those of Gu et al. (2020) who report similar means and standard deviations. Their models include macroeconomic variables and interactions between firm characteristics and factors as opposed to our model, where we only include firm-specific characteristics. Tobek and Hronec (2021) also report comparable results on the U.S. sample, albeit with slightly lower means and Sharpe ratios. This could be due to the fact that they include anomalies only after the publication date. Looking at longer horizons, our results are consistent with Cakici et al. (2023), who also find decreasing returns at longer horizons. However, as they use quintile sorts with only a single sub-portfolio, their results at longer horizons are more risky and not directly comparable.

Portfolios that were formed using longer horizon predictions have lower mean returns. This is more pronounced in a case without transaction costs, as when transaction costs are included, longer horizons are less costly to trade. However, after accounting for transaction costs, the Sharpe ratios are increasing with longer horizons, thus offering better risk-adjusted returns than the one-month portfolio.

Long-only component of strategies has a higher mean return but also a higher variance and deeper maximum drawdowns compared to the long-short strategy. In case we had short-selling restrictions, long-only portfolios at longer horizons offer similar risk-adjusted returns to our benchmark. The short component of portfolios is not profitable on its own, with negative mean returns at all horizons after accounting for transaction costs; however, it serves as a hedge during more turbulent periods.

The turnover of the one-month strategy is 120%. This means that we sell (buy) roughly 60% of firms from both the long and the short side of our portfolio and buy (sell) different firms when rebalancing. Longer horizons have lower turnover by construction, and it is approximately *h* times smaller than the turnover of the one-month portfolio.

Additional performance measures, Sortino ratio, conditional value at risk at 99%, Alpha and Beta can be seen in Table A.1. Alpha and Beta are calculated with respect to U.S. market returns. In Figure A.1 can be seen cumulative returns for decile portfolios. There is a drop in profitability after approximately 2005. We discuss this decrease in profitability and how it affects different horizon portfolios in subsection D.1.

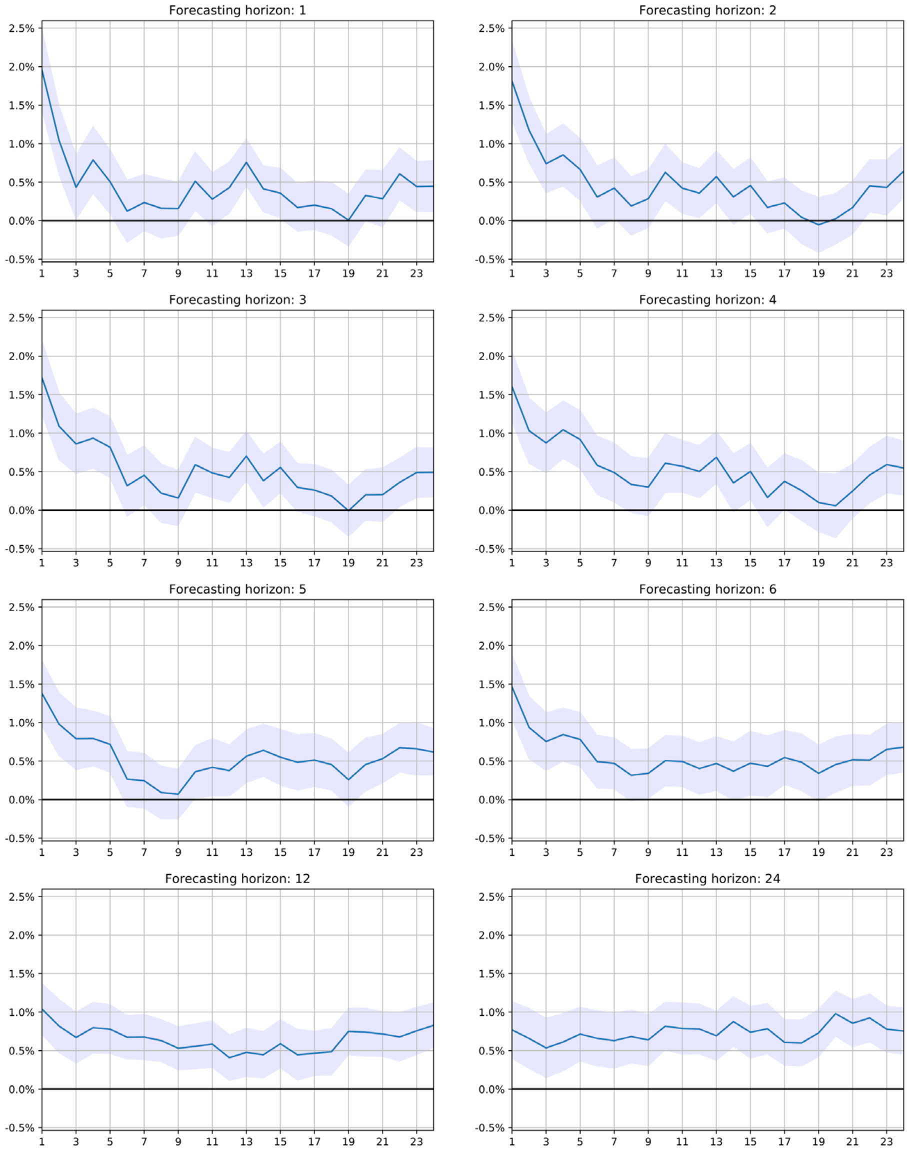
Table 3 Performance of Long-Short Decile Portfolios in the U.S.

	Without transaction costs				*With transaction costs*				
	Mean	***Std***	***Sharpe***	***MDD***	***Mean***	***Std***	***Sharpe***	***MDD***	***Turnover***
				Panel A: Long-short portfolio					
1	1.76	5.23	1.16	-30.27	1.13	5.14	0.76	-37.31	120.20
2	1.33	4.23	1.09	-32.89	1.03	4.19	0.85	-36.76	58.43
3	1.11	3.74	1.03	-25.69	0.91	3.73	0.84	-27.31	40.36
4	1.06	3.37	1.09	-17.71	0.90	3.37	0.93	-20.15	32.03
5	0.88	3.09	0.99	-39.00	0.75	3.10	0.84	-41.96	26.68
6	0.89	2.87	1.07	-26.09	0.77	2.87	0.94	-29.24	23.02
9	0.76	2.44	1.08	-27.16	0.69	2.44	0.98	-29.15	16.15
12	0.73	2.18	1.15	-24.75	0.67	2.19	1.06	-26.16	12.78
24	0.82	2.46	1.15	-24.87	0.79	2.46	1.11	-25.67	6.73
				Panel B: Long only component of the strategy					
1	1.74	7.16	0.84	-53.23	1.42	7.10	0.69	-58.81	126.81
2	1.44	7.22	0.69	-65.59	1.29	7.21	0.62	-67.86	59.86
3	1.28	7.07	0.63	-67.06	1.18	7.07	0.58	-68.67	40.60
4	1.32	7.04	0.65	-63.21	1.24	7.04	0.61	-64.75	31.99
5	1.20	7.07	0.59	-64.57	1.14	7.07	0.56	-65.45	26.86
6	1.25	7.00	0.62	-63.72	1.19	7.00	0.59	-64.50	23.05
9	1.22	6.99	0.60	-65.81	1.18	6.99	0.58	-66.60	16.01
12	1.24	6.85	0.63	-64.52	1.21	6.85	0.61	-65.07	12.61
24	1.32	6.05	0.76	-54.24	1.31	6.05	0.75	-54.30	6.61
				Panel C: Short only component of the strategy					
1	0.01	8.02	0.01	-84.12	-0.30	7.98	-0.13	-86.39	113.49
2	-0.10	7.82	-0.04	-82.20	-0.26	7.81	-0.12	-83.65	56.93
3	-0.16	8.04	-0.07	-82.33	-0.27	8.03	-0.12	-83.36	40.08
4	-0.26	7.96	-0.11	-81.69	-0.35	7.96	-0.15	-84.30	32.02
5	-0.31	8.04	-0.13	-83.22	-0.38	8.03	-0.17	-86.26	26.47
6	-0.36	8.20	-0.15	-85.79	-0.42	8.20	-0.18	-88.02	22.96
9	-0.48	8.13	-0.21	-89.74	-0.52	8.13	-0.22	-90.86	16.27
12	-0.54	7.83	-0.24	-90.70	-0.57	7.84	-0.25	-91.50	12.94
24	-0.60	7.36	-0.28	-91.24	-0.61	7.37	-0.29	-91.65	6.86

Notes: The table shows the performance of long-short decile portfolios in the U.S. for the period between 1995 and 2018. Monthly mean returns, standard deviation, annualized Sharpe ratio and maximum drawdown for strategies labelled 1 to 24 are reported. The label corresponds to the horizon *h* for which we obtain the predictions and, at the same time, the holding period for a given portfolio. In Panel A are the results of the long-short portfolio. The results are decomposed into long and short components in Panel B and Panel C. The displayed values are in percentages except for the Sharpe ratio.

One may ask whether we cannot simply use one-month predictions and increase the rebalancing frequency to decrease transaction costs. The answer is that it is better to use predictions at the horizon of the desired holding period, with the exception of a holding period of two months where the difference is minimal. The Sharpe ratio and the mean are higher (mean by approximately 0.2% per month), and the standard deviation is lower for those portfolios. This holds both for the case with and without transaction costs.

Figure 1 Average Gross Returns Up to Two Years after Rebalancing



Notes: The average monthly return *x* months after rebalancing. Returns of long-short decile portfolios for the U.S. sample for the period between 1995 and 2018 are used. Portfolio returns are without transaction costs. Confidence intervals around the means are presented.

This is related to results presented in Figure 1. For multiple forecasting horizons, we show how returns vary each month for up to two years after rebalancing the portfolio. We report returns without transaction costs. It shows us that the one-month forecasting horizon is most profitable the first month after rebalancing, and then profitability lowers sharply. We can see that for a forecasting horizon of one month, the optimal rebalancing frequency is one or two months. For longer horizons, about the first five months are significant, with decreasing returns for longer holding-period months. For horizons 12 and 24, we have a significantly positive return each month. It shows us that the underlying models are indeed learning for their intended horizon. This is consistent with Leung et al. (2021) who also observe slower signal decay for 6-month return predictions compared to one month.

The importance of using multiple sub-portfolios with staggered rebalancing schedule is most pronounced for portfolios with longer holding periods. For example, we look at the performance of a 12-month decile single sub-portfolio with transaction costs. Changing only the month when we rebalance, the mean monthly returns vary between 0.47% and 0.86%, with the Sharpe ratio being between 0.67 to 0.85. In the 24-month portfolio case, single sub-portfolio mean returns range from 0.42% to 0.95%, Sharpe ratio is between 0.5 and 1.21. Utilizing multiple sub-portfolios enables us to eliminate the rebalancing timing risk, leading to less risky portfolios.

Double Sorting Portfolios

Double sorting portfolios were constructed by combining two predictions made at different horizons and rebalanced each month (holding period $b = 1$). We combine a one-month forecasting horizon with longer horizons (2, 3, 6, 9, 12, and 24 months). Equal weights are used. Cutoff points 0.15 for shorts, and 0.85 for the long side are used. The cutoffs were selected so that we have a similar number of firms in our portfolio as in the long-short decile case, allowing us to better compare with our benchmark. The average number of firms in a portfolio is between 180 and 340. The number of firms is lower when sorting on two more distant horizons as the number of common firms decreases.

Table 4 Double-Sorted Portfolios Performance in the U.S.

	Without transaction costs				*With transaction costs*				
	Mean	*Std*	*Sharpe*	*MDD*	*Mean*	*Std*	*Sharpe*	*MDD*	*Turnover*
1 - 2	1.80	5.07	1.23	-28.31	1.20	4.97	0.84	-32.16	115.63
1 - 3	2.02	5.15	1.36	-27.43	1.43	5.05	0.98	-28.11	112.26
1 - 6	1.95	4.90	1.38	-22.36	1.36	4.82	0.98	-23.19	110.43
1 - 9	2.02	4.77	1.47	-23.25	1.45	4.70	1.07	-23.65	110.08
1 - 12	2.00	4.78	1.45	-25.09	1.43	4.73	1.05	-25.30	109.39
1 - 24	2.09	4.55	1.59	-21.89	1.50	4.50	1.15	-23.99	113.03

Notes: The table shows the profitability of a double-sorted long-short portfolio in the U.S. between 1995 and 2018. Portfolio labels (1-2 to 1-24) show which two horizon predictions were used in double sorting. Results are shown with and without transaction costs. Monthly mean returns, standard deviation, annualized Sharpe ratio, and maximum drawdown are reported. Reported values are in percentages, with the exception of the Sharpe ratio.

In Table 4 are the results of double-sorted portfolios. The best-performing portfolio is a 1-24 horizon combination. After transaction costs, it has a mean return of 1.50%, an increase of 0.4% per month compared to the benchmark. At the same time, we have a lower standard deviation and maximum drawdown -24% while the benchmark has a figure almost twice as large. The other double-sorted portfolios are either slightly better or better than the benchmark. The turnover of double-sorted strategies is slightly lower than that of the benchmark. This suggests that we incur approximately the same transaction costs as our benchmark, indicating that the advantages do not stem from transaction cost differences but rather from improved firm selection.

Additional performance metrics for double-sorted portfolios are reported in Table A.2. Cumulative returns of double-sorted strategies in comparison with the benchmark can be seen in Figure A.2. The benchmark strategy is underperforming compared to the double-sorted portfolios.

Overall, double sorting portfolios can be considered better than the one-month long-short decile sorting benchmark for the U.S. Combining short-horizon predictions with longer ones brings better returns and decreased risk.

