ECONSTOR
Make Your Publications Visible.

A Service of
ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Nechvatalova, Lenka

Working Paper

Multi-Horizon Equity Returns Predictability via Machine Learning

IES Working Paper, No. 2/2021

Provided in Cooperation with:
Charles University, Institute of Economic Studies (IES)

Suggested Citation: Nechvatalova, Lenka (2021) : Multi-Horizon Equity Returns Predictability via Machine Learning, IES Working Paper, No. 2/2021, Charles University in Prague, Institute of Economic Studies (IES), Prague

This Version is available at:
https://hdl.handle.net/10419/247369

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

INSTITUTE
OF ECONOMIC STUDIES
Faculty of Social Sciences
Charles University

MULTI-HORIZON EQUITY RETURNS PREDICTABILITY VIA MACHINE LEARNING

Lenka Nechvatalova

IES Working Paper 2/2021

Institute of Economic Studies,
Faculty of Social Sciences,
Charles University in Prague

[UK FSV – IES]

Opletalova 26
CZ-110 00, Prague
E-mail : ies@fsv.cuni.cz
http://ies.fsv.cuni.cz

Institut ekonomických studií
Fakulta sociálních věd
Univerzita Karlova v Praze

Opletalova 26
110 00 Praha 1

E-mail : ies@fsv.cuni.cz
http://ies.fsv.cuni.cz

Disclaimer: The IES Working Papers is an online paper series for works by the faculty and students of the Institute of Economic Studies, Faculty of Social Sciences, Charles University in Prague, Czech Republic. The papers are peer reviewed. The views expressed in documents served by this site do not reflect the views of the IES or any other Charles University Department. They are the sole property of the respective authors. Additional info at: ies@fsv.cuni.cz

Copyright Notice: Although all documents published by the IES are provided without charge, they are licensed for personal, academic or educational use. All rights are reserved by the authors.

Citations: All references to documents served by this site must be appropriately cited.

Bibliographic information:
Nechvatalova L. (2021): "Multi-Horizon Equity Returns Predictability via Machine Learning" IES Working Papers 2/2021. IES FSV. Charles University.

This paper can be downloaded at: http://ies.fsv.cuni.cz

Multi-Horizon Equity Returns Predictability via Machine Learning

Lenka Nechvatalova[a]

[a] Institute of Economic Studies, Charles University
Institute of Information Theory and Automation, Czech Academy of Sciences
Prague, Czech Republic

February 2021

Abstract:
We examine the predictability of expected stock returns across horizons using machine learning. We use neural networks, and gradient boosted regression trees on the U.S. and international equity datasets. We find that predictability of returns using neural networks models decreases with longer forecasting horizon. We also document the profitability of long-short portfolios, which were created using predictions of cumulative returns at various horizons, before and after accounting for transaction costs. There is a trade-off between higher transaction costs connected to frequent rebalancing and greater returns on shorter horizons. However, we show that increasing the forecasting horizon while matching the rebalancing period increases risk-adjusted returns after transaction cost for the U.S. We combine predictions of expected returns at multiple horizons using double-sorting and buy/hold spread, a turnover reducing strategy. Using double sorts significantly increases profitability on the U.S. sample. Buy/hold spread portfolios have better risk-adjusted profitability in the U.S.

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

Acknowledgements: Lenka Nechvátalová gratefully acknowledges support from the Czech Science Foundation under the 19-28231X (EXPRO) project.

1 Introduction

The central question of asset pricing is why do different assets have different expected returns. There are various explanations, but the most conceptual is the risk-based explanation. Investors get a different amount of compensation for bearing different amount of risk. Expected returns predictions, or compensation for the undertaken risk, should take into account all available information, but it is not clear what information is truly relevant and what functional form should resulting predictability take. Over the last forty years, the risk-based explanation failed to find convincing empirical support as hundreds of anomalies, or patterns in stock returns not identified by the prevailing asset pricing models, were identified. Hundreds of predictors identified in the academic literature alone, however, does not make it a *big data* problem in a classical sense. Successful applications of machine learning in other fields like natural language processing, image recognition or gaming are based on two aspects which simply do not hold in finance — an abundance of data and high "signal-to-noise" ratios. When forecasting returns, we do not have an abundance of data at our disposal. It is true that we have plenty of potentially predictive variables; however, we are significantly constrained by the number of observations we can learn from. We do not have many observations of the dependent variable or label. It is the number of observations that limits how complex models we can train. To make things even worse, we are also in the case of a low signal-to-noise ratio, which captures the level of predictability in the domain of interest. Its low levels in financial markets is a direct consequence of investors incentives. Any time there would be an arbitrage, investors should exploit it immediately and in a process erase any level of predictability not associated with the risk compensation.

So how does a successful application of machine learning in finance look like? Gu et al. (2020) show that combining anomalies via machine learning-based predictive regression into one signal achieves unprecedented out-of-sample expected returns predictability[1]. 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. Similar approaches are also employed by 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)[2].

However, most of the empirical results in the asset pricing literature, including the recent work using machine learning, are based on one-month forecasting horizon[3]. Investors

1. This predictability is typically based on the individual firm characteristics, and it is common to use 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, 1992). For the comprehensive list of anomalies documented in the literature, please see the large replication study of Hou et al. (2020).

2. Applications of machine learning in finance was successful also in other areas than asset pricing. Khandani et al. (2010) apply machine learning to construct models of consumer credit risk that improve classification rates of credit card delinquencies and defaults. Butaru et al. (2016) model consumer credit risk finding differences between drivers of delinquency across different banks. Sirignano et al. (2018) develop a deep learning model of mortgage risk using a large number of loan-specific as well as macroeconomic variables finding strongly nonlinear relationships. Heaton et al. (2017) apply deep learning hierarchical decision models to portfolio selection.

3. Studies, employing machine learning on the cross-section to predict expected returns, which do not consider longer horizons than one-month include: Bryzgalova et al. (2020), Tobek and Hronec (2020), Chen et al. (2020), Feng et al. (2020) Kozak et al. (2020), or Messmer (2017).

are not horizon agnostic, though. Only an investor with a logarithmic utility function would allocate his portfolio the same way for the 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. There are three main contributions. Firstly, we document that predictability of returns decreases with longer forecasting horizon. Secondly, we look at the long-short portfolios' profitability based on the predictive regressions across horizons after accounting for transaction costs. There is a trade-off between greater profitability on shorter horizons and higher transaction costs resulting from more frequent rebalancing. Without transaction costs, the risk-adjusted profitability is approximately the same for all horizons in the U.S., and declines internationally. After accounting for transactions costs, the risk-adjusted profitability increases with longer horizons in the U.S. It remains roughly the same for all horizons internationally. Thirdly, we combine predictions for multiple horizons and find that correctly combining them using double sorts significantly improves profitability on the U.S. universe.

We examine expected multi-horizon stock returns around the globe. Using 153 anomalies from Tobek and Hronec (2020) as variables in predictive regressions for stock returns, we inspect the predictability of stock returns from one month to two years ahead and find decreasing predictability with longer horizons. This result matches the conclusion of Lewellen (2015), who uses a smaller number of variables and the least-squares approach on one month, six months and one-year horizons. On the other hand, it differs from the conclusion of Gu et al. (2020) who, though being closer to this study in methodology, document increasing predictability on a longer horizon (one-year) compared to a shorter horizon (one-month). This could be a result of different anomalies being used or difference in the investment universe, where our universe is significantly more liquid. The role of horizons is also studied by Kamara et al. (2015) who focus on the pricing of risk factors on different forecasting horizons. Their results suggest that risk at longer horizons is more relevant for persistent factors. Avramov et al. (2020) use a machine learning-based approach to create a fundamental based measure which is measuring the distance between fundamentals and prior moving averages. They show that their measure is significant not only on the one-month horizon but also on longer horizons (with profitability decreasing at longer horizons).