Buy/hold Spread Portfolios

Portfolios using the buy/hold spread strategy are constructed with 10%/20% cutoffs. We combine predictions from two different forecasts, using a longer horizon as a buy signal and one-month predictions as a hold signal. We expect lower turnover as it is harder to trade into a position than to trade out of it.

Portfolios performance is reported in Table 5. We refer to strategies by the buy and hold horizons that are used. Buy/hold portfolios have, on average, between 240 (for a two-year portfolio) and 290 (for a one-month portfolio) firms. Thus, it is comparable to the number of firms in the decile and double-sorted portfolios.

Table 5 Buy/Hold Spread Portfolio Performance in the U.S.

		Without transaction costs				*With transaction costs*				
buy	*hold*	*Mean*	*Std*	*Sharpe*	*MDD*	*Mean*	*Std*	*Sharpe*	*MDD*	*Turnover*
1	1	1.61	4.93	1.13	-36.74	1.15	4.85	0.82	-41.59	81.74
2	1	1.60	4.82	1.15	-32.18	1.20	4.77	0.87	-36.37	71.01
3	1	1.65	4.57	1.25	-23.36	1.29	4.51	0.99	-24.23	62.16
4	1	1.53	4.53	1.17	-21.90	1.18	4.48	0.91	-22.77	58.56
5	1	1.41	4.17	1.17	-21.60	1.07	4.13	0.90	-24.43	56.79
6	1	1.58	4.05	1.35	-19.91	1.25	4.00	1.08	-20.31	54.87
9	1	1.39	3.53	1.36	-20.87	1.08	3.50	1.07	-24.25	51.99
12	1	1.28	3.23	1.37	-19.07	0.97	3.20	1.05	-22.69	50.37
24	1	1.28	2.87	1.54	-25.02	0.97	2.84	1.18	-28.85	49.59

Notes: The profitability of long-short buy/hold spread portfolios in the U.S. for the 1995 to 2018 period. We use a buy/hold spread 10%/20% and report the results both without transaction costs and with transaction costs. Buy and hold column shows which horizons were used in the portfolio creation. Monthly mean returns, standard deviation, annualized Sharpe ratio, and maximum drawdown are reported. All values are reported in percentages except for the Sharpe ratio.

The turnover of the portfolios is approximately 50% lower than that of the benchmark. Longer horizon portfolios have lower turnovers, as fewer firms are the same as the one-month forecast. Strategies have a lower mean and risk than our benchmark before costs. Mean returns decrease with longer horizons. Sharpe ratio increases with longer horizon portfolios. After costs, buy/hold spread portfolios have better risk-adjusted profitability than the benchmark. Portfolio 24-1 has the highest Sharpe ratio, with slightly lower returns than the benchmark.

Comparing double sorting and buy/hold spread portfolios, they have similar Sharpe ratios, which increase with longer horizons. Double sorting is able to achieve significantly higher returns than the benchmark without increasing the variance. Buy/hold spread portfolios at lower horizons provide slightly higher returns and slightly lowered standard deviation, while on longer horizons, they offer a less risky alternative to double-sorted portfolios.

In Table A.3 are additional performance metrics. Cumulative returns of long-short buy/hold spread portfolios, compared with our benchmark model, can be seen in Figure A.3.

Blitz et al. (2023) also use buy/hold spread portfolios, but they do not combine multiple forecasts. They find decreasing returns and Sharpe ratios for longer-horizon portfolios without transaction costs.

We replicate this and find decreasing returns with longer horizons but before-costs Sharpe ratios are mostly similar for all horizons. We use a more liquid universe of firms, which could explain this difference. After-cost strategies have a Sharpe ratio between 0.7 and 0.95, outperforming a one-month benchmark. However, they have lower returns and lower Sharpe ratios compared to our combined forecast buy/hold spread portfolios. This adds support to our decision to combine two forecasts together.

We show that extending the rebalancing frequency while keeping the forecasting horizon equal increases risk-adjusted profitability for the U.S. sample. Combining two predictions using double sorting has performance gains compared to the benchmark. Buy/hold portfolios offer a lower-risk alternative to double-sorting portfolios.

3.2 International Evidence

Using an international dataset increases the sample size and should prevent data-snooping or overfitting concerns. However, there are possible problems with including international data. The countries may have different institutional settings, laws or accounting standards. The data preprocessing procedure we follow should lower these concerns. We train a feedforward neural network model on the international dataset (U.S. and 22 other developed countries) to obtain predictions of cumulative returns at different horizons. We form portfolios in the same way as U.S. portfolios and evaluate their performance.

Decile Portfolios

Long-short decile portfolios are created using the international sample forecasts. We keep the forecasting horizons equal to the rebalancing frequency. The average number of firms in a sub-portfolio is 600, 2.5 times more than in the U.S. setting.

In Table 6 are reported results of long-short decile portfolios with and without transaction costs. If we invested in a simple buy-and-hold of the MSCI world index, we would get 0.45% mean monthly returns and a Sharpe ratio of 0.37. One month portfolio has a mean return of 1.82% with a Sharpe ratio of 1.92 without transaction costs. The mean return is similar to our U.S. benchmark model; however, the standard deviation is almost halved for the international portfolio. Similarly, the mean return of 1.07% of the international portfolio, after transaction costs, is almost equal to that of the U.S. but with variance greatly reduced. Lower variance might be

Table 6 Performance of Long-Short Decile Portfolios - International Sample

	Without transaction costs				*With transaction costs*				
	Mean	***Std***	***Sharpe***	***MDD***	***Mean***	***Std***	***Sharpe***	***MDD***	***Turnover***
				Panel A: Long-short portfolio					
1	1.82	3.29	1.92	-23.31	1.07	3.18	1.17	-27.11	123.41
2	1.34	3.40	1.36	-26.96	0.97	3.37	1.00	-30.15	60.33
3	1.07	3.15	1.18	-21.91	0.82	3.13	0.91	-24.67	41.18
4	0.95	2.93	1.13	-25.31	0.75	2.92	0.89	-30.43	32.71
5	0.97	2.59	1.29	-23.50	0.80	2.58	1.07	-26.63	27.26
6	0.89	2.43	1.26	-26.95	0.74	2.42	1.06	-29.73	23.46
9	0.93	2.36	1.36	-23.98	0.83	2.35	1.22	-25.60	16.52
12	0.90	2.38	1.30	-27.40	0.81	2.38	1.19	-28.48	12.94
24	0.79	2.51	1.09	-34.37	0.75	2.51	1.03	-35.06	6.85
				Panel B: Long only component of the strategy					
1	1.35	5.56	0.84	-49.33	0.95	5.53	0.60	-53.09	130.05
2	1.14	5.88	0.67	-61.63	0.95	5.87	0.56	-64.54	62.48
3	1.00	5.85	0.59	-61.20	0.87	5.85	0.52	-63.35	42.17
4	1.00	5.82	0.60	-59.17	0.90	5.82	0.54	-59.69	33.16
5	1.02	5.75	0.61	-60.85	0.94	5.75	0.56	-61.29	27.56
6	0.97	5.71	0.59	-61.24	0.90	5.71	0.55	-61.62	23.70
9	1.03	5.61	0.64	-59.14	0.98	5.61	0.60	-59.42	16.52
12	1.06	5.51	0.66	-59.40	1.02	5.51	0.64	-59.60	12.86
24	1.11	5.29	0.73	-57.19	1.09	5.29	0.71	-57.28	6.79
				Panel C: Short only component of the strategy					
1	0.48	6.10	0.27	-64.06	0.10	6.05	0.06	-71.58	116.96
2	0.24	6.28	0.13	-67.58	0.05	6.26	0.03	-72.43	58.24
3	0.13	6.28	0.07	-66.69	-0.00	6.27	-0.00	-70.10	40.20
4	0.02	6.20	0.01	-70.64	-0.08	6.19	-0.05	-73.91	32.25
5	0.01	6.20	0.00	-71.61	-0.08	6.18	-0.05	-74.28	26.93
6	-0.01	6.16	-0.01	-70.96	-0.09	6.15	-0.05	-73.50	23.19
9	-0.03	6.05	-0.02	-70.88	-0.09	6.04	-0.05	-72.63	16.48
12	-0.08	5.95	-0.04	-71.77	-0.12	5.94	-0.07	-73.13	13.01
24	-0.30	5.75	-0.18	-78.13	-0.33	5.75	-0.20	-78.69	6.91

Notes: The table shows the performance of long-short decile portfolios on the international sample from 1995 to 2018. Monthly mean returns, standard deviation, annualized Sharpe ratio and maximum drawdown for strategies labelled 1 to 24 are reported. The label represents the horizon *h* for which we obtain the predictions and, at the same time, the holding period for a given portfolio. In Panel A are the results of a long-short portfolio. The results are decomposed into long and short components in Panel B and Panel C. The displayed values are in percentages except for the Sharpe ratio.

because of the larger sample or diversification. The one-month strategy turnover is 120%, comparable to the U.S. benchmark portfolio turnover. The results for the one-month predicting horizon on the international dataset are consistent with the results of Tobek and Hronec (2021).

Other portfolios on the longer horizon have similar or lower Sharpe ratios and lower returns than the one-month strategy when we account for transaction costs. For example, the nine-month portfolio has the same Sharpe ratio as a one-month portfolio and returns lower by 0.24% after transaction costs, offering a lower-risk alternative.

Looking at the long and short-leg components of strategies separately, there is a difference in contribution to return between international and U.S. cases. Internationally, short legs are more successful. For shorter horizons, the international portfolios' short legs have positive mean returns even after accounting for transaction costs.

In Table A.4 are presented additional performance measures. Alpha and Beta are calculated with respect to international market returns. In Figure A.4 are cumulative returns of these portfolios with and without transaction costs compared to the one-month international benchmark portfolio. Similarly to the U.S. sample, there is a visible decrease in profitability over time. After 2005, the longer-horizon portfolios start to have better risk-adjusted returns compared to the one-month portfolio. This decrease in profitability over time is analyzed in subsection D.1.

Double Sorting Portfolios

Double-sorted portfolios are created using forecasts from two models with different forecasting horizons. Cutoff points are 0.15 for the short side and 0.85 for the long side. The average number of firms in a double-sorted portfolio is between 490 (for 1-24 portfolio) and 920 (for 1-2 portfolio).

Performance measured of double-sorted portfolios are in Table 7. Portfolios 1-12 and 1-24 have the highest Sharpe ratios, close to that of our international benchmark. They also have a similar Sharpe ratio as double-sorted portfolios in the U.S. sample but offer a lower return. Turnover is comparable with the benchmark.

Table 7 Double-Sorted Portfolios Performance - International Sample

	Without transaction costs				*With transaction costs*				
	Mean	*Std*	*Sharpe*	*MDD*	*Mean*	*Std*	*Sharpe*	*MDD*	*Turnover*
1 - 2	1.68	3.51	1.66	-29.46	0.96	3.42	0.98	-35.17	117.87
1 - 3	1.69	3.58	1.64	-27.06	1.00	3.50	0.99	-32.87	114.12
1 - 6	1.66	3.49	1.65	-29.06	0.98	3.40	1.00	-34.95	112.66
1 - 9	1.77	3.41	1.79	-28.05	1.09	3.31	1.14	-33.87	111.64
1 - 12	1.83	3.34	1.89	-23.41	1.14	3.24	1.22	-29.72	111.64
1 - 24	1.86	3.38	1.91	-16.04	1.16	3.28	1.23	-20.67	113.88

Notes: The table shows the profitability of a double-sorted long-short portfolio using the international sample for the period between 1995 and 2018. The portfolio label shows which two forecasting horizons were used in double sorting. Results are shown with and without transaction costs. The holding period of portfolios is one month. Monthly mean returns, standard deviation, annualized Sharpe ratio, and maximum drawdown are presented. Reported values are in percentages, with the exception of the Sharpe ratio.

Table A.5 reports Sortino ratio, conditional value at risk, Alpha and Beta for double-sorted portfolios. Cumulative returns of double-sorted portfolios and of benchmark model are in Figure A.5.

Buy/hold Spread Portfolios

Long-short buy/hold spread (10%/20%) portfolios were constructed using predictions made on the international sample. The average number of firms in a portfolio is between 570 and 650, with the number of firms being lower with longer forecasting horizons.

Results are reported in Table 8. Portfolios 9-1 and 12-1 have the highest Sharpe ratio, which is similar to that of the one-month international benchmark. It offers slightly lower returns. Other portfolios have similar returns as 9-1 but higher variance. Compared to double sorting portfolios, it has lower returns but similar Sharpe ratios. Turnover of buy/hold spread strategies is lower than benchmark turnover and comparable with the buy/hold spread strategy in the U.S.

Table 8 Buy/hold Spread Portfolio Performance - International Sample

		Without transaction costs				*With transaction costs*				
buy	*hold*	*Mean*	*Std*	*Sharpe*	*MDD*	*Mean*	*Std*	*Sharpe*	*MDD*	*Turnover*
1	*1*	1.84	3.21	1.98	-17.95	1.05	3.15	1.15	-25.51	78.66
2	*1*	1.67	3.46	1.67	-19.82	0.98	3.41	0.99	-26.46	67.08
3	*1*	1.60	3.40	1.63	-19.00	0.98	3.36	1.01	-24.97	58.71
4	*1*	1.52	3.28	1.61	-22.38	0.92	3.24	0.99	-27.31	55.15
5	*1*	1.51	3.09	1.70	-21.87	0.93	3.05	1.06	-27.01	53.24
6	*1*	1.43	3.05	1.62	-19.86	0.87	3.01	1.00	-26.79	51.32
9	*1*	1.51	2.84	1.84	-22.24	0.99	2.81	1.22	-27.12	47.81
12	*1*	1.44	2.77	1.80	-26.12	0.92	2.73	1.17	-31.37	46.85
24	*1*	1.33	3.06	1.51	-33.52	0.82	3.03	0.94	-38.85	46.79

Notes: The profitability of long-short buy/hold spread portfolios on the international universe for the period between 1995 and 2018. We use a buy/hold spread 10%/20% and report the results both without transaction costs and with transaction costs. Monthly mean returns, standard deviation, annualized Sharpe ratio, and maximum drawdown are presented. All values are reported in percentages except for the Sharpe ratio.