Next, we construct the long-short portfolios using cumulative returns predictions at different horizons. Gu et al. (2020) and Tobek and Hronec (2020) focus on the monthly horizon only and document strong predictability in the cross-section of returns by using a number of machine learning methods. We replicate their results on the liquid universe of global stocks and use them as our benchmark. Avramov et al. (2021) show, that profitability of machine learning-based strategies using stock characteristics as predictors, significantly weakens after considering transaction costs. Other papers examining transaction costs when combining multiple anomalies are for example Frazzini et al. (2012) and DeMiguel et al. (2020)[4].

One of the remedies of high transaction costs is rebalancing less often, and adjusting the forecasting horizon to match the rebalancing frequency. The mean returns decrease

4. Papers focusing on transactions costs for individual anomalies include Korajczyk and Sadka (2004), Novy-Marx and Velikov (2019) and Chen and Velikov (2017).

with the horizon even when we take the transaction costs into account. However, longer horizons offer better Sharpe ratios in the U.S. Internationally, longer horizons offer a less-risky alternative compared to the one-month horizon.

Instead of relying on the forecast for one specific horizon, it is possible to use expectations over different horizons to potentially achieve higher out-of-sample risk-adjusted returns. We combine predictions for two different horizons via double sorting. We independently sort stocks based on predicted cumulative returns from two different horizons. Then we go long stocks which are in the top 15% on both horizons and go short stocks which are in the bottom 15% on both horizons. In the U.S., this leads to large performance gains over our benchmark, i.e. portfolios based on the one-month ahead forecasts. Internationally, this also leads to higher 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 buy stocks based on predictions on a certain horizon and hold stocks based on predictions on a one-month horizon. Buy/hold spread portfolios in the U.S. have better risk-adjusted profitability when we buy based on longer horizons and hold stocks based on one-month horizon. 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 multihorizon 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

For the United States equity data, we use CRSP/Compustat Merged Database from the Center for Research in Security Prices. For international equity data, we use Datastream from Refinitiv. We also use the U.S. consumer price index to estimate transaction costs and three month U.S. T-bill rate, which will be used for anomaly calculation, both from Datastream. We also use Market minus risk-free rate for the U.S. and developed markets from the data library provided by French (2020).

The dataset is filtered and preprocessed, to fix known errors in the databases and exclude non-equity firms. We also need to restrict the investment universe in order to avoid thinly traded stocks. This way, we mitigate, to some extent, the effect of market microstructure noise in our results. The preprocessing of the dataset and the liquidity filters are described in more detail in Appendix C.

We calculate 153 anomalies that were published in the academic literature. We follow the list of anomalies and their construction from Tobek and Hronec (2020). All of the anomalies are firm-specific with monthly frequency. The anomalies are cross-sectionally ranked with respect to firm's region, and missing observations are imputed with median value. This is done to avoid problems with outliers and is a common procedure used for example in Gu et al. (2020), 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 estimation of transaction costs can be found in subsection B.3.

2.2 Stochastic discount factor and predictive regressions

This section introduces the theoretical setting, the connection between the predictive regressions for stock returns and stochastic discount factor. According to the law of one price, Equation 1 should hold for any return $R_{t+1,i}$.

$$E_t\left[M_{t+1}R_{i,t+1}\right] = 0 \tag{1}$$

where M_{t+1} is a stochastic discount factor. Similarly, Equation 2 should also hold for any excess return $R^e_{t+1,i} = R_{t+1,i} - R^f_{t+1}$, where R^f_t is a risk-free rate.

$$E_t\left[M_{t+1}R^e_{i,t+1}\right] = 0 \tag{2}$$

Equation 2 is equivalent to the Equation 3, which shows that expected excess return for a generic asset i is a function of the systematic risk exposure $\beta_{t,i}$ and the price of risk λ_t.

$$E_t\left[r_{i,t+1}\right] = \beta'_{i,t}\lambda_t \tag{3}$$

where $\beta_{t,i} = -\frac{\mathrm{Cov}_t\left(R^e_{t+1,i}, M_{t+1}\right)}{\mathrm{Var}_t(M_{t+1})}$ and $\lambda_t = \frac{Var_t(M_{t+1})}{E_t[M_{i,t+1}]}$.

According to this formulation, investors are only being compensated for holding systematic risk and not an idiosyncratic one. Further, the stochastic discount factor can also be represented as transformed tangency portfolio, i.e. portfolio on the mean-variance efficient frontier with the highest Sharpe ratio[5]. In general, a stochastic discount factor can be obtained as a portfolio, which satisfies the fundamental asset pricing Equation 1, or 2 when working with the excess returns. Portfolio weights of stochastic discount factor can be seen in Equation 4.

$$\omega_t = \mathbb{E}_t\left[R^e_{t+1}{R^e_{t+1}}^{\top}\right]^{-1}\mathbb{E}_t\left[R^e_{t+1}\right] \tag{4}$$

These optimal weights represent one of the mean-variance efficient portfolios and can be obtained by combining Equation 5 with Equation 3.

$$M_{t+1} = 1 - \omega_t^T R^e_{t+1} \tag{5}$$

In other words, we have a portfolio representation of stochastic discount factor or traded factor $F_{t+1} = \omega_t^\top R^e_{t+1}$. Equation 3 can be rewritten as the so-called beta representation shown in Equation 6

$$\mathbb{E}_t\left[R^e_{t+1,i}\right] = \frac{\mathrm{Cov}_t\left(R^e_{t+1,i}, F_{t+1}\right)}{\mathrm{Var}_t\left(F_{t+1}\right)} \cdot \mathbb{E}_t\left[F_{t+1}\right] = \beta_{t,i}\mathbb{E}_t\left[F_{t+1}\right] \tag{6}$$

5. For more details, see Cochrane (2009) or Campbell (2017).

which forms a basis for the one-factor model in Equation 7.

$$R^e_{t+1,i} = \beta_{t,i} F_{t+1} + \epsilon_{t+1,i} \tag{7}$$

Predictive regressions for stock returns such as Lewellen (2015) or Gu et al. (2020) are concerned with the conditional mean estimation as in Equation 8.

$$R_{i,t+1} = \mathrm{E}_t \left(R_{i,t+1} \right) + \epsilon_{i,t+1} \tag{8}$$

$$\mathrm{E}_t \left(R_{i,t+1} \right) = g^{ML}(Z_{i,t}) \tag{9}$$

where stocks are indexed as $i = 1, ..., N_t$, months by $t = 1, ..., T$, $Z_{i,t}$ are stock characteristics or predictive signals and g^{ML} 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^{ML} is a simple linear combination of ordinary least squares. Directly addressing shortcoming of the ordinary least squares approach, Gu et al. (2020) use variety of machine learning methods, such as Elastic net, Random Forests, Gradient Boosted Trees and Neural Networks, to represent the function g^{ML}. Machine learning methods aim to explicitly allow non-linearity, interaction of predictive variables, regularization.

In this study, we focus on the conditional mean estimation, as in equations 8 and 9, and our predictive regression setting is closest to the Gu et al. (2020). There are two main differences. First, we focus on neural networks only as opposed to conducting the horse-race of multiple machine learning models. It is already a well-documented fact, that neural networks are the most powerful tools for explaining the cross-section of stock returns, see Gu et al. (2020), Tobek and Hronec (2020) and Chen et al. (2020). We further include gradient boosted trees as a form of robustness for our results. Second difference and also the one being our main contribution is extending the forecasting horizon from one month[6] to multiple horizons.

We also explicitly allow heterogeneity in predictability across horizons as can be seen in Equation 10.

$$\mathrm{E}_t \left(R^e_{i,t+h} \right) = g^{ML}_h (R^e_{i,t+h}) + \epsilon_{i,t+h} \tag{10}$$

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

Relationship between estimating the conditional mean of stock returns and stochastic discount factor also holds in the multi-period setting. Extending the stochastic discount factor from the one-period setting to multiple horizons is straightforward. The multi-period stochastic discount factor is simply the product of single-period stochastic discount

6. Majority of results documented by the empirical asset pricing literature are based on the data with monthly frequency and the same forecasting horizon.

factors.

$$M_{t,t+H} = \prod_{h=1}^{H} M_{t,t+h-1,t+h} \tag{11}$$

If the stochastic discount factor correctly conditionally prices the one-period returns, then it also correctly prices the multi-horizon returns. It follows from the law of iterated expectations[7] and the fact that multi-horizon returns are also products of single-period returns, i.e. $R_{t,t+H} = \prod_{h=1}^{H} R_{t+h-1,t+h}$

$$\begin{aligned} E[M_{t-h,t+1}R_{t-h,t+1}] &= E[M_{t-h,t}R_{t-h,t}M_{t,t+1}R_{t,t+1}] = \\ &= E[M_{t-h,t}R_{t-h,t}E_t[M_{t,t+1}R_{t,t+1}] = E[M_{t-h,t}R_{t-h,t}], \end{aligned} \tag{12}$$

The law of one price is therefore valid in multiple horizons and holds for the excess returns as well.

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 (2020).

To obtain the out-of-sample predictions, we train multiple models to include data that was available at the moment. Our first model is predicting returns for the year 1995. To do this, we take data from the beginning of the dataset to 1994 and split them in proportion 7:3 (keeping the time ordering) into train and validation samples. These two will be used when training the model. We use our model to make predictions for the year 1995. Our testing sample is beginning in December and ends in November. This means that we are using December characteristics to predict January returns in case $h = 1$. We repeat this procedure for the years 1996 to 2018, each time training a new model, to obtain our out-of-sample predictions.

We use a feed-forward neural network that is described in subsection B.1. The goal of the model is to aggregate all of the available input, anomalies, and condense them into one real-valued output.

As our labels, we use cumulative returns at horizon h that were cross-sectionally demeaned. We estimate the model for each horizon separately. Normalized anomalies serve as an input into the 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.

Every time model is trained, we perform a hyperparameter search. The hyperparameter space we search follows Tobek and Hronec (2020). We extend the hyperparameter space to cover more degrees of models complexity which could also vary across horizons. We test 6 different network architectures. The network can have either 1, 2 or 3 hidden layers. We also choose between wide and narrow network. The wide network has 150 nodes in each hidden layer, and the narrow has 32 nodes in the first hidden layer, 16 nodes in the second hidden layer and 8 in the third hidden layer (if the layer is present). Batch size is 256 or 1024. Dropout rate tested are 0.1, 0.01, and 0.001. Learning rate tested are 0.1,

7. $\mathrm{E}(X) = \mathrm{E}(\mathrm{E}(X \mid Y))$.

0.01, and 0.001. The number of epochs is fixed at 25. We keep Adam optimization betas at 0.9 and 0.999. The patience of early stopping is set to 5. The ensemble of five models is used, each with different random seed initialization. Reducing learning rate on plateau patience is applied after each epoch with learning rate halved if there is no improvement. The best set of hyperparameters, which is determined by the lowest mean square error, is used.

2.4 Portfolio formation

To assess the economic significance of our forecasts, we construct multiple portfolios. We use three portfolio construction methods, all of them based on portfolio sorting, a frequently used method in asset pricing[8].

Decile sorting

In case of only one set of predictions, each month, we cross-sectionally sort stocks based on the returns predictions. To construct a long-short portfolio, we buy firms that are in the highest predicted return decile, and short firms from the lowest decile for each month. We use equal-weighting, as in our empirical part we only focus on the universe of most liquid stocks[9].

Double sorting

A way to use two forecasts together, e.g. using different forecasting horizons or combining different models, is double sorting. Stocks are independently sorted into three groups based on each forecast separately. A long-short portfolio is constructed by buying the firms that are in the high expected return group for both forecasts and shorting firms that belong to the low expected return group for both forecasts.

Buy/hold spread strategy

Buy/hold spread, also referred to as banding, is a transaction cost mitigating technique. The strategy aim is to reduce turnover. It works by having a stricter rule to trade into position than to trade out of it. For example, 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%. Similarly, we sell the lowest 10% of the stocks and hold them until there are no longer in the bottom 20%.

While Novy-Marx and Velikov (2019) use this technique on only one model, or characteristic, we extend this to combine two different models. We use two models with different predicting horizons and use the one with the longer horizon as a buy signal and the shorter horizon as a hold signal. The reasoning behind this is that we will buy (sell) 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 this approach, we do not have consistency between buy and hold signal as in one signal case where it holds that if a firm

8. See, a survey of Green et al. (2013).
9. See Section C.3.

is in buy category, then it is also in hold category. We thus adjust the rule to remove firm from the portfolio when the two signals have opposing suggestions about the side of the trade (buy signal would buy while hold signal would sell and vice versa).

Returns calculation

Independent of the type of portfolio we are constructing we have target actions assigned to each firm at a given month. These actions are buy, sell, hold or nothing/remove from the portfolio. Portfolio value is calculated iteratively, as trading needs to reflect the current weights of the portfolio. This way, transaction costs can be accounted for properly as we know the exact size of the trade.

As a turnover reducing strategy, staggered portfolio rebalancing can be used. It works by prolonging holding period of strategy and rebalancing less often. Using this technique and having a holding period longer than the forecasting period will result in a staleness of the signal as we keep the firm in the portfolio past the intended period for which we forecasted.

In case of having a holding period longer than one month, we create multiple trajectories to use all of the information available. To create trajectories, we divide available capital at the start of investing into b parts. Each trajectory will function as a separate portfolio. Final portfolio value is obtained as a sum of values of trajectories. Portfolio returns are obtained by weighting returns from trajectories using the value of trajectory. The number of trajectories b will be equal to the holding period. The trajectories will be rebalanced in a staggered manner - each month one of the trajectories, the one that was rebalanced b months ago, is rebalanced to reflect the current target actions. This way, all of the information is used.

Now we calculate returns for each trajectory. Target weights w_{it}^* are assigned to each firm at each month. When we rebalance the trajectory, we divide available capital using equal-weighting between the firms. For decile sorting and double sorting, we fully reflect the current target actions. In 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 not rebalancing, we use the current holding of stocks, that is we use normalized weight from the end of the previous month.

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

$$w_{it} = w_{it}^* - ts_{it} \cdot tc_{it} \tag{13}$$

$$ts_{it} = w_{it} - w_{i(t-1)}^{end,norm} \tag{14}$$

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

$$w_{it}^{norm} = \frac{2w_{it}}{\sum_i |w_{it}|} \tag{15}$$

In case we remove 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 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.

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. The forecasts are from 1995 to 2018 (277 months). 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, apart from feedforward neural networks, gradient boosted regression trees were also used to obtain the forecasts, with results presented in Appendix D.

We follow the approach of Lewellen (2015) who assess the predictive ability of forecasts using regression of realized returns on predictions. Table 1 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, for most horizons significant, and decreasing 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 expected return. R^2 is decreasing with the horizon, for both the U.S. and the international sample suggesting that the predictability decreases with longer horizons. R^2 is higher for international dataset.

Table 1: Predictive ability of return forecasts

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.

	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

The decreasing predictability with longer horizons conclusion matches that of Lewellen (2015). Gu et al. (2020), being closer to our approach, have the opposite conclusion, they report higher R^2 for yearly predictions than for monthly ones.

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 horizon of predictions used to calculate the weights of the portfolio. It is intuitive rebalancing frequency as we have predictions for cumulative h-months returns of a given stock which allows us to fully utilize the predictions. This way, the turnover and transaction costs are lowered significantly for longer horizons.

Table 2 presents the mean, standard deviation, Sharpe ratio and maximum drawdown both for portfolios without and with transaction costs included. One month long-short portfolio results (that will serve as our benchmark) 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 (2020) also report comparable results on U.S. sample, albeit with slightly lower means and Sharpe ratios. This could be due to the fact that they include anomalies only after publication date.

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 one-month portfolio.

Long-only component of strategies has a higher mean return but also a higher variance and deeper maximum drawdowns compared to long-short strategy. 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 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 2003. The mean after this year is around 0.6-0.7% with transaction costs for all horizons and the standard deviation is lower. This break affects shorter horizons more. The longer horizons have higher Sharpe ratios and lower drawdown compared to $h = 1$.

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 holding

Table 2: Performance of long-short decile portfolios in the U.S.

The table shows the performance of long-short decile portfolios in the U.S. for the period between 1995 to 2018. Monthly mean returns, standard deviation, 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 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.