Additional performance metrics for portfolios are reported in Table A.6. Beta coefficients are all close to zero. In Figure A.6 are cumulative returns of buy/hold spread strategies and of benchmark strategy.

Overall, portfolios made using the international dataset offer lower-risk opportunities compared to the U.S. sample. A one-month long-short decile portfolio performs well, even after accounting for transaction costs, which are higher on the international sample than in the U.S. There are comparable portfolios available when considering longer horizons or combinations of horizons.

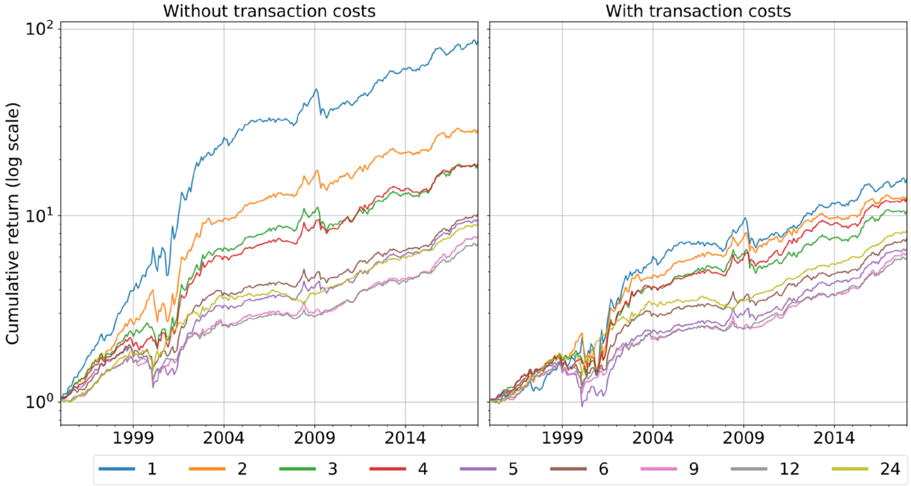
4. Conclusion

Our study investigates the predictability of global stock returns using machine learning techniques across various forecasting horizons. We find that predictability diminishes with longer horizons, both in the U.S. and internationally. We show that machine learning portfolios remain profitable even after accounting for transaction costs, especially with longer horizons. We use neural networks and gradient-boosted trees in predictive regressions for stock returns using 153 anomalies documented in the literature as variables. We document that the predictability of returns decreases with longer forecasting horizons both in the U.S. and internationally. We work with a highly liquid universe and estimated transaction costs to mitigate concerns that profitability is concentrated in small, difficult-to-arbitrage stocks and diminishes after factoring in transaction costs. We construct a number of portfolios using longer-horizon forecasts, allowing us to reduce turnover or combine multiple horizons. After accounting for the transaction costs, longer horizons long-short portfolios offer better risk-adjusted returns in the U.S. This holds even in more recent times. Post-2005, while overall profitability decreased, longer horizons consistently yielded higher Sharpe ratios compared to one-month portfolios in both the U.S. and international markets. Leveraging return predictions for multiple horizons via double-sorted portfolios leads to profitability improvement in the U.S. Finally, we employ a turnover-reducing strategy, buy/hold spread, and show higher risk-adjusted profitability in the U.S.

APPENDIX

A. Additional tables and figures

Figure A.1 Cumulative Returns of Long-Short Decile Portfolios in the U.S.



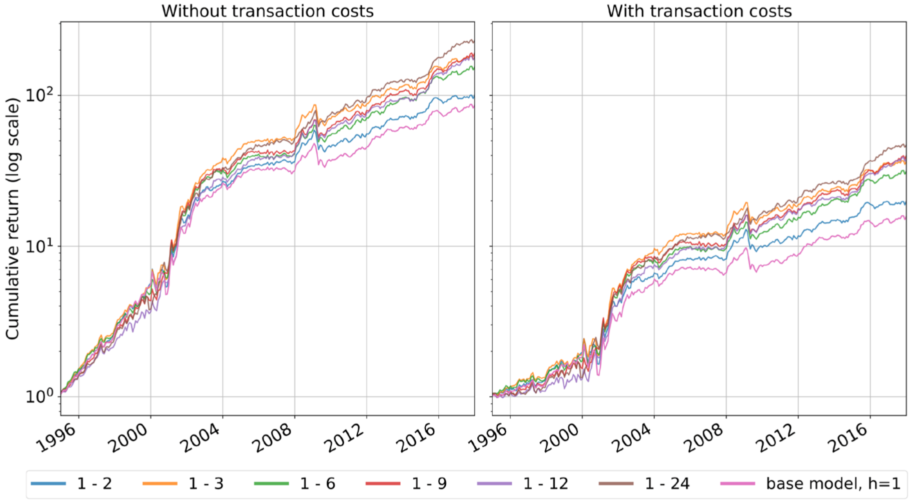
Notes: The figure shows cumulative returns of long-short decile portfolios without and with transaction costs on the U.S. sample. The portfolio label is the forecasting horizon in months and holding period of the strategy.

Table A.1 Performance Measures of Long-Short Decile Portfolios in the U.S.

	Without transaction costs				*With transaction costs*			
	Sortino	***CVaR 99%***	***Alpha***	***Beta***	***Sortino***	***CVaR 99%***	***Alpha***	***Beta***
1	2.27	-9.63	1.92	-0.22	1.33	-10.56	1.29	-0.22
2	1.97	-8.50	1.43	-0.13	1.43	-8.94	1.12	-0.13
3	1.88	-7.38	1.20	-0.12	1.46	-7.66	1.00	-0.12
4	2.01	-6.52	1.13	-0.10	1.64	-6.82	0.97	-0.10
5	1.57	-6.84	0.94	-0.08	1.29	-7.06	0.81	-0.08
6	1.83	-5.90	0.97	-0.11	1.56	-6.08	0.86	-0.11
9	1.74	-5.23	0.80	-0.06	1.53	-5.35	0.73	-0.06
12	1.90	-4.65	0.74	-0.02	1.71	-4.73	0.68	-0.02
24	2.07	-4.59	0.83	-0.02	1.97	-4.65	0.80	-0.02

Notes: Sortino ratio, conditional value at risk at 99%, Alpha and Beta (in comparison with U.S. market returns) are reported for long-short decile portfolios for the period between 1995 and 2018. Portfolio label is the forecasting horizon and the holding period for the portfolio.

Figure A.2 Cumulative Returns of Long-Short Double Sorting Portfolios in the U.S.



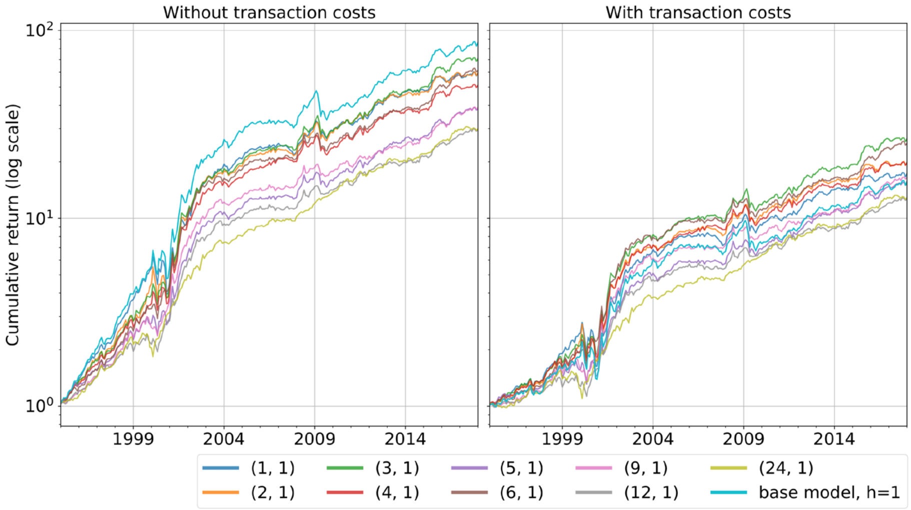
Notes: The figure shows cumulative returns on a logarithmic scale of the double-sorting strategy and of the long-short decile portfolio at horizon one in the U.S. The two numbers correspond to horizons on which we double-sorted. The holding period is one month.

Table A.2 Double-Sorted Portfolios Performance Metrics - in the U.S.

	Without transaction costs				***With transaction costs***			
	Sortino	***CVaR 99%***	***Alpha***	***Beta***	***Sortino***	***CVaR 99%***	***Alpha***	***Beta***
1 - 2	2.57	-8.87	1.94	-0.19	1.56	-9.70	1.35	-0.20
1 - 3	3.02	-8.50	2.19	-0.23	1.95	-9.27	1.60	-0.23
1 - 6	3.21	-7.65	2.14	-0.26	1.99	-8.43	1.56	-0.27
1 - 9	3.35	-7.58	2.24	-0.30	2.13	-8.35	1.67	-0.30
1 - 12	3.11	-8.07	2.22	-0.30	1.99	-8.83	1.65	-0.31
1 - 24	3.35	-7.83	2.21	-0.18	2.13	-8.78	1.63	-0.18

Notes: Sortino ratio, conditional value at risk at 99%, Alpha and Beta (with regards to market returns in the U.S.) are presented for long-short double-sorting portfolios for the period between 1995 and 2018. Portfolio labels are the two forecasting horizons which were used in double sorting.

Figure A.3 Cumulative Returns of Buy/Hold Spread Portfolios in the U.S.



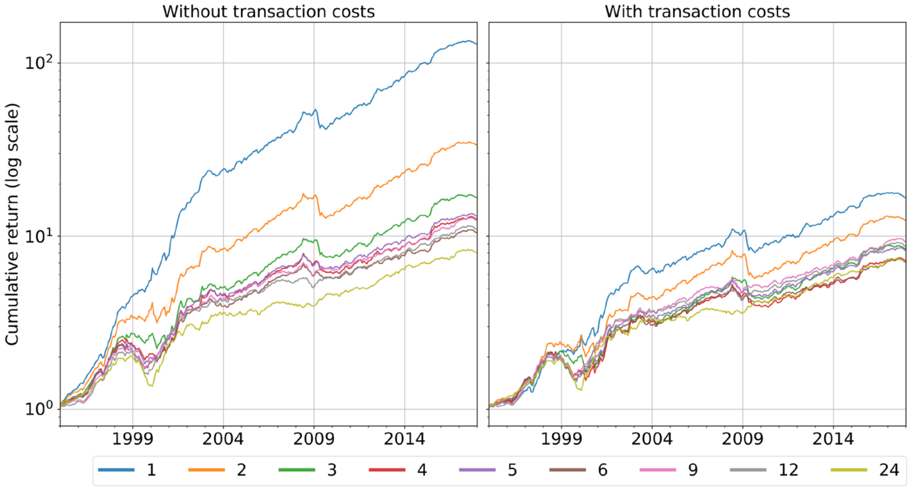
Notes: Cumulative returns of long-short buy/hold spread portfolios compared with the benchmark model. We use buy/hold spread 10%/20%. The portfolio label signifies the horizon based on which we buy and hold stocks, respectively.

Table A.3 Buy/Hold Spread Portfolio Performance Metrics - U.S. Sample

		Without transaction costs				***With transaction costs***			
buy	***hold***	***Sortino***	***CVaR 99%***	***Alpha***	***Beta***	***Sortino***	***CVaR 99%***	***Alpha***	***Beta***
1	*1*	2.10	-9.17	1.71	-0.14	1.42	-9.74	1.25	-0.14
2	*1*	2.32	-8.87	1.71	-0.15	1.62	-9.49	1.31	-0.15
3	*1*	2.68	-7.79	1.77	-0.17	1.96	-8.24	1.41	-0.17
4	*1*	2.46	-7.59	1.65	-0.17	1.77	-8.12	1.30	-0.17
5	*1*	2.26	-7.98	1.51	-0.14	1.60	-8.47	1.17	-0.14
6	*1*	2.94	-6.67	1.74	-0.23	2.17	-7.06	1.41	-0.23
9	*1*	2.91	-6.19	1.54	-0.21	2.08	-6.54	1.23	-0.21
12	*1*	2.95	-5.44	1.42	-0.19	2.03	-5.85	1.11	-0.19
24	*1*	3.10	-5.04	1.35	-0.11	2.14	-5.45	1.05	-0.11

Notes: Sortino ratio, conditional value at risk at 99%, Alpha and Beta (in comparison with U.S. market returns) are reported for long-short buy and holds spread for the period between 1995 and 2018. We use 10%/20% buy/hold spread cutoffs. The portfolio label signifies the horizon based on which we buy and the horizon based on which we hold stocks, respectively.