	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

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.

This is related to results presented in Figure 1. We show for multiple forecasting horizons how are returns varying each month, for up to two years, after rebalancing the portfolio. We show this for the case without transaction costs. In case we would want to include transaction costs, the first month returns would be lowered and then the month where we rebalance the portfolio. It shows us that 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 first five months are significant with decreasing returns for longer holding-period months. For the horizons 12 and 24, we have significantly positive return each month. It shows us that the underlying models are indeed learning for their intended horizon.

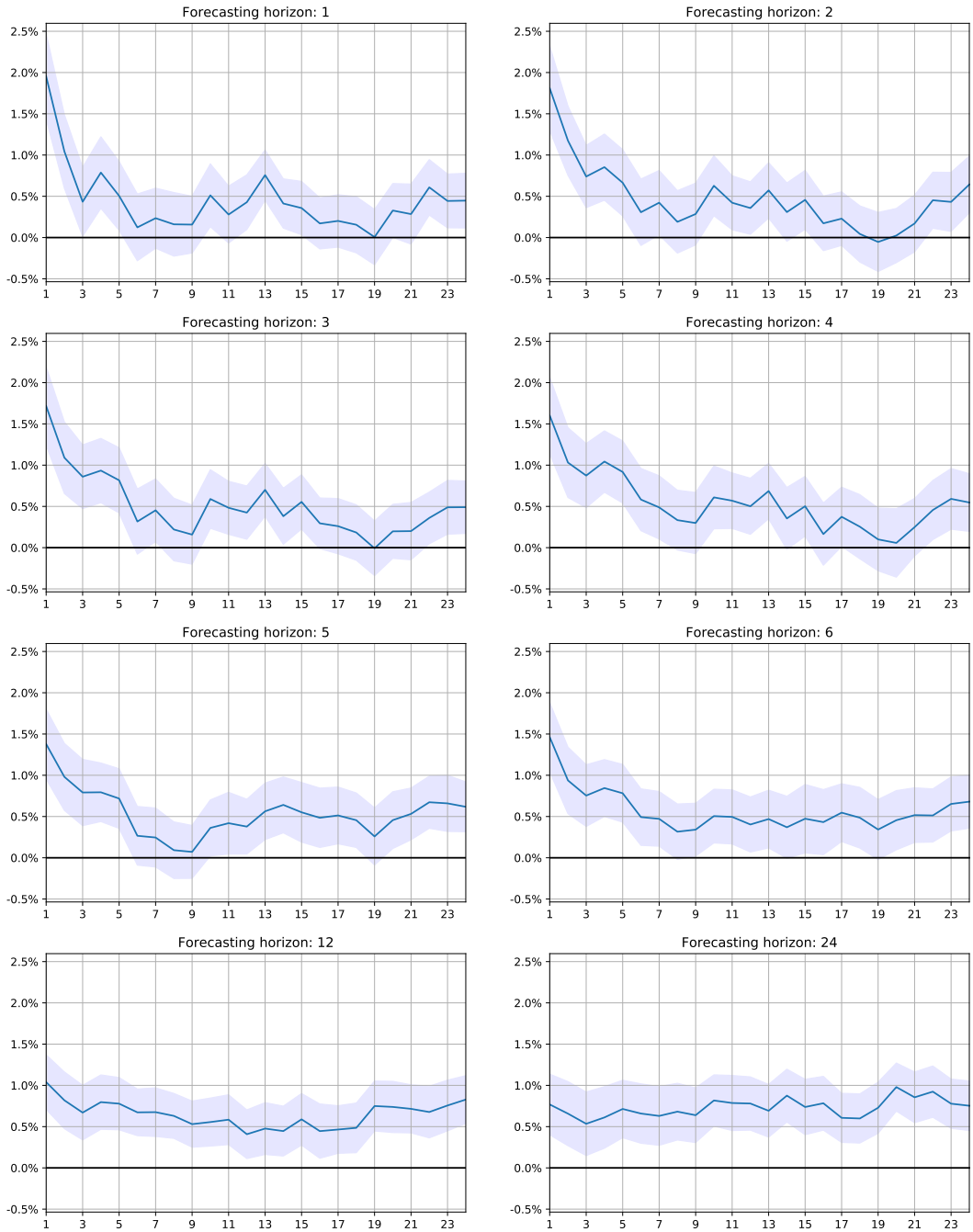

Figure 1: Average gross returns up to two years after rebalancing

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.

Double sorting portfolios

Double sorting portfolios were constructed by combining two predictions made at different horizon and rebalanced each month (holding period $b = 1$). We combine one-month forecasting horizon with longer horizons (2, 3, 6, 12, 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 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 3: Double-sorted portfolios performance in the U.S.

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, Sharpe ratio and maximum drawdown are reported. Reported values are in percentages with the exception of the Sharpe ratio.

	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

In Table 3 are the results of double-sorted portfolios. The best performing portfolio is 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 it almost twice as big. 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 means that we pay roughly the same transaction costs as our benchmark and that the benefits are not due to transaction costs differences but rather by 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 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 buy/hold spread strategy are constructed. We use 10%/20% buy/hold spread cutoffs. The portfolios should have lower turnover compared to long-short decile strategy as it is harder to trade into position than to trade out of it. We use predictions at various horizons as a buy signal and one-month predictions as a hold signal. Another benefit of this strategy could be from combining multiple predictions into one portfolio.

Portfolios performance is reported in Table 4. We refer to strategies by the buy and hold horizons that are used. Buy/hold portfolios have, on average, between 240 (for two-year portfolio) and 290 (for one-month portfolio) firms. Thus it is comparable to the number of firms in the decile portfolios and the double-sorted portfolios.

Table 4: Buy/hold spread portfolio performance in the U.S.

The profitability of long-short buy/hold spread portfolios in the U.S. for 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 show which horizons were used in the portfolio creation. Monthly mean returns, standard deviation, Sharpe ratio and maximum drawdown are reported. All values are reported in percentages except for the Sharpe ratio.

		Without transaction costs				With transaction costs				
		Mean	Std	Sharpe	MDD	Mean	Std	Sharpe	MDD	Turnover
buy	hold									
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

The 1-1 portfolio has lower turnover by 40% compared to the benchmark, but it has similar, slightly better performance. The benefit of reduced turnover is cancelled out by decreased performance because of a less strict cutoff. Turnover is decreasing with longer horizons. When we do not take transaction costs into account, the strategies are lower-mean, lower-risk with comparable Sharpe ratios to our benchmark. Portfolio 24-1 has the highest Sharpe ratio. It has slightly lower mean returns than our benchmark and a turnover of only 50%. Portfolio 1-6 is a more risky alternative with Sharpe ratio 1.08 but with higher returns and lower variance compared to the benchmark.

In Table A.3 is reported Sortino ratio, conditional value at risk, Alpha and Beta. Cumulative returns of long-short buy/hold spread portfolios compared with our base model, long-short decile portfolio at the one-month horizon, can be seen in Figure A.3. When transaction costs are not taken into account, the benchmark model is superior to buy/hold spread strategies. When transaction costs are present, buy/hold spread portfolios are better as they have lower turnover and thus lower transaction costs.

Comparing double sorting and buy/hold spread portfolios, double sorting is able to achieve significantly higher returns than 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 with interesting return-risk balance.

We show that for the U.S. sample extending the rebalancing frequency while keeping the forecasting horizon equal increases risk-adjusted profitability. 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 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 setting, laws or accounting standards. The preprocessing procedure we follow should lower these concerns. We train a feedforward neural network model on the international dataset (including the U.S.) 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 portfolio is created using the international sample forecasts. Forecasting horizon of presented portfolios is equal to the rebalancing frequency. The average number of firms in a portfolio is 600, 2.5 times more than in the U.S. setting.

In Table 5 are reported results of long-short decile portfolios with and without transaction costs. 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 as to our U.S. benchmark model; however, the standard deviation is almost halved for the international portfolio. Similarly, the mean return 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 because of the larger sample or diversification. The one-month strategy turnover is 120%, comparable to 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 (2020).

Other portfolios on the longer horizon have similar or lower Sharpe ratios and lower returns than 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 return lower by 0.24% after transaction costs, offering a lower-risk alternative.

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

Table 5: Performance of long-short decile portfolios - international sample

The table shows the performance of long-short decile portfolios on the international sample from 1995 to 2018. Monthly mean returns, standard deviation, 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 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.

	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

Overall, longer horizons decile portfolios are not superior compared to one-month strategy, though some horizons offer a lower return-risk alternative. Comparing with the evidence from the U.S., we note that including international data is beneficial by lowering the variance and keeping returns comparable to U.S. strategy.

In Table A.4 are presented additional performance measures. In Figure A.4 are cumulative returns of these portfolios with and without transaction costs in comparison with one-month international benchmark portfolio.

Double sorting portfolios

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

In Table 6 are results of double-sorted portfolios. Portfolios 1-12 and 1-24 have the highest Sharpe ratios, close to that of our international benchmark. They also have similar Sharpe ratio as double-sorted portfolios in the U.S. sample but offer a lower return. Turnover is similar to that of the benchmark one-month portfolio.

Table 6: Double-sorted portfolios performance - international sample

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

	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

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 using various forecasting horizons. 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 7. 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 7: Buy/hold spread portfolio performance - International sample

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

		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

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. Benchmark model and 1-1 buy/hold spread strategy perform the best both before and after transaction costs.

Overall, portfolios made using international dataset offer lower-risk opportunities compared to the U.S. sample. 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 we consider longer horizons or combination of horizons.

In our analysis, we are not investigating the option to combine the long and short components from different models. However, this approach might be appealing from the investor's perspective.

4 Conclusion

The fundamental question of asset pricing is why do different assets have different expected returns. Over the last forty years, hundreds of anomalies, or potential predictors of stock returns, were identified. Machine learning approach proved to be the best suited to address the problems of ambiguity of functional form as well as the level of dimensionality in the predictive regressions for stock returns. We examine the predictability of expected stock returns across multiple 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 predictability of returns using machine learning-based predictive regressions decreases with longer forecasting horizons. The reason behind decreasing predictability on longer horizons remains an open question and could be approached by examining the horizon-specific variable importance. We further address the critique that the profitability of machine learning-based portfolios disappears after the transaction costs. After accounting for the transaction costs, reducing the rebalancing frequency while matching the corresponding forecast horizon increases the risk-adjusted returns of machine learning-based portfolios. We also leverage return predictions for multiple horizons via double-sorted portfolios and achieve profitability improvement on the U.S. universe of stocks. Finally, we employ a turnover reducing strategy, buy/hold spread, and show higher risk-adjusted profitability in the U.S.

References

Avramov, Doron, Si Cheng, and Lior Metzker. 2021. "Machine learning versus economic restrictions: Evidence from stock return predictability." *Available at SSRN 3450322.*

Avramov, Doron, Kaplanski Guy, and Subrahmanyam Avanidhar. 2020. "Post-Fundamentals Drift in Stock Prices: A Machine Learning Approach." *Available at SSRN 3507512.*

Bali, Turan G, Robert F Engle, and Scott Murray. 2016. *Empirical asset pricing: The cross section of stock returns.* Chap. 7, 103–121. John Wiley & Sons.

Bryzgalova, Svetlana, Markus Pelger, and Jason Zhu. 2020. "Forest through the trees: Building cross-sections of stock returns." *Available at SSRN 3493458.*

Butaru, Florentin, Qingqing Chen, Brian Clark, Sanmay Das, Andrew W Lo, and Akhtar Siddique. 2016. "Risk and risk management in the credit card industry." *Journal of Banking & Finance* 72:218–239.

Campbell, John Y. 2017. *Financial decisions and markets: a course in asset pricing.* Princeton University Press.

Chen, Andrew Y, and Mihail Velikov. 2017. "Accounting for the anomaly zoo: A trading cost perspective." *Available at SSRN 3073681.*

Chen, Luyang, Markus Pelger, and Jason Zhu. 2020. "Deep learning in asset pricing." *Available at SSRN 3350138.*

Chen, Tianqi, and Carlos Guestrin. 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, Kee H, and Hao Zhang. 2014. "A simple approximation of intraday spreads using daily data." *Journal of Financial Markets* 17:94–120.

Cochrane, John H. 2009. *Asset pricing: Revised edition.* Princeton university press.

DeMiguel, Victor, Alberto Martin-Utrera, Francisco J Nogales, and Raman Uppal. 2020. "A transaction-cost perspective on the multitude of firm characteristics." *The Review of Financial Studies* 33 (5): 2180–2222.

Fama, Eugene, and Kenneth French. 1992. "The cross-section of expected stock returns." *Journal of Finance* 47 (2): 427–465.

Feng, Guanhao, Nick Polson, and Jianeng Xu. 2020. "Deep Learning in Characteristics-Sorted Factor Models." *Available at SSRN 3243683.*

Föllmer, Hans, and Alexander Schied. 2011. *Stochastic finance: an introduction in discrete time.* Walter de Gruyter.

Fong, Kingsley YL, Craig W Holden, and Ondrej Tobek. 2018. "Are volatility over volume liquidity proxies useful for global or US research?" *Kelley School of Business Research Paper,* nos. 17-49.

Frazzini, Andrea, Ronen Israel, and Tobias J Moskowitz. 2012. "Trading costs of asset pricing anomalies." *Fama-Miller working paper,* 14–05.

French, Kenneth R. 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, Joachim, Andreas Neuhierl, and Michael Weber. 2017. *Dissecting characteristics nonparametrically.* Technical report. National Bureau of Economic Research.

Friedman, Jerome H. 2001. "Greedy function approximation: a gradient boosting machine." *Annals of statistics,* 1189–1232.

Giglio, Stefano, and Dacheng Xiu. 2019. "Asset pricing with omitted factors." *Chicago Booth Research Paper,* nos. 16-21.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning.* MIT press.

Green, Jeremiah, John RM Hand, and X Frank Zhang. 2013. "The supraview of return predictive signals." *Review of Accounting Studies* 18 (3): 692–730.

Griffin, John M, Patrick J Kelly, and Federico Nardari. 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, Shihao, Bryan Kelly, and Dacheng Xiu. 2020. "Empirical asset pricing via machine learning." *The Review of Financial Studies* 33 (5): 2223–2273.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

Heaton, JB, NG Polson, and Jan Hendrik Witte. 2017. "Deep learning for finance: deep portfolios." *Applied Stochastic Models in Business and Industry* 33 (1): 3–12.

Hou, Kewei, Chen Xue, and Lu Zhang. 2020. "Replicating anomalies." *The Review of Financial Studies* 33 (5): 2019–2133.

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

Ioffe, Sergey, and Christian Szegedy. 2015. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." *arXiv preprint arXiv:1502.03167.*

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

Kamara, Avraham, Robert A Korajczyk, Xiaoxia Lou, and Ronnie Sadka. 2015. "Horizon pricing." *Journal of Financial and Quantitative Analysis (JFQA)* 51:1769–1793.

Kelly, Bryan T, Seth Pruitt, and Yinan Su. 2019. "Characteristics are covariances: A unified model of risk and return." *Journal of Financial Economics* 134 (3): 501–524.

Khandani, Amir E, Adlar J Kim, and Andrew W Lo. 2010. "Consumer credit-risk models via machine-learning algorithms." *Journal of Banking & Finance* 34 (11): 2767–2787.

Kingma, Diederik P, and Jimmy Ba. 2014. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980.*

Korajczyk, Robert A, and Ronnie Sadka. 2004. "Are momentum profits robust to trading costs?" *The Journal of Finance* 59 (3): 1039–1082.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh. 2020. "Shrinking the cross-section." *Journal of Financial Economics* 135 (2): 271–292.

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

Messmer, Marcial. 2017. "Deep learning and the cross-section of expected returns." *Available at SSRN 3081555.*

Novy-Marx, Robert, and Mihail Velikov. 2019. "Comparing cost-mitigation techniques." *Financial Analysts Journal* 75 (1): 85–102.

Schmidt, Peter S, Urs Von Arx, Andreas Schrimpf, Alexander F Wagner, and Andreas Ziegler. 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.