Figure A.4 Cumulative Returns of Long-Short Decile Portfolios on International Sample



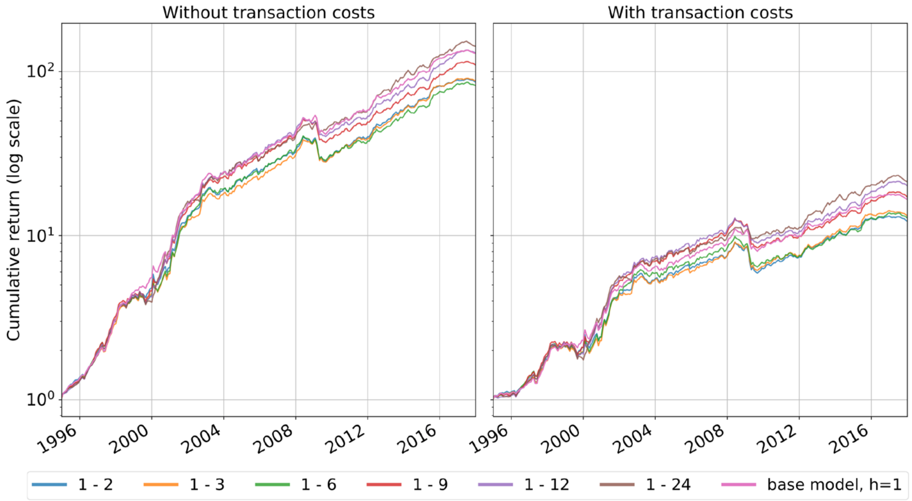
Notes: The figure shows cumulative returns of long-short decile portfolios without and with transaction costs. The portfolio label is the forecasting horizon in months and the holding period of the strategy.

Table A.4 Performance Measures for Long-Short Decile Portfolios - International Sample

	Without transaction costs				***With transaction costs***			
	Sortino	***CVaR 99%***	***Alpha***	***Beta***	***Sortino***	***CVaR 99%***	***Alpha***	***Beta***
1	4.53	-4.95	1.89	-0.12	2.26	-5.71	1.13	-0.11
2	2.53	-6.64	1.39	-0.08	1.70	-7.09	1.02	-0.08
3	2.17	-6.29	1.11	-0.06	1.56	-6.63	0.86	-0.06
4	2.09	-5.43	0.97	-0.03	1.55	-5.73	0.77	-0.03
5	2.54	-4.52	0.98	-0.03	1.98	-4.79	0.82	-0.03
6	2.42	-4.32	0.91	-0.03	1.93	-4.52	0.76	-0.03
9	2.76	-4.10	0.93	0.01	2.36	-4.25	0.82	0.01
12	2.61	-4.29	0.88	0.03	2.30	-4.40	0.80	0.03
24	2.13	-4.35	0.76	0.06	1.98	-4.41	0.71	0.06

Notes: Sortino ratio, conditional value at risk at 99%, Alpha and Beta (in comparison with international market returns) are reported for long-short decile portfolios for the period between 1995 and 2018. The portfolio label is the forecasting horizon and the holding period.

Figure A.5 Cumulative Returns of Long-Short Double Sorting Portfolios - International Universe



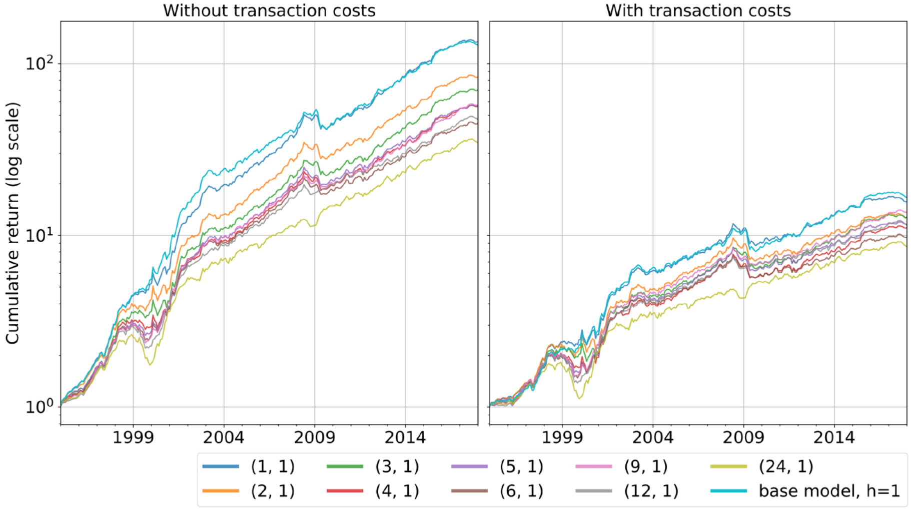
Notes: The figure shows cumulative returns of the double sorting strategy in comparison with the long-short decile portfolio at horizon one, both for the international universe. Portfolios are plotted before and after accounting for transaction costs. The portfolio label signifies the two horizons that are used to double-sort. The holding period is one month.

Table A.5 Double-Sorted Portfolios Performance Metrics - International Sample

	Without transaction costs				***With transaction costs***			
	Sortino	***CVaR 99%***	***Alpha***	***Beta***	***Sortino***	***CVaR 99%***	***Alpha***	***Beta***
1 - 2	3.70	-5.74	1.75	-0.12	1.83	-6.47	1.03	-0.12
1 - 3	3.70	-5.88	1.77	-0.14	1.88	-6.61	1.07	-0.13
1 - 6	3.59	-5.83	1.75	-0.16	1.82	-6.56	1.07	-0.15
1 - 9	4.16	-5.37	1.86	-0.16	2.18	-6.10	1.17	-0.15
1 - 12	4.72	-4.82	1.90	-0.13	2.46	-5.52	1.21	-0.12
1 - 24	4.63	-5.15	1.91	-0.08	2.42	-5.91	1.20	-0.07

Notes: Sortino ratio, conditional value at risk at 99%, Alpha and Beta (with regards to the international market returns) long-short double sorting portfolios for the period between 1995 and 2018. Portfolio labels are the two forecasting horizons which were used in double sorting.

Figure A.6 Cumulative Returns of Buy/Hold Spread Portfolios on International Sample



Notes: Cumulative returns of long-short buy/hold spread portfolios in comparison with the base model, both on the international universe. We use a buy/hold spread of 10%/20%. The portfolio label signifies the horizon based on which we buy and hold stocks, respectively.

Table A.6 Buy/Hold Spread Portfolio Performance Metrics - International Sample

		Without transaction costs				***With transaction costs***			
buy	***hold***	***Sortino***	***CVaR 99%***	***Alpha***	***Beta***	***Sortino***	***CVaR 99%***	***Alpha***	***Beta***
1	*1*	4.81	-4.92	1.87	-0.05	2.25	-5.77	1.08	-0.05
2	*1*	3.76	-5.64	1.69	-0.03	1.87	-6.36	1.00	-0.03
3	*1*	3.69	-5.62	1.63	-0.05	1.92	-6.32	1.00	-0.04
4	*1*	3.53	-5.40	1.54	-0.03	1.84	-6.08	0.94	-0.03
5	*1*	3.86	-4.94	1.53	-0.04	2.02	-5.58	0.95	-0.03
6	*1*	3.65	-4.87	1.45	-0.04	1.89	-5.46	0.89	-0.04
9	*1*	4.47	-4.22	1.53	-0.04	2.45	-4.80	1.00	-0.03
12	*1*	4.02	-4.52	1.44	0.00	2.20	-5.05	0.92	0.01
24	*1*	3.19	-5.13	1.30	0.05	1.71	-5.67	0.79	0.06

Notes: Sortino ratio, conditional value at risk at 99%, Alpha and Beta (in comparison with the international market returns) for long-short buy/hold spread portfolio made on the international universe for the period between 1950 and 2018. We use a buy/hold spread of 10%/20%. The portfolio label signifies the horizon based on which we buy and hold stocks, respectively.

B. Methodology

B.1 Machine learning

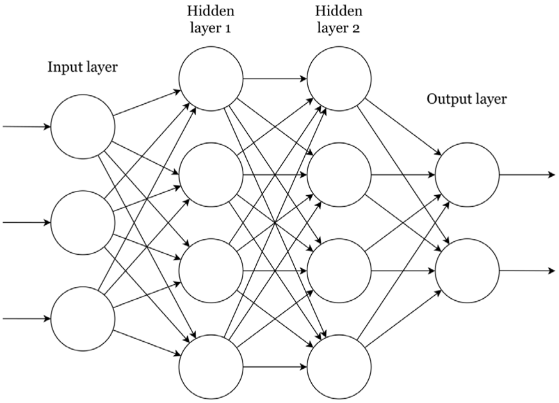
This section gives an overview of the feedforward neural networks, gradient-boosted trees, and algorithms that will be used. For more details, see Goodfellow et al. (2016) or Hastie et al. (2009).

Feedforward Neural Network

Feedforward neural network consists of an input layer of raw predictors, one or multiple hidden layers and an output layer. Each layer is composed of nodes, also called neurons. The nodes can be fully connected to all nodes in the previous and next layer or only to some of them.

Figure B.1 shows an example of a neural network that is fully connected, has three inputs, two hidden layers, each with four neurons and an output layer with two outputs.

Figure B.1 Example of Multilayer Fully Connected Neural Network



Neuron i is defined as:

$$y_i = \varphi(s_i + b_i), \;\; s_i = \sum_{j=1}^{m} w_{ij} x_j$$

with $x_1, \ldots x_m$ being neuron inputs, $w_{i1}, \ldots w_{im}$ are synaptic weights, b_i is bias term for a given neuron, $\varphi(\cdot)$ is the activation function, and y_i is the output of the neuron i.

A Commonly used activation function, and the one that we will be using, is called rectified linear unit (ReLU), and it is defined as:

$$ReLU(x) \; = \begin{cases} 0 & if \; x < 0 \\ x & otherwise \end{cases}$$

Other used types of activation functions are, for example, sigmoid, hyperbolic tangent, piece-wise linear or threshold activation functions.

Optimization

Machine learning minimizes **loss function**[8]. For a function with one input, the derivative $f'(x)$gives us the slope of $f(x)$at x telling us in which direction to move. We might encounter multiple problems that will make it impossible to reach the global minimum using this procedure. Those are local minimums or saddle points. In the case of working with multiple inputs, we need to work with gradients, and we move in the direction of the steepest descent - known as **gradient descent**.

Stochastic gradient descent (SGD) is an extension of gradient descent. With larger datasets, the time to move even one step in the right direction using gradient descent takes too long as we need to use the entire dataset to compute the gradient. Instead, we approximately estimate the gradient using a small and random sample called a minibatch. The approximation greatly speeds up the optimization and allows us to work with large datasets.

We will be using an extension of stochastic gradient descent, namely **Adam** optimization algorithm (short for adaptive moments) proposed by Kingma and Ba (2014). It is based on computing adaptive estimates of the first and second moments of gradients.

When we move in the direction of the steepest descent, the size of the step, ϵ, is called a **learning rate**. It is a positive scalar, and there are different methods of choosing the learning rate. The simplest one is to set it to a small constant. To speed up the convergence, it is common to decrease the learning rate during the learning process. We will use decaying learning rate, more specifically reducing learning rate on a plateau by a fixed factor when after a certain number of epochs, there was no improvement to validation error. The epoch term means that the network has seen the entire dataset once. Other learning rate decay schemes include linear decay until reaching a fixed minimum or exponential decay.

When using a training feedforward neural network or obtaining predictions, **forward propagation** is employed. Forward propagation is the calculation of the final output of the model, given the inputs. This includes calculating the output value of each node in the network so that we can obtain the final output. With predictions and real values available, we compute the loss $\mathcal{L}(\theta)$.

As a loss function, we are using mean squared error:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{n=1}^{N}\left(\theta_n - \hat{\theta}_n\right)^2$$

where N is the batch size, θ_n is the target value, and $\hat{\theta}_n$ is the estimated value of *n*th observation.

The **backpropagation algorithm** efficiently calculates the gradient of the loss function with respect to the parameters of the network. The efficiency comes from using the chain rule and from iterative calculation backwards through the network, which avoids unnecessary calculations. The calculated gradient lets us see which node is responsible for most of the error and lets us change the parameters

[8] Also called the objective function, criterion, or error function.

accordingly. We adjust the weights by a learning rate multiplied by the gradient of the loss function with respect to a given weight.

Regularization Techniques

Regularization of neural networks controls the kind of functions we allow our model to take or specify which functions are preferred. Regularization is a modification to the neural network with the aim of reducing generalization error to prevent overfitting. We use regularisation techniques: early stopping, batch normalization, ensemble and dropout.

Early stopping is a form of regularization. When we train the model, the training error reduces over time; however, the validation error is rising after a certain time, signalling overfit. Early stopping is a rule to stop the learning when, after a certain number of epochs, given by the patience parameter, the improvement to the validation error is lower than the specified threshold. We set this threshold to zero so that we stop learning when there is no improvement.

Batch normalization by Ioffe and Szegedy (2015) is used to prevent an internal covariate shift. Internal covariate shift means that the distribution of inputs to the layer changes during the learning as the parameters of preceding layers change. It poses a problem as the layers need to continuously adapt to the changing distribution, and small changes to the parameters could be greatly amplified further in the network. Batch normalization addresses this by normalizing the input of each layer for each minibatch during the training. It allows us to use higher learning rates and works as a regularization.

Ensembles are used to lower the generalization error by averaging several models. We train the model multiple times with different starting seeds and average the predictions from them to get the final prediction. The ensemble will work at least as well as any individual models, and if models make independent errors, the ensemble will be better. The different initialization works to get at least partially independent errors. The disadvantage of using ensembles in machine learning is their computational cost.

Dropout is a technique developed by Srivastava et al. (2014) to prevent overfitting similarly to an ensemble but using only one model. It provides an efficient way to combine many network architectures by randomly dropping nodes and their connections from the network as we train it. It prevents the nodes from co-adapting too much. At each step, the node is activated with probability p and connected to the next layer with weight w. When we predict, we use a single unthinned network that has smaller weights to account for the time the node was not activated during the training.

Gradient Boosted Regression Trees

Gradient-boosted regression trees employ decision trees and a technique called gradient boosting. Decision trees can be divided into classification trees, where the leaf contains the class to which the data supplied belongs and regression trees, where the leaves are real numbers. Classification and Regression Tree (CART) is a term which covers both categories. CART creates binary trees - each non-terminal node is split into two nodes. The benefits of trees include intuitive interpretation and the fact that they allow for both numerical values and categorical values in one model.