Sharpe, William F. 1963. "A simplified model for portfolio analysis." *Management science* 9 (2): 277–293.

Sharpe, William F. 1966. "Mutual fund performance." *The Journal of business* 39 (1): 119–138.

Sharpe, William F. 1994. "The sharpe ratio." *Journal of portfolio management* 21 (1): 49–58.

Sirignano, Justin, Apaar Sadhwani, and Kay Giesecke. 2018. "Deep learning for mortgage risk." *arXiv preprint arXiv:1607.02470.*

Sloan, Richard G. 1996. "Do stock prices fully reflect information in accruals and cash flows about future earnings?" *Accounting review,* 289–315.

Sortino, Frank A, and Lee N Price. 1994. "Performance measurement in a downside risk framework." *the Journal of Investing* 3 (3): 59–64.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. "Dropout: a simple way to prevent neural networks from overfitting." *The journal of machine learning research* 15 (1): 1929–1958.

Tobek, Ondrej, and Martin Hronec. 2020. "Does it Pay to Follow Anomalies Research? Machine Learning Approach With International Evidence." *Journal of Financial Markets (forthcoming).*

A Additional tables and figures

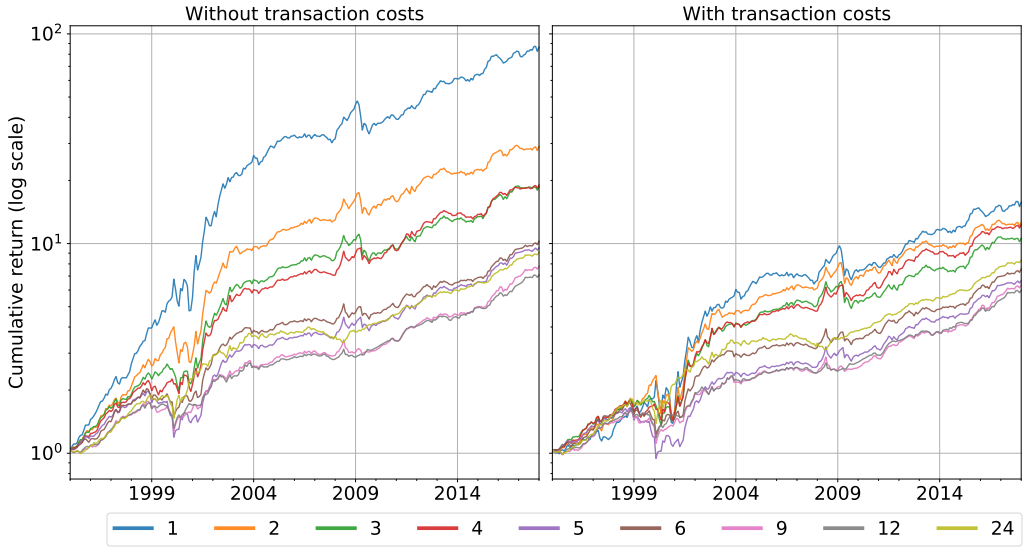

Figure A.1: Cumulative returns of long-short decile portfolios in the U.S.

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.

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.

	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

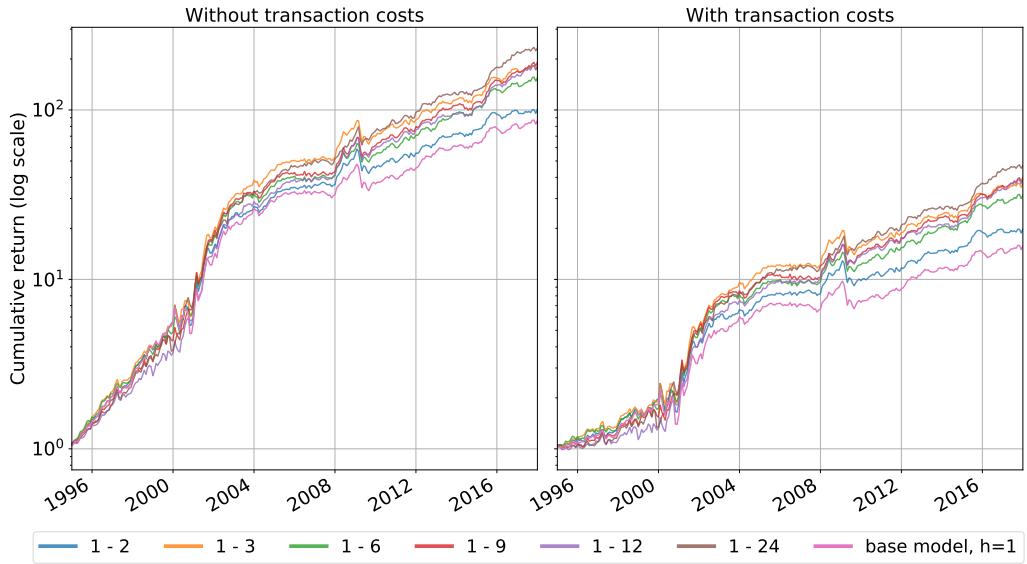

Figure A.2: Cumulative returns of long-short double sorting portfolios in the U.S.

The figure shows cumulative returns on a logarithmic scale of the double-sorting strategy and of 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.

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.

	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

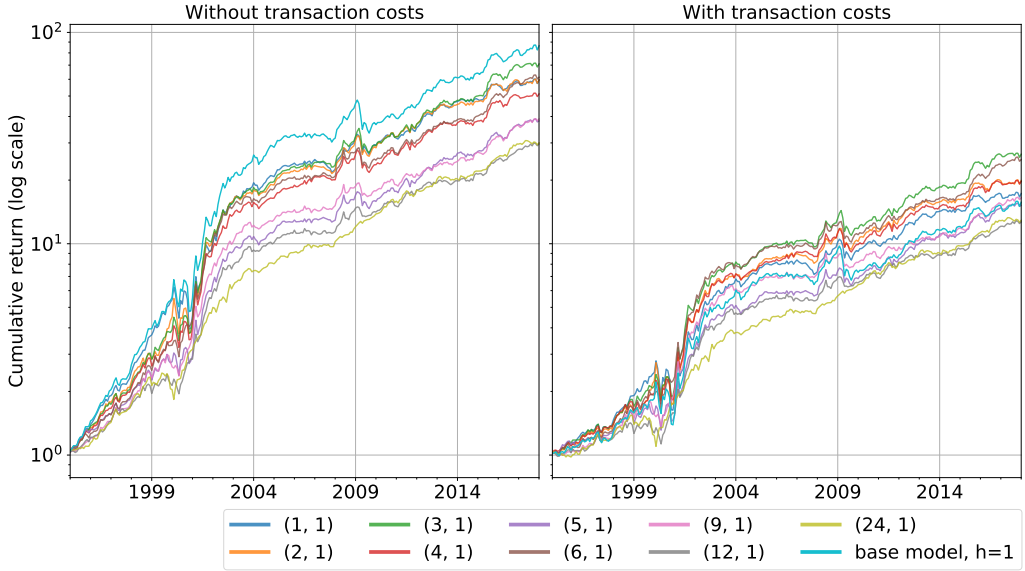

Figure A.3: Cumulative returns of buy/hold spread portfolios in the U.S.

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

Table A.3: Buy/hold spread portfolio performance metrics - U.S. sample

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. Portfolio label signifies horizon based on which we buy and horizon based on which we hold stocks respectively.

		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

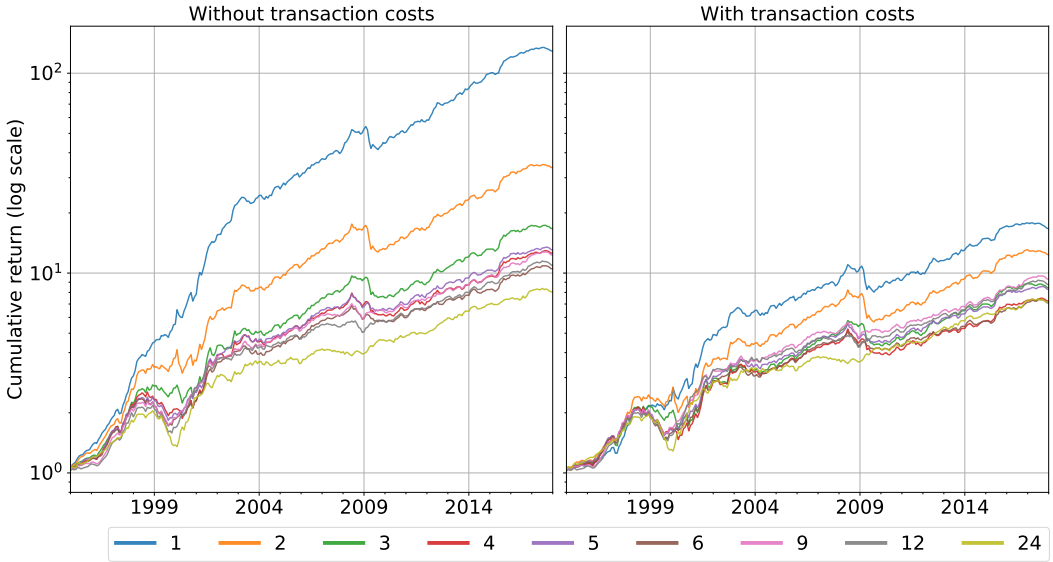

Figure A.4: Cumulative returns of long-short decile portfolios on international sample

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

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. Portfolio label is the forecasting horizon and the holding period.

	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

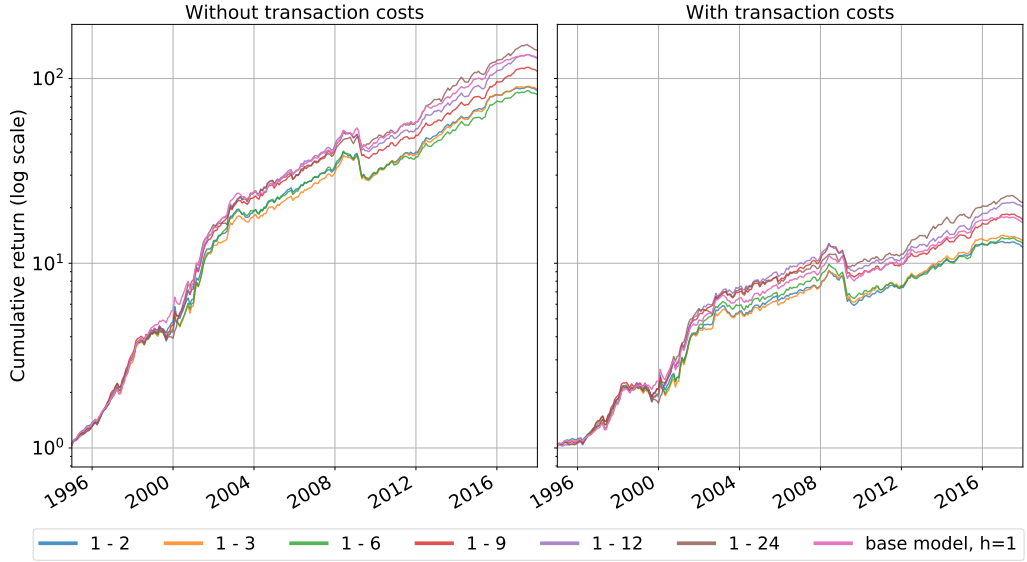

Figure A.5: Cumulative returns of long-short double sorting portfolios - international universe

The figure shows cumulative returns of the double sorting strategy in comparison with long-short decile portfolio at horizon one, both for the international universe. Portfolios are plotted before and after accounting for transaction costs. 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

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.

	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

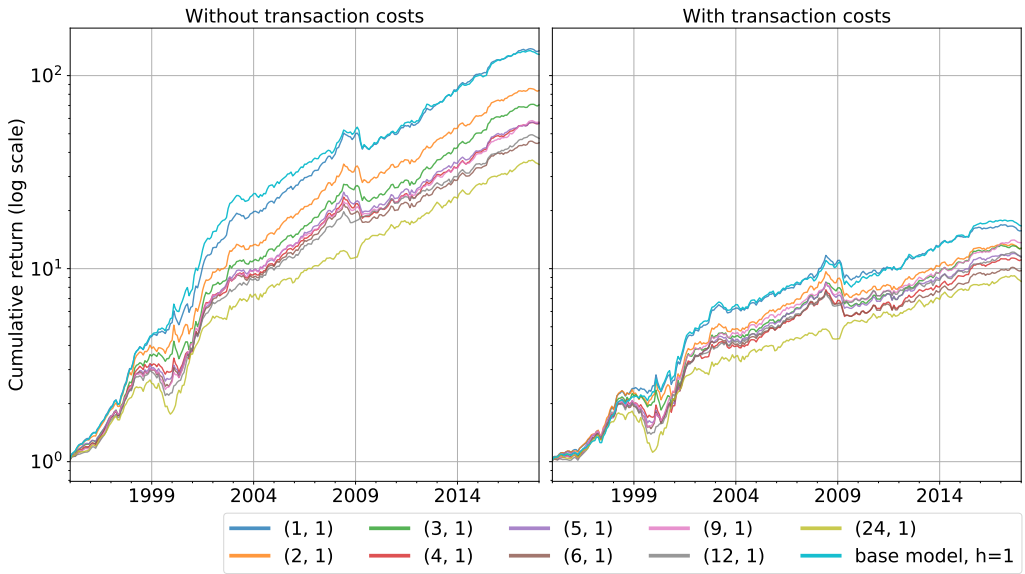

Figure A.6: Cumulative returns of buy/hold spread portfolios on international sample

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

Table A.6: Buy/hold spread portfolio performance metrics - International sample

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 buy/hold spread of 10%/20%. Portfolio label signifies horizon based on which we buy and hold stocks, respectively.

		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

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 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.

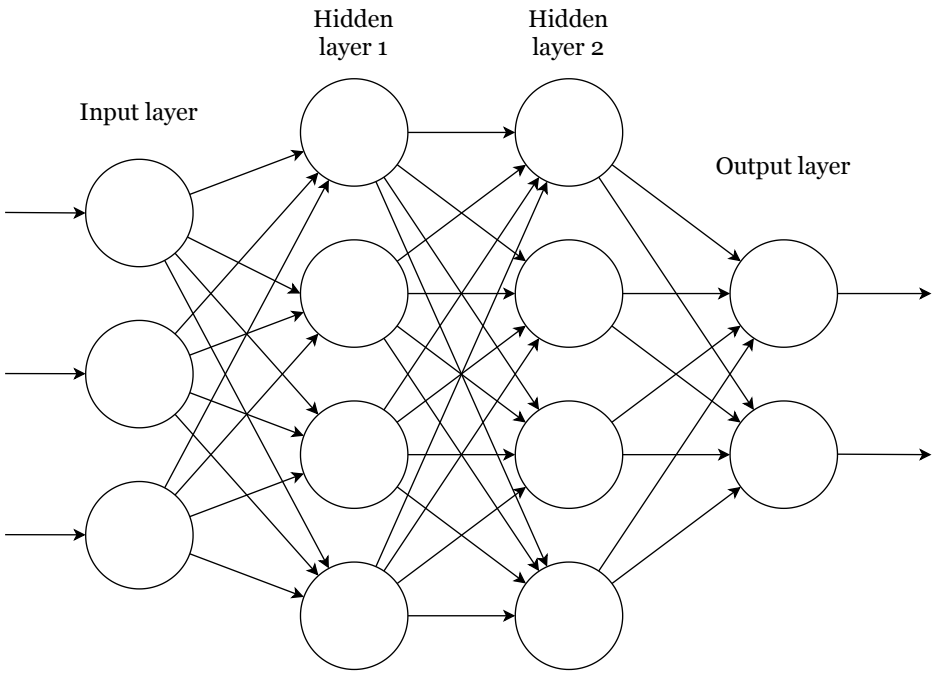

Figure B.1: Example of multilayer fully connected neural network.