As only one tree is usually not sufficiently strong to be used alone, techniques were developed to combine multiple trees, called ensemble models. Examples are boosted trees, random forests or rotation forests.

Boosted regression trees were first proposed by Friedman (2001). Gradient boosting is a machine learning technique that uses an ensemble of models that are iteratively learned. In this iterative learning, each added model is working to correct the mistakes of the current ensemble model. These ensemble models are often, but not necessarily, trees.

We use the implementation of boosted trees called XGBoost (Extreme Gradient Boost) by Chen and Guestrin (2016). It employs computing of second-order gradients to improve the performance, allowing regularization to improve generalization.

The tree is defined as:

$$f_t(x) = w_{q(x)}, w \in R^T, q{:}\, R^d \rightarrow \{1,2,\cdots,T\}$$

where w is a vector of scores on leaves, and q is a function which assigns each observation to the corresponding leaf. T is the number of leaves.

A tree ensemble with K additive functions then forms the final model and final predictions.

$$\hat{y}_i = const. + \sum_{k=1}^{K} \nu f_k(x_i)\,, \quad f_k \in \mathcal{F}$$

Where $\mathcal{F}$ is the space of all CART. f_k is an independent tree. Each added tree is multiplied by the shrinkage parameter v. *const.* is our starting point before fitting the first tree.

Our loss function is the following:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2$is the regularization term which penalizes the complexity and avoids overfitting. l is the differentiable convex training loss function; in our case, we will use mean square error.

The model is trained using additive strategy. At iteration t (out of a total of K) the prediction is:

$$\hat{y}_i^{(t)} = \phi(x_i) = \hat{y}_i^{(t-1)} + \nu f_t(x_i)$$

where ν is the shrinkage parameter that shrinks the influence of the tree that is being added to avoid overfitting. It also allows subsequent trees room for improvement of the model.

When learning, at *t*-th iteration, we fit tree f_t which minimizes

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

Note that the goal of f_t is to minimize loss with respect to residuals from the previous predictions $\hat{y}_i^{(t-1)}$ while taking into account the regularization term.

f_t from this equation can be approximated by the second-order Taylor approximation

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

where g_i, h_i are first and second-order derivations of the loss function.

I_R and I_L are instances of sets of right and left nodes, I is their union. To evaluate whether to split the node or not, we compare I_R and I_L with the I to see whether there is a loss reduction after splitting. More formally

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i\right)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma$$

where γ is the regularization term on the additional leaf. We select the best split based on $\mathcal{L}_{split}$, and if it is positive, we add the branch to the tree.

When fitting the tree, it is not feasible to search through all of the possible splits. We instead have a certain number of quantiles on a characteristic, and we test only these splits in our search.

The XGBoost also employs feature subsampling, which prevents overfitting and also speeds up the optimization. Excluding a random portion of characteristics in each tree allows us to get more diverse models by ensuring that not all of the trees are split on the dominant characteristic (i.e. firm size).

B.2 Performance Evaluation

The most apparent metrics are the mean and standard deviation of returns. The downside of using standard deviation to be mindful is that positive returns are treated the same way as negative ones.

Sharpe ratio is defined as the difference between average return and risk-free rate for a given period divided by the standard deviation of the rate of return. Formally:

$$SR_k = \frac{E\left[R_k - R_f\right]}{\sqrt{var(R_k)}}$$

Proposed by Sharpe (1966) under the name reward-to-variability ratio, it became a commonly used measure of performance. Sharpe (1994) proposes an extension to the Sharpe ratio so we can also compare to the benchmark changing over time.

$$SR_k = \frac{E\left[R_k - R_f\right]}{\sqrt{var(R_k - R_b)}}$$

Sharpe ratio's weak point is that it takes standard deviation as risk, disregarding whether it is upside or downside volatility and treating both the same.

Sharpe ratio is usually presented in the annualized form. It can be calculated by multiplying the Sharpe ratio with the square root of 12 in case we are using monthly data. Sharpe ratio may be negative in some cases, and it has been shown to be unreliable in judging the performance of a strategy in these cases (Scholz, 2007). In spite of this, we will be using the annualized Sharpe ratio because of its simplicity and widespread use in related literature.

A crucial and often overlooked fact is that the Sharpe ratio is also simply a rescaled t-statistic for the statistical significance of the mean being different from zero. A t-statistic can be obtained from the Sharpe ratio by multiplying by the square root of the number of observations and dividing by the square root of 12 in case the ratio was annualized. When comparing different strategies with the same number of observations, the ratios are proportional to the t-statistic.

To counter some of the problems of the Sharpe ratio, we include Sortino Ratio. It is a modification of the Sharpe ratio by Sortino and Price (1994) that penalizes only returns that are below the minimum acceptable return (MAR). This way, only the variation below MAR is counted in the denominator. The Sortino ratio is calculated as:

$$\text{Sortino}_k = \frac{E[R_k - MAR]}{\sqrt{\frac{1}{T}\sum_{t=1}^{T} min\left(0; R_{t,k} - MAR\right)^2}}$$

The denominator measures downside deviation. A minimum acceptable return of 0% will be used when using the Sortino ratio.

So far mentioned metrics do not consider the tail risk of a portfolio. The Value at Risk (VaR) is a measure of the risk of loss that tells us how much we can lose with a specified confidence level $\alpha \in (0,1)$ in a set time period. From Föllmer and Schied (2011):

$$\text{VaR}_\alpha(X) = \inf\{x \in \mathbb{R}: P(X + x < 0) \leq 1 - \alpha\}$$

VaR is not a coherent measure as it fails to hold the subadditivity axiom of coherence. Meaning that the VaR of holding a portfolio is not necessarily equal to or lower than the sum of the VaRs of individual components.

Conditional Value at Risk (Rockafellar, Uryasev, et al., 2000), for which the subadditivity holds, is defined as:

$$\text{CVaR}_\alpha = \frac{1}{\alpha}\int_0^\alpha VaR_\alpha(X)d\alpha$$

It gives the average value at risk at level $\alpha \in (0,1)$ of a position X. For example, CVaR 99% is the expected return on the portfolio in 1% of the worst cases.

Portfolio drawdown (underwater) is defined as a drop in portfolio value compared to the achieved maximum in the past. With $R_p(w_1, \dots, w_n, t)$being the cumulative portfolio return over portfolio holding time drawdown is defined as

$$D(w, t) = \max_{0 \le \tau \le t}\left\{R_p(w, \tau)\right\} - R_p(w, t)$$

Maximum Drawdown up to time *T* is:

$$MDD(T) = \max_{0 \le \tau \le T}\{D(w, \tau)\}$$

The maximum drawdown represents the largest peak-to-trough decline observed. Although average drawdown can also be utilized, it is less prevalent than maximum drawdown[9].

To compare our results with a benchmark, we use a single-index model developed by Sharpe (1963), which is an asset pricing model measuring the risk and return of a portfolio relative to another portfolio. It is defined as

$$R_{s,t} - R_f = \alpha_i + \beta_i\left(R_{M,t} - R_f\right) + \epsilon_{i,t}$$

Where R_s is the return of our portfolio, R_M is the market return, and R_f is the risk-free rate. The two coefficients, Alpha and Beta, are of interest as they tell us the abnormal return and exposure to market movements.

B.3 Transaction cost proxies

Turnover

The turnover, the percentage of monthly change of holdings, is defined as:

$$Turnover_t = \frac{1}{ge}\sum_{i=1}^{n}|ts_{it}|$$

where ge is gross exposure, the sum of long and short positions divided by the capital, and ts_{it} is the trade size for firm i at a given month. A turnover of 200% means that the entire portfolio was liquidated, and new stocks were bought for both the long side and the short side of the portfolio. Turnover of a portfolio is indicative of transaction costs paid. However, some portfolios may select especially costly firms to trade while keeping the turnover low.

We are using our preprocessed daily dataset to estimate transaction costs for each firm in a given month. Closing quoted spread (Chung and Zhang, 2014) and volatility over volume (Fong et al., 2018) proxies are used.

[9] for example in Avramov et al. (2020), Gu et al. (2020), and Tobek and Hronec (2021)

Closing Quoted Spread

Closing quoted spread proxy by Chung and Zhang (2014) is defined as:

$$QS = \frac{1}{T}\sum_{t=1}^{T}\frac{2(ask - bid)}{ask \; + \; bid}$$

with bid being the closing bid, ask being the closing ask, and T being the number of days for a given month. If the daily value of QS is missing or negative, it is not included in the calculation of the average. The downside of the quoted spread is that it is not available for the whole sample period in all of the regions as it requires closing bid and ask, which is frequently not available in the earlier periods.

Volatility over volume (% spread)

Volatility over volume (VoV) (% spread) proxy was introduced by Fong et al. (2018), and it is defined as:

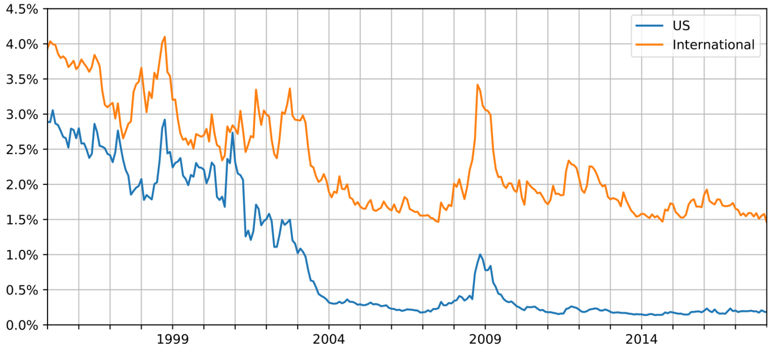
$$VoV(\%\, spread) = 8\frac{\sigma^{2/3}}{avg\ vol^{1/3}}$$

with σ being the standard deviation of daily returns, $avg\ vol$ being the average daily trading volume for a given month. The trading volume is in U.S. dollars and is deflated to 2000 prices. It roughly measures the fixed component of transaction costs.

Fong et al. (2018) benchmarked it to other transaction cost proxies, showing that only the closing quoted spread outperformed this proxy. VoV proxy has fewer missing observations than the quoted spread as it uses returns and volume only and not closing bid and ask.

To estimate transaction costs, we use closing quoted spread (Chung and Zhang, 2014). Missing observations are filled in with volatility over volume (Fong et al., 2018). For the remaining observations with missing values, we assume transaction costs of 5%. Average estimated transaction costs over time are displayed in Figure B.3.

Figure B.3 Average Estimated Transaction Costs



Notes: Estimated transaction costs cross-sectional average for the U.S. and international sample (with the U.S. excluded).

C. Data Preprocessing and Filtering

C.1 U.S. Data Processing

CRSP/Compustat Merged Database from the Center for Research in Security Prices is used. It is a comprehensive, survivorship bias-free and accurate database. CRSP at daily and monthly frequency is used, daily is used for estimating transaction costs and monthly for returns and characteristics calculation. COMPUSTAT fundamental data are used at a yearly frequency. Quarterly fundamentals are available; however, the international coverage of quarterly data is problematic, so we do not use them to keep the U.S. and international datasets comparable. The dataset includes stocks that are (or were) listed on the New York Stock Exchange (NYSE), the Nasdaq Stock Exchange (NASDAQ) or the American Stock Exchange (AMEX), among others. The sample used is from the period between 1963 and 2018.

Handling of CRSP and COMPUSTAT data mostly follows Bali et al. (2016). For the monthly dataset, we need to ensure that we only include securities that were available to trade on the last day of the month *t*. We thus include only firms with a starting date at the latest on the last day of the month *t*, and the ending date has to be on the last day of the month or later. The preprocessing of daily and monthly datasets is otherwise the same.

To get U.S. shares only, we filter based on the SHRCD share code being 10 or 11. To include only common equity firms in our dataset, we select firms with exchange code (EXCHCD) 1, 2 or 3.

Market capitalization is calculated as the absolute value of the number of shares outstanding (SHROUT) times the price of the stock at the end of the month (ALTPRC). ALTPRC is used as the PRC variable is missing or zero if the stock was not traded. Absolute value is taken as CRSP reports a negative price, equal to the average of bid and ask if the stock was not traded that day. If SHROUT or ALTPRC is missing, we mark market capitalization as missing.

As for returns, most of the time return (RET) variable can be used with the exception of the last month when the firm is active. When the firm delists, the RET does not correspond to the real return that an investor would get. If the stock is delisted, but the investor does not liquidate the position (this behaviour is expected as it is a sudden change without much warning in many cases), he ends up with untradeable stock. The CRSP includes delisting returns DLRET, the reason for the delisting and the date of delisting.

C.2 International Data Processing

As a source of international cross-sectional equity data, we use Datastream. We use a sample from January 1980 to 2018. The starting year is limited by the coverage of fundamental data in the Worldscope database. Datastream comprises several databases which we will use. Daily pricing data (unadjusted price, total return index, market value, number of shares outstanding, unadjusted volume, dividends and others), yearly fundamental data from Worldscope database (i.e. accruals, inventory or earnings) and I/B/E/S Estimates (Institutional Brokers Estimate System) are used. Where currency is needed, we use U.S. dollars.

One of the reasons why the research is focused on the United States equity market is the high reliability of the data available. For the U.S., we have available CRSP and COMPUSTAT datasets, which are well-checked and reliable. Having a reliable international dataset is valuable as we can provide evidence that anomalies found in U.S. data are not data snooping.

In order to get the dscodes (identification of firm listings), we use constituent lists provided by Datastream and Worldscope. These lists include Datastream research lists, Datastream dead lists and Worldscope coverage lists for each country. These lists contain around 230 thousand dscodes. This number is, however, greatly reduced when we filter our dataset.

We perform static screening (using only static variables) with the goal of removing duplicates and ensuring we include only common equity firms. We keep only firms marked as major listings. This excludes listings of secondary share classes of a firm. We also keep only listings that are traded on the domestic market. By doing this, we get only one listing per firm. Stocks with the type of instrument other than equity are then filtered out. This filters some of the non-equity listings (bonds, options, etc.); however, this indicator variable is not entirely reliable.

We sort industries, using variable INDN, which provides the name of the industry, into common and uncommon equity and exclude listings which belong to uncommon equity. Examples of filtered-out industries are investment trusts, real estate investment trusts, mutual funds or exchange-traded notes. We search the name of the firm for suspicious word parts to filter out non-common equity further. If the name of the firm contains suspicious words, it is checked manually. For the list of word parts, see Griffin et al. (2010). Some of the words are checked on all firms, and some are country-specific as some of the countries have different ways to mark preferred shares, non-voting shares and others. We exclude a firm if it does not have pricing or fundamentals coverage.

We continue with dynamic screening, which is to eliminate errors in daily and then monthly pricing data. Daily pricing data are padded, meaning that if stock is not traded on a given day, the last available price is reported. We delete observations after the firm is delisted. This is done by trimming observations when the return index in the original currency does not change at the end of the series for each firm. The last observation of the firm is treated as a delisting return because Datastream does not report separate delisting return as CRSP. The order of magnitude of our variables is adjusted so that they are the same as in the CRSP dataset.

We need to preprocess data first on a daily frequency so that they can be used for transaction cost calculations and then create a monthly dataset that will be used in the models. We drop observations with a missing return index. We calculate the daily return from the return index. Return is set to missing in cases when daily returns are higher than 500% or when the price is more than 100,000 dollars. Datastream was rounding prices to the nearest penny before decimalization. This causes nontrivial differences in calculated returns when prices are small. Because of this, we set the return to missing for a price that is less than 0.1 USD. Alternative price screens of 1 USD or 0.5 USD work as well (Ince and Porter, 2006). In cases when the return index is smaller than 0.01, we set the corresponding return to missing, as these cases are heavily affected by rounding. We fix cases when the return is abnormal, but there is a reversal the next day. This is when the daily return is over 200%, but the two-day return is less than 110%.

We divide dividends by a fixed value if the dividend is greater than half the adjusted price. Schmidt et al. (2015) documents that dividend data for some European countries are erroneous. They observe dividends, which are unusually large, about ten times the actual price of the stock. If we used these dividends to calculate returns, we would get unreasonably high returns on the day of the dividend payment. As these dividends are usually a fraction of usual dividends, it is concluded that a decimal error occurred.

Monthly returns are calculated using the return index. For transforming other variables to monthly frequency, either the last available value for a given month is used or the sum over the month in case of volume traded. We compare the return index provided by Datastream with returns that we calculate using price and dividend. If the difference between Datastream returns and returns we constructed is larger than 0.5 in absolute terms, we set returns to missing. We compare the market value reported by Datastream with a self-created market value that we calculate by multiplying the unadjusted price with the number of shares outstanding. If the difference between those two numbers is greater than 0.5 in absolute terms, we set the market value to missing. Monthly returns higher than 2000% are discarded. If R_t or R_{t-1} is higher than 300% and $(1 + R_t)(1 + R_{t-1}) - 1$ is less than 50%, then both returns are set to missing. Monthly returns before the year 2000 are winsorized in each region as a way to limit outliers. Data below the first percentile are set to the first percentile value, and data above 99th percentile are set to 99th percentile value.

C.3 Investment Universe - Liquidity Filter

As our investment universe, we have a sample of 23 developed countries: Australia, Austria, Belgium, Denmark, Finland, France, Germany, Greece, Hong Kong, Ireland, Italy, Japan, Luxembourg, the Netherlands, New Zealand, Norway, Portugal, Singapore, Spain, Sweden, Switzerland, the United Kingdom, and the United States. These countries are sorted into four regions: U.S., Europe, Japan, and Asia Pacific.

We apply a liquidity filter allowing us to avoid micro-caps stocks, which are highly illiquid, and trading would be costly or even impossible (i.e. shorting some firms). We sort firms based on market capitalization and then exclude a portion of low-market capitalization firms each month. In each region, we exclude the least capitalized firms so that the sum of the market capitalization of those firms is 5% of the total market capitalization for that region.

We also employ a similar filter that is based on trading volume over the last 12 months. We exclude low-traded firms so that the sum of their trading volume makes 5% of the total traded volume of the given region. In case trading volume is missing for a firm, we exclude this firm if it belongs to the lowest 10% based on market capitalization.

For stocks that are not in the U.S., we also require that they have a market capitalization larger than the lowest decile NYSE market cap for a given month. This filtering is to ensure that non-U.S. firms have capitalization comparable to U.S. stocks.

Additionally, the firms need to have a price larger than one dollar, in the case of Asia Pacific region $0.1, at the end of the previous month.

D. Robustness Checks

D.1 Portfolio Performance over Time

A common concern is the decreasing profitability of portfolios over time. Therefore, we present a subsample analysis for the long-short decile portfolios. A possible split point could be the decimalization of the U.S. stock market in 2001 (Avramov et al., 2023; Cakici et al., 2023) or the financial crisis (Leung et al., 2021). Portfolios created using machine learning methods often have a decline in profitability after around 2003-2004 (Tobek and Hronec, 2021; Gu et al., 2020; Blitz et al., 2023). Given this evidence and our own results (see Figure A.1 and A.4), we use the end of 2004 as our split-point.

Table D.1 compares the performance of long-short decile portfolios in two subperiods, 1995-2004 and 2005-2018. Table D.2 provides additional performance metrics. Consistent with the literature, the profitability of decile portfolios is diminished after 2005. The first subperiod has considerably higher transaction costs, which makes this drop in profitability less severe for portfolios with transaction costs included. This decrease is more pronounced for shorter forecasting horizons.

In the U.S., risk-adjusted profitability is decreased only in the case of shorter forecasting horizons. For example, for the 12-month portfolio with transaction costs on the U.S. sample, the Sharpe ratio went from 0.93 to 1.35. International portfolios have lower Sharpe ratios for all horizons in the second period. International portfolios have been outperforming the U.S. portfolios in the first period but have more similar performance after 2005. We also see higher risk-adjusted returns for longer-horizon portfolios after 2005. This pattern wasn’t visible when looking at the whole sample.

However, risk-adjusted profitability is decreased only in the case of shorter forecasting horizons. For longer forecasting horizons (5-24 months in the U.S., 3-24 months internationally), there is no change or, in some cases, even improvement of risk-adjusted return in the second period. For example, for the 12-month portfolio with transaction costs on the U.S. sample, the Sharpe ratio went from 0.93 to 1.35. Overall, we can conclude that longer horizon portfolios can be interesting from the investors’ perspective, as they are profitable even in more recent years.

Table D.1 Profitability of Long-Short Decile Portfolios over Time

	Without transaction costs				*With transaction costs*			
	Mean	***Std***	***Sharpe***	***MDD***	***Mean***	***Std***	***Sharpe***	***MDD***
Panel A: U.S. 1995-2004								
1	3.08	6.94	1.54	-29.39	1.87	6.93	0.93	-36.46
2	2.16	5.46	1.37	-32.93	1.59	5.47	1.01	-36.38
3	1.73	4.64	1.29	-24.58	1.34	4.65	1.00	-27.02
4	1.52	4.24	1.24	-15.49	1.22	4.25	0.99	-19.64
5	1.11	3.92	0.98	-39.01	0.86	3.93	0.76	-41.73
6	1.22	3.54	1.20	-26.08	1.02	3.55	0.99	-29.00
9	0.90	3.01	1.04	-27.16	0.76	3.02	0.87	-28.99
12	0.87	2.81	1.07	-24.75	0.76	2.81	0.93	-26.06
24	1.13	3.36	1.16	-24.86	1.07	3.37	1.10	-25.61
Panel B: U.S. 2005-2018								
1	0.75	3.04	0.85	-30.27	0.65	3.04	0.74	-30.96
2	0.70	2.82	0.86	-21.69	0.65	2.82	0.80	-21.99
3	0.65	2.79	0.80	-25.71	0.61	2.79	0.76	-25.90
4	0.71	2.49	1.00	-17.72	0.69	2.49	0.96	-17.88
5	0.71	2.27	1.09	-13.75	0.69	2.27	1.06	-13.85
6	0.63	2.21	0.99	-16.24	0.61	2.21	0.96	-16.70
9	0.66	1.89	1.21	-14.15	0.64	1.89	1.18	-14.43
12	0.62	1.55	1.38	-8.59	0.61	1.55	1.35	-8.75
24	0.58	1.38	1.45	-14.54	0.57	1.38	1.44	-14.68
Panel C: International 1995-2004								
1	2.86	3.94	2.52	-9.42	1.69	3.88	1.51	-12.24
2	2.00	4.30	1.61	-23.56	1.43	4.28	1.16	-25.22
3	1.52	4.02	1.31	-18.23	1.13	4.02	0.98	-23.69
4	1.37	3.74	1.27	-25.31	1.07	3.74	0.99	-30.43
5	1.39	3.25	1.49	-23.50	1.14	3.24	1.22	-26.63
6	1.28	3.04	1.45	-26.95	1.06	3.03	1.21	-29.73
9	1.35	2.94	1.59	-23.98	1.20	2.94	1.41	-25.60
12	1.32	3.08	1.48	-27.40	1.19	3.08	1.34	-28.48
24	1.14	3.33	1.19	-34.37	1.07	3.33	1.11	-35.06
Panel D: International 2005-2018								
1	1.03	2.42	1.47	-23.31	0.59	2.42	0.85	-27.11
2	0.84	2.41	1.20	-26.96	0.62	2.41	0.89	-30.15
3	0.74	2.22	1.15	-21.91	0.59	2.22	0.92	-24.67
4	0.63	2.06	1.06	-23.44	0.51	2.07	0.86	-26.08
5	0.64	1.90	1.17	-17.17	0.54	1.90	0.98	-19.08
6	0.59	1.79	1.14	-16.32	0.50	1.80	0.97	-18.12
9	0.60	1.74	1.20	-16.06	0.54	1.74	1.08	-16.53
12	0.57	1.60	1.24	-12.18	0.53	1.60	1.14	-12.45
24	0.52	1.58	1.14	-7.03	0.50	1.58	1.08	-7.49

Notes: The table shows the performance of long-short decile portfolios in the U.S. (Panel A and B) and internationally (Panel C and D) for periods 1995-2004 and 2005-2018. Monthly mean returns, standard deviation, annualized Sharpe ratio and maximum drawdown for strategies labelled 1 to 24 are reported. The label corresponds to the horizon h for which we obtain the predictions and, at the same time, the holding period for a given portfolio. The displayed values are in percentages except for the Sharpe ratio.

Table D.2 Performance metrics of Long-Short Decile Portfolios over Time

	Without transaction costs				*With transaction costs*			
	Sortino	*CVaR 99%*	*Alpha*	*Beta*	*Sortino*	*CvaR 99%*	*Alpha*	*Beta*
Panel A: U.S. 1995-2004								
1	3.14	-12.46	3.25	-0.23	1.72	-13.55	2.05	-0.25
2	2.54	-11.79	2.26	-0.14	1.75	-12.32	1.69	-0.14
3	2.60	-8.40	1.82	-0.13	1.89	-8.84	1.44	-0.13
4	2.37	-8.03	1.57	-0.08	1.81	-8.33	1.27	-0.08
5	1.52	-9.61	1.18	-0.10	1.14	-9.89	0.94	-0.10
6	2.11	-7.14	1.33	-0.15	1.69	-7.35	1.13	-0.15
9	1.60	-6.90	0.95	-0.07	1.31	-7.03	0.81	-0.07
12	1.69	-6.42	0.90	-0.04	1.44	-6.52	0.78	-0.04
24	2.03	-7.03	1.17	-0.05	1.90	-7.11	1.11	-0.05
Panel B: U.S. 2005-2018								
1	1.31	-6.78	0.90	-0.21	1.11	-6.89	0.79	-0.20
2	1.37	-5.86	0.79	-0.13	1.26	-5.91	0.74	-0.13
3	1.21	-6.08	0.73	-0.12	1.14	-6.12	0.70	-0.11
4	1.68	-4.71	0.80	-0.12	1.61	-4.73	0.78	-0.12
5	1.83	-4.47	0.76	-0.07	1.76	-4.50	0.74	-0.07
6	1.59	-4.56	0.68	-0.08	1.54	-4.58	0.66	-0.08
9	2.11	-3.70	0.69	-0.04	2.05	-3.71	0.68	-0.04
12	2.62	-2.84	0.63	-0.01	2.56	-2.85	0.62	-0.01
24	3.00	-2.34	0.57	0.01	2.96	-2.34	0.57	0.01
Panel C: International 1995-2004								
1	7.54	-4.89	2.91	-0.10	3.50	-5.98	1.74	-0.10
2	3.21	-7.91	2.02	-0.03	2.09	-8.56	1.45	-0.03
3	2.53	-7.38	1.53	-0.02	1.75	-7.82	1.14	-0.01
4	2.45	-6.38	1.38	-0.02	1.78	-6.78	1.08	-0.02
5	3.03	-4.75	1.42	-0.06	2.31	-5.06	1.17	-0.06
6	2.87	-4.76	1.31	-0.08	2.24	-5.02	1.09	-0.08
9	3.40	-4.58	1.36	-0.01	2.86	-4.76	1.20	-0.01
12	3.03	-5.21	1.31	0.02	2.65	-5.35	1.18	0.02
24	2.33	-5.48	1.12	0.05	2.15	-5.56	1.05	0.05
Panel D: International 2005-2018								
1	2.45	-4.92	1.10	-0.12	1.28	-5.43	0.67	-0.11
2	1.88	-5.37	0.91	-0.11	1.32	-5.61	0.69	-0.11
3	1.84	-4.81	0.79	-0.09	1.41	-4.98	0.65	-0.09
4	1.78	-4.12	0.65	-0.03	1.37	-4.28	0.53	-0.03
5	2.08	-3.77	0.64	-0.01	1.68	-3.89	0.54	-0.01
6	2.01	-3.44	0.59	0.00	1.66	-3.54	0.50	0.00
9	2.15	-3.36	0.59	0.02	1.88	-3.43	0.53	0.02
12	2.29	-2.95	0.55	0.04	2.04	-3.00	0.50	0.04
24	2.13	-2.77	0.48	0.06	2.00	-2.79	0.45	0.06

Notes: Additional performance measures of long-short decile portfolios in the U.S. (Panel A and B) and internationally (Panel C and D) for periods 1995-2004 and 2005-2018. Sortino ratio, conditional value at risk at 99%, Alpha and Beta (with regards to market returns in the U.S./internationally) are presented. The label corresponds to the horizon h for which we obtain the predictions and, at the same time, the holding period for a given portfolio.

D.2 Gradient Boosted Regression Trees Results

As a robustness check, we use gradient-boosted regression trees[10] instead of feedforward neural networks. Sample splitting is the same as with neural networks. We use a hyperparameter search to select optimal parameters that perform well out-of-sample. For the optimal number of trees, we test 50, 100, 200, 300, 400, and 500. The maximum depth of each tree between one and nine is considered, and learning rates are 0.01, 0.025, 0.05, and 0.1.

We obtain predictions of cumulative returns at various horizons using gradient-boosted regression trees on the U.S. sample. We constructed portfolios in the same way as with neural networks. Results of long-short decile portfolios at various horizons, with a holding period equal to the forecasting horizon used, are presented in Table D.3. Results for the one-month horizon have a comparable Sharpe ratio to neural networks, but it is slightly more volatile. Looking at longer horizons, the two-month portfolio has a higher Sharpe ratio after accounting for transaction costs, benefiting from the reduced turnover of the strategy. The short leg of portfolios is not profitable with transaction costs, similar to neural networks portfolios in the U.S. However, in this case, a long-only component is more profitable and has a higher Sharpe ratio than a long-short strategy. Short only component seems ineffective in this case. More performance metrics for portfolios are in Table D.4. Betas of portfolios are around -0.40, almost double that of neural networks. In Figure D.1 are cumulative returns of long-short decile portfolios. Without transaction costs one-month, then two-month portfolios dominate. When we account for transaction costs, the two-month portfolio is better.

Double-sorted portfolios were made with cutoffs of top 15% and bottom 15%. In Table D.5 is shown the performance of double-sorted long-short portfolios. The portfolio 1-2 has a slightly higher Sharpe ratio than the one-month decile portfolio. Double-sorted portfolios have higher mean returns. Cumulative returns of double-sorted long-short portfolios in comparison with one-month long-short decile portfolio are in Figure D.2. Additional performance metrics are in Table D.6. Betas are more negative than in the case of decile portfolios.

Performance of long-short buy/hold spread of 10%/20% is presented in Table D.7. Portfolio 2-1 has the highest Sharpe ratio and mean, higher than the one-month decile portfolio. Additional metrics for these portfolios are in Table D.8. Cumulative returns of buy/hold spread portfolios are in Figure D.3. Portfolio 2-1 outperforms the benchmark model (one-month long-short decile portfolio).

[10] See subsection B.1 for more details.

Table D.3 Performance of Long-Short Decile Portfolios in the U.S.

	Without transaction costs				*With transaction costs*				
	Mean	***Std***	***Sharpe***	***MDD***	***Mean***	***Std***	***Sharpe***	***MDD***	***Turnover***
				Panel A: Long-short portfolio					
1	1.84	5.59	1.14	-47.63	1.23	5.48	0.78	-49.68	121.56
2	1.75	5.67	1.07	-41.85	1.44	5.63	0.89	-42.98	59.46
3	1.14	5.19	0.76	-39.71	0.93	5.17	0.62	-40.62	40.77
4	1.15	5.44	0.73	-38.67	0.98	5.43	0.63	-43.32	32.35
5	1.01	5.34	0.65	-46.94	0.87	5.32	0.57	-50.32	26.84
6	0.90	4.99	0.63	-52.73	0.79	4.98	0.55	-55.25	23.13
9	0.98	4.09	0.83	-45.34	0.90	4.07	0.76	-47.43	16.69
12	0.98	4.36	0.78	-63.57	0.92	4.36	0.73	-64.64	12.62
24	0.86	4.74	0.63	-67.57	0.83	4.75	0.61	-68.14	6.14
				Panel B: Long only component of the strategy					
1	1.78	6.13	1.00	-54.38	1.48	6.10	0.84	-54.91	121.22
2	1.68	6.08	0.95	-52.37	1.53	6.06	0.87	-52.69	59.12
3	1.32	5.53	0.83	-52.41	1.22	5.52	0.76	-52.84	40.30
4	1.26	5.28	0.83	-53.69	1.18	5.27	0.78	-53.97	32.00
5	1.21	5.56	0.75	-59.39	1.14	5.56	0.71	-59.59	26.52
6	1.18	5.38	0.76	-54.92	1.13	5.38	0.73	-55.12	22.60
9	1.22	5.61	0.76	-55.82	1.19	5.60	0.73	-55.90	16.18
12	1.23	5.69	0.75	-54.12	1.20	5.69	0.73	-54.23	12.02
24	1.20	5.47	0.76	-50.79	1.18	5.47	0.75	-50.85	5.84
				Panel C: Short only component of the strategy					
1	0.07	7.98	0.03	-84.96	-0.28	7.92	-0.12	-85.96	121.81
2	0.06	8.04	0.03	-82.67	-0.12	8.02	-0.05	-83.31	59.61
3	-0.20	8.25	-0.08	-83.78	-0.32	8.24	-0.14	-84.71	41.12
4	-0.13	8.39	-0.05	-82.93	-0.22	8.38	-0.09	-83.74	32.58
5	-0.21	8.38	-0.09	-83.68	-0.29	8.37	-0.12	-84.31	27.06
6	-0.30	8.27	-0.12	-85.54	-0.36	8.27	-0.15	-86.02	23.55
9	-0.24	8.14	-0.10	-82.30	-0.29	8.14	-0.13	-82.72	17.13
12	-0.27	7.99	-0.12	-80.40	-0.31	7.99	-0.13	-82.28	13.18
24	-0.39	7.52	-0.18	-85.69	-0.41	7.52	-0.19	-86.36	6.47

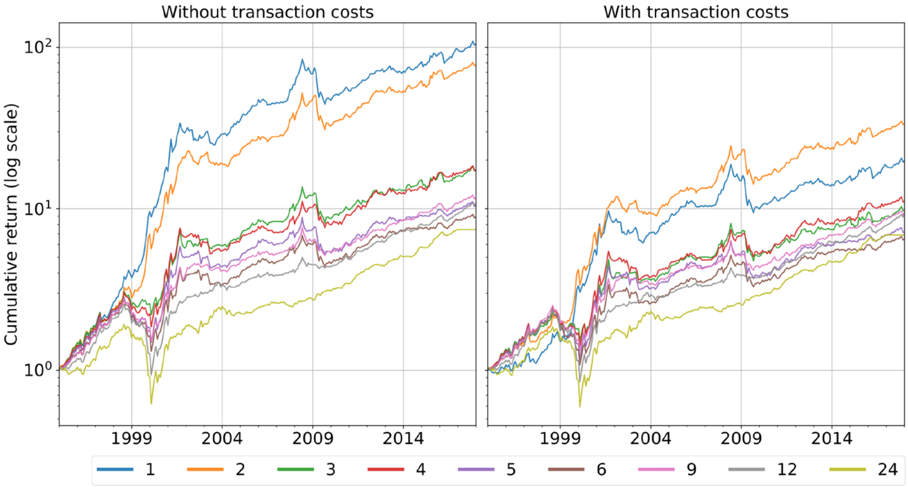
Notes: The table shows the performance of long-short decile portfolios in the U.S. for the period between 1995 and 2018. Monthly mean returns, standard deviation, annualized Sharpe ratio and maximum drawdown for strategies labelled 1 to 24 are reported. The label corresponds to the horizon h for which we obtain the predictions and, at the same time, the holding period for a given portfolio. In Panel A are the results of the long-short portfolio. The results are decomposed into long and short components in Panel B and Panel C. The displayed values are in percentages except for the Sharpe ratio.

Table D.4 Performance Measures of Long-Short Decile Portfolios in the U.S.

	Without transaction costs				*With transaction costs*			
	Sortino	***CVaR 99%***	***Alpha***	***Beta***	***Sortino***	***CVaR 99%***	***Alpha***	***Beta***
1	2.19	-10.53	2.12	-0.37	1.35	-11.10	1.50	-0.37
2	1.93	-11.71	2.04	-0.39	1.53	-12.13	1.73	-0.39
3	1.27	-11.55	1.48	-0.47	1.00	-11.82	1.27	-0.47
4	1.27	-11.53	1.54	-0.54	1.06	-11.74	1.38	-0.54
5	1.08	-12.10	1.35	-0.47	0.92	-12.29	1.22	-0.47
6	0.98	-12.04	1.25	-0.47	0.84	-12.22	1.13	-0.47
9	1.35	-9.02	1.23	-0.34	1.22	-9.17	1.15	-0.34
12	1.18	-10.01	1.18	-0.28	1.09	-10.14	1.12	-0.28
24	0.93	-10.95	0.98	-0.17	0.90	-11.03	0.95	-0.17

Notes: Sortino ratio, conditional value at risk at 99%, Alpha and Beta (in comparison with U.S. market returns) are reported for long-short decile portfolios for the period between 1995 and 2018. The portfolio label is the forecasting horizon and the holding period for the portfolio.

Figure D.1 Cumulative Returns of Long-Short Decile Portfolios in the U.S.



Notes: The figure shows cumulative returns of long-short decile portfolios without and with transaction costs on the U.S. sample. The portfolio label is the forecasting horizon in months and the holding period of the strategy.

Table D.5 Double-Sorted Portfolios Performance in the U.S.

	Without transaction costs				*With transaction costs*				
	Mean	***Std***	***Sharpe***	***MDD***	***Mean***	***Std***	***Sharpe***	***MDD***	***Turnover***
1 - 2	2.05	6.17	1.15	-42.60	1.46	6.05	0.84	-43.44	116.60
1 - 3	1.93	6.11	1.10	-48.97	1.34	6.00	0.78	-50.85	114.60
1 - 6	1.72	6.35	0.94	-56.98	1.13	6.24	0.63	-58.65	114.89
1 - 9	1.92	6.40	1.04	-55.77	1.34	6.28	0.74	-57.51	116.45
1 - 12	2.04	6.04	1.17	-40.56	1.44	5.93	0.84	-42.87	118.54
1 - 24	1.97	6.47	1.06	-53.40	1.35	6.41	0.73	-63.92	122.07

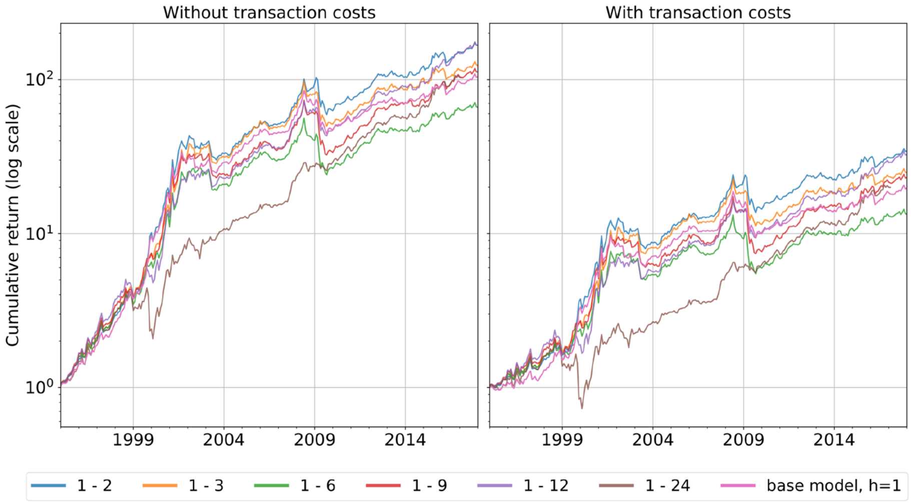
Notes: The table shows the profitability of a double-sorted long-short portfolio in the U.S. between 1995 and 2018. Portfolio labels (1-2 to 1-24) show which two horizon predictions were used in double sorting. Results are shown with and without transaction costs. Monthly mean returns, standard deviation, annualized Sharpe ratio, and maximum drawdown are reported. Reported values are in percentages, with the exception of the Sharpe ratio.

Table D.6 Double-Sorted Portfolios Performance Metrics - in the U.S.

	Without transaction costs				*With transaction costs*			
	Sortino	***CVaR 99%***	***Alpha***	***Beta***	***Sortino***	***CVaR 99%***	***Alpha***	***Beta***
1 - 2	2.23	-11.99	2.38	-0.46	1.49	-12.62	1.79	-0.45
1 - 3	2.03	-12.43	2.31	-0.52	1.32	-13.05	1.72	-0.52
1 - 6	1.62	-13.65	2.14	-0.57	1.00	-14.30	1.55	-0.57
1 - 9	1.83	-13.51	2.35	-0.58	1.20	-14.18	1.76	-0.58
1 - 12	2.19	-11.91	2.43	-0.53	1.42	-12.76	1.82	-0.53
1 - 24	1.75	-13.20	2.26	-0.41	1.10	-14.40	1.63	-0.41

Notes: Sortino ratio, conditional value at risk at 99%, Alpha and Beta (with regards to market returns in the U.S.) are presented for long-short double-sorting portfolios for the period between 1995 and 2018. Portfolio labels are the two forecasting horizons which were used in double sorting.

Figure D.2 Cumulative Returns of Long-Short Double Sorting Portfolios in the U.S.



Notes: The figure shows cumulative returns on a logarithmic scale of the double-sorting strategy and of the long-short decile portfolio at horizon one in the U.S. The two numbers correspond to horizons on which we double-sorted. The holding period is one month.

Table D.7 Buy/Hold Spread Portfolio Performance in the U.S.

		Without transaction costs				*With transaction costs*				
buy	***hold***	***Mean***	***Std***	***Sharpe***	***MDD***	***Mean***	***Std***	***Sharpe***	***MDD***	***Turnover***
1	1	1.69	5.49	1.07	-44.34	1.24	5.43	0.79	-45.85	84.51
2	1	1.84	5.85	1.09	-42.53	1.43	5.77	0.86	-43.95	71.74
3	1	1.53	5.88	0.90	-45.65	1.15	5.82	0.69	-46.88	63.39
4	1	1.54	6.06	0.88	-46.21	1.18	6.01	0.68	-47.36	61.13
5	1	1.47	5.95	0.86	-44.82	1.12	5.90	0.66	-46.00	57.91
6	1	1.31	5.55	0.82	-43.94	0.97	5.50	0.61	-45.27	55.23
9	1	1.33	5.49	0.84	-41.09	1.02	5.45	0.65	-42.56	51.59
12	1	1.42	5.41	0.91	-46.73	1.13	5.36	0.73	-53.80	50.97
24	1	1.36	5.81	0.81	-57.34	1.07	5.80	0.64	-63.66	48.51

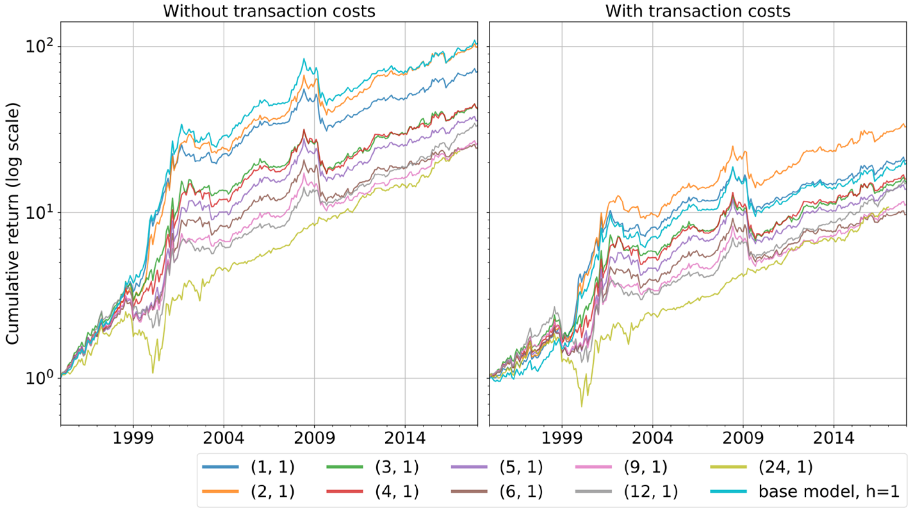
Notes: The profitability of long-short buy/hold spread portfolios in the U.S. for the 1995 to 2018 period. We use a buy/hold spread 10%/20% and report the results both without transaction costs and with transaction costs. Buy and hold columns show which horizons were used in the portfolio creation. Monthly mean returns, standard deviation, annualized Sharpe ratio, and maximum drawdown are reported. All values are reported in percentages except for the Sharpe ratio.

Table D.8 Buy/Hold Spread Portfolio Performance Metrics - U.S. sample

		Without transaction costs				*With transaction costs*			
buy	*hold*	*Sortino*	*CVaR 99%*	*Alpha*	*Beta*	*Sortino*	*CVaR 99%*	*Alpha*	*Beta*
1	*1*	2.07	-10.31	1.93	-0.32	1.42	-10.81	1.48	-0.32
2	*1*	2.10	-11.13	2.13	-0.40	1.55	-11.57	1.72	-0.40
3	*1*	1.58	-11.92	1.88	-0.48	1.14	-12.47	1.51	-0.48
4	*1*	1.54	-12.69	1.97	-0.58	1.13	-13.23	1.60	-0.58
5	*1*	1.48	-12.62	1.88	-0.56	1.08	-13.20	1.53	-0.55
6	*1*	1.39	-11.74	1.71	-0.54	0.99	-12.20	1.37	-0.54
9	*1*	1.44	-11.54	1.71	-0.52	1.06	-11.93	1.40	-0.51
12	*1*	1.61	-11.30	1.77	-0.48	1.22	-11.79	1.47	-0.47
24	*1*	1.40	-12.24	1.57	-0.31	1.04	-12.85	1.28	-0.31

Notes: Sortino ratio, conditional value at risk at 99%, Alpha and Beta (in comparison with U.S. market returns) are reported for long-short buy and holds spread for the period between 1995 and 2018. We use 10%/20% buy/hold spread cutoffs. The portfolio label signifies the horizon based on which we buy and the horizon based on which we hold stocks, respectively.

Figure D.3 Cumulative Returns of Buy/Hold Spread Portfolios in the U.S.



Notes: Cumulative returns of long-short buy/hold spread portfolios compared with the benchmark model. We use buy/hold spread 10%/20%. The portfolio label signifies the horizon based on which we buy and hold stocks, respectively.

REFERENCES

Avramov D, Si C, Metzker L (2023): Machine Learning vs. Economic Restrictions: Evidence from Stock Return Predictability. *Management Science* 69 (5):2587–2619.

Avramov D, Kaplanski G, Subrahmanyam A (2020): Post-Fundamentals Drift in Stock Prices: A Machine Learning Approach. *Available at SSRN 3507512.*

Baba Yara, Fahiz. 2020. "Machine Learning and Return Predictability Across Firms, Time and Portfolios." *Available at SSRN 3696533.*

Bali TG, Engle RF, Murray S (2016): *Empirical Asset Pricing: The Cross Section of Stock Returns.* Chap. 7, 103–121. John Wiley & Sons.

Bhandari LC (1988): Debt/Equity Ratio and Expected Common Stock Returns: Empirical Evidence. *The journal of finance* 43 (2): 507–528.

Blitz D, Hanauer MX, Hoogteijling T, Howard C (2023): The Term Structure of Machine Learning Alpha. *Available at SSRN.*

Bryzgalova S, Pelger M, Zhu J (2020): Forest through the Trees: Building Cross-Sections of Stock Returns. *Available at SSRN 3493458.*

Cakici N, Fieberg C, Metko D, Zaremba A (2023): Predicting Returns with Machine Learning Across Horizons, Firms Size, and Time. *Journal of Financial Data Science, Forthcoming.*

Chen L, Pelger M, Zhu J (2020): Deep Learning in Asset Pricing. *Available at SSRN 3350138.*

Chen T, Guestrin C (2016): Xgboost: A Scalable Tree Boosting System. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,* 785–794.

Chung KH, Zhang H (2014): A Simple Approximation of Intraday Spreads Using Daily Data. *Journal of Financial Markets* 17:94–120.

DeMiguel V, Martin-Utrera A, Nogales FJ, Uppal R (2020): A Transaction-Cost Perspective on the Multitude of Firm Characteristics. *The Review of Financial Studies* 33(5):2180–2222.

Fama E, French K (1992a): The Cross-Section of Expected Stock Returns. *Journal of Finance* 47 (2): 427–465.

Fama EF, French KR (1992b): The Cross-Section of Expected Stock Returns. *The Journal of Finance* 47 (2): 427–465.

Follmer H, Schied A (2011): *Stochastic Finance: An Introduction in Discrete Time.* Walter de Gruyter.

Fong KYL, Holden CW, Tobek O (2018): Are Volatility Over Volume Liquidity Proxies Useful for Global or US Research? *Kelley School of Business Research Paper,* nos. 17-49.

Frankel R, Lee CMC (1998): Accounting Valuation, Market Expectation, and Cross-Sectional Stock Returns. *Journal of Accounting and economics* 25(3):283– 319.

Frazzini A, Israel R, Moskowitz TJ (2012): Trading Costs of Asset Pricing Anomalies. *Fama-Miller working paper,* 14–05.

French KR (2020): Kenneth R. French - Data Library. *Tuck-MBA program web server. http://mba. tuck. dartmouth. edu/pages/faculty/ken. french/data library. html (accessed March 14, 2020).*

Freyberger J, Neuhierl A, Weber M (2017): *Dissecting Characteristics Nonparametrically.* Technical Report. National Bureau of Economic Research.

Friedman JH (2001): "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of statistics,* 1189–1232.

Giglio S, Xiu D (2019): Asset Pricing with Omitted Factors. *Chicago Booth Research Paper,* nos. 16-21.

Goodfellow I, Bengio Y, Courville A (2016): *Deep Learning.* MIT press.

Green J, Hand JRM, Zhang XF (2013): The Supraview of Return Predictive Signals. *Review of Accounting Studies* 18(3):692–730.

Griffin JM, Kelly PJ, Nardari F (2010): Do Market Efficiency Measures Yield Correct Inferences? A Comparison of Developed and Emerging Markets.” *The Review of Financial Studies* 23(8):3225–3277.

Gu S, Kelly B, Xiu D (2020): Empirical Asset Pricing Via Machine Learning. *The Review of Financial Studies* 33 (5): 2223–2273.

Hastie T, Tibshirani R, Friedman J (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Science & Business Media.

Hawkins EH, Chamberlin SC, Daniel WE (1984): Earnings Expectations and Security Prices. *Financial Analysts Journal* 40(5):24–38.

Heston SL, Sadka R (2008): Seasonality in the Cross-Section of Stock Returns. *Journal of Financial Economics* 87(2):418–445.

Hoffstein C, Faber N, Braun S (2020): Rebalance Timing Luck: The (Dumb) Luck of Smart Beta. *Available at SSRN 3673910.*

Hou K, Xue C, Zhang L (2020): Replicating Anomalies. *The Review of Financial Studies* 33(5):2019–2133.

Ince OS, Porter RB (2006): Individual Equity Return Data from Thomson Datastream: Handle with Care! *Journal of Financial Research* 29(4):463–479.

Ioffe S, Szegedy C (2015): Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv preprint arXiv:1502.03167.*

Jegadeesh N (1990): Evidence of Predictable Behavior of Security Returns. *The Journal of Finance* 45(3):881–898.

Jegadeesh N, Titman S (1993): Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *The Journal of finance* 48(1):65–91.

Kelly BT, Pruitt S, Su Y (2019): Characteristics are Covariances: A Unified Model of Risk and Return. *Journal of Financial Economics* 134(3):501–524.

Kingma DP, Ba J (2014): Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980.*

Kozak S, Nagel S, Santosh S (2020): Shrinking the Cross-Section. *Journal of Financial Economics* 135 (2): 271–292.

La Porta R (1996): Expectations and the Cross-Section of Stock Returns. *The Journal of Finance* 51(5):1715–1742.

Leung E, Lohre H, Mischlich D, Shea Y, Stroh M (2021): The Promises and Pitfalls of Machine Learning for Predicting Stock Returns. *The Journal of Financial Data Science.*

Lewellen J (2015): The Cross-section of Expected Stock Returns.” *Critical Finance Review* 4(1):1–44.

Moskowitz TJ, Grinblatt M (1999): Do Industries Explain Momentum? *The Journal of Finance* 54(4):1249–1290.

Novy-Marx R, Velikov M (2019): Comparing Cost-Mitigation Techniques. *Financial Analysts Journal* 75(1):85–102.

Ortiz-Molina H, Phillips GM (2014): Real Asset Illiquidity and the Cost of Capital. *Journal of Financial and Quantitative Analysis* 49(1):1–32.

Rockafellar RT, Uryasev S et al. (2000): Optimization of Conditional Valueat-Risk. *Journal of risk* 2:21–42.

Schmidt PS, Von Arx U, Schrimpf A, Wagner AF, Ziegler A (2015): On the Construction of Common Size, Value and Momentum Factors in International Stock Markets: A Guide with Applications. *CCRS Working Paper Series,* nos. 01/11.

Scholz H (2007): Refinements to the Sharpe Ratio: Comparing Alternatives for Bear Markets. *Journal of Asset Management* 7:347–357.

Sharpe WF (1963): A Simplified Model for Portfolio Analysis. *Management Science* 9(2):277–293.

Sharpe WF (1966): Mutual Fund Performance. *The Journal of Business* 39(1):119–138.

Sharpe WF (1994): The Sharpe Ratio. *Journal of Portfolio Management* 21 (1): 49–58.

Sloan RG. (1996): Do Stock Prices Fully Reflect Information in Accruals and Cash Flows about Future Earnings? *Accounting Review,* 289–315.

Sortino FA, Price LN (1994): Performance Measurement in a Downside Risk Framework." *The Journal of Investing* 3(3):59–64.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014): Dropout: A Simple Way to Prevent Neural Networks from Overfitting." *The journal of machine learning research* 15(1):1929–1958.

Titman S, Wei KCJ, Xie F (2004): Capital Investments and Stock Returns." *Journal of Financial and Quantitative Analysis* 39(4):677–700.

Tobek O, Hronec M (2021): Does it Pay to Follow Anomalies Research? Machine Learning Approach With International Evidence. *Journal of Financial Markets* 56:100588.