Neuron i is defined as:

$$y_i = \varphi\left(s_i + b_i\right), \quad s_i = \sum_{j=1}^{m} w_{ij} x_j \tag{16}$$

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, $\varphi(.)$ is the activation function, and y_i is the output of the neuron i.

Commonly used activation function, and the one that we will be using, is called rectified

linear unit (ReLU) and it is defined as:

$$\mathrm{ReLU}(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{otherwise} \end{cases} \tag{17}$$

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**[10]. For a function with one input, the derivative $f'(x)$ provides us with 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 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 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 first and second moments of gradients.

When we move in the direction of the steepest descend, 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 fixed minimum or exponential decay.

When using 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. It is shown in Equation 18.

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{n=1}^{N}(\theta_n - \hat{\theta}_n)^2 \tag{18}$$

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

10. Also called objective function, criterion, or error function.

The **backpropagation algorithm** efficiently calculates the gradient of the loss function with respect to 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 allows us to see which node is responsible for most of the error and lets us change the parameters 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 is for controlling the kind of functions we allow our model to take or specifying which functions are preferred. Regularization is a modification to the neural network with the aim of reducing generalization error, to prevent overfitting. Regularization techniques we use are 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 with normalizing of the input of each layer for each minibatch during the training. It allows us to use higher learning rates, and it also 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 seed 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 in a similar way as 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 is preventing the nodes to co-adapt 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 of these categories. CART creates binary trees - each non-terminal node is split into two nodes. Benefits of trees include intuitive interpretation or the fact that it allows 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 forest or rotation forest.

Boosted regression trees were first proposed by Friedman (2001). Gradient boosting is a machine learning technique which works by using 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, allows regularization to improve generalization.

The tree is defined as

$$f_t(x) = w_{q(x)}, w \in R^T, q : R^d \to \{1, 2, \cdots, T\} \tag{19}$$

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.

$$\hat{y}_i = const. + \sum_{k=1}^{K} \nu f_k(\mathbf{x}_i), \quad f_k \in \mathcal{F} \tag{20}$$

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

Our loss function is the following:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{21}$$

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:

$$\hat{y}_i^{(t)} = \phi(x_i) = \hat{y}_i^{(t-1)} + \nu f_t(x_i) \tag{22}$$

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

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)\right) + \Omega(f_t) \tag{23}$$

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

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}^{(t-1)}\right) + g_i f_t\left(\mathbf{x}_i\right) + \frac{1}{2} h_i f_t^2\left(\mathbf{x}_i\right) \right] + \Omega\left(f_t\right) \tag{24}$$

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 node or not, we compare I_R and I_L with the I to see whether there is loss reduction after splitting. More formally

$$\mathcal{L}_{split} = \frac{1}{2} \left[\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i\right)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \tag{25}$$

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 of 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:

$$SR_k = \frac{E\left[R_k - R_f\right]}{\sqrt{\operatorname{var}\left(R_k\right)}} \tag{26}$$

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 Sharpe ratio so we can also compare to the benchmark changing over time.

$$SR_k = \frac{E\left[R_k - R_b\right]}{\sqrt{var(R_k - R_b)}} \tag{27}$$

Sharpe ratio 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.

A crucial and often overlooked fact is that Sharpe ratio is also simply a rescaled t-statistic for statistical significance of mean being different from zero. 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 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 Sharpe ratio, we include Sortino Ratio. It is a modification of Sharpe ratio by Sortino and Price (1994) that penalizes only returns that are below minimum acceptable return (MAR). This way only the variation below MAR is counted in the denominator. Sortino ratio is calculated as:

$$\text{Sortino}_k = \frac{E\left[R_k - MAR\right]}{\sqrt{\frac{1}{T}\sum_{t=1}^{T} \min\left(0; R_{t,k} - MAR\right)^2}} \tag{28}$$

Denominator measures downside deviation. Minimum acceptable return of 0% will be used when using Sortino ratio.

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

$$\text{VaR}_\alpha(X) = \inf\{x \in \mathbb{R} : P(X + x < 0) \leq 1 - \alpha\} \tag{29}$$

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 VaRs of individual components.

Conditional Value at Risk (CVaR), for which the subadditivity holds, is defined as:

$$\text{CVaR}_\alpha = \frac{1}{\alpha}\int_0^\alpha VaR_\alpha(X)d\alpha \tag{30}$$

This measure is sometimes called Expected Shortfall. It gives the average value at risk at level $\alpha \in (0, 1)$ of a position X. For example, $CvaR_{1\%}$ 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 $R_p(w_1, \ldots, w_n, t)$ being the cumulative portfolio return over portfolio holding time drawdown is defined as

$$D(\boldsymbol{w}, t) = \max_{0 \leq \tau \leq t}\{R_p(\boldsymbol{w}, \tau)\} - R_p(\boldsymbol{w}, t) \tag{31}$$

Maximum Drawdown up to time T is:

$$MDD(T) = \max_{0 \leq \tau \leq T}\{D(\boldsymbol{w}, \tau)\} \tag{32}$$

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

$$R_{s,t} - R_f = \alpha_i + \beta_i(R_{M,t} - R_f) + \epsilon_{i,t} \tag{33}$$

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:

$$Turnover_t = \frac{1}{ge} \sum_i |ts_{it}| \tag{34}$$

where ge is gross exposure, the sum of long and short positions divided by the capital, and ts_{it} is the trade size. 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 at a given month. Closing quoted spread (Chung and Zhang, 2014) and volatility over volume (Fong et al., 2018) proxies are used.

Closing quoted spread

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

$$QS = \frac{1}{T} \sum_{t=1}^{T} \frac{2(ask - bid)}{ask + bid} \tag{35}$$

with bid being the closing bid, ask is the closing ask and T is the number of days for a given month. If the daily value of QS is missing or negative, it is not included in the calculating 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:

$$VoV(\%\ spread) = 8 \frac{\sigma^{2/3}}{avg\ vol^{1/3}} \tag{36}$$

with σ being the standard deviation of daily returns, $avg\ vol$ being 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 and showed that only closing quoted spread outperformed this proxy. VoV proxy has less missing observations than quoted spread as it uses returns and volume only and not closing bid and ask.

We combine the two proxies them by using closing quoted spread and in case of missing observation we fill in with volatility over volume and then with 5%. Average estimated transaction costs over time are displayed in Figure B.2.

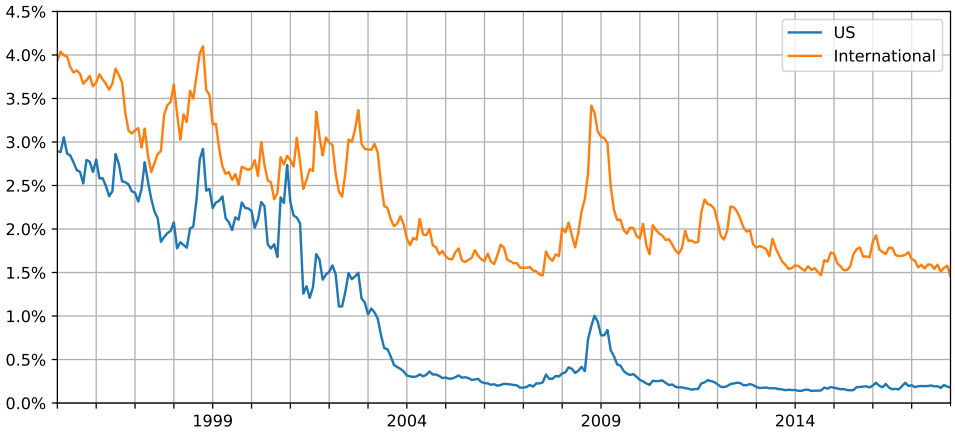

Figure B.2: Average estimated transaction costs

Estimated transaction costs cross-sectional average for the U.S. and international sample (with the U.S. excluded). Transaction costs are estimated using closing quoted spread (Chung and Zhang, 2014) and volatility over volume (Fong et al., 2018).

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 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 starting date at the latest on the last day of the month t and ending date has to be on the last day of the month or later. Preprocessing of daily and monthly dataset is otherwise the same.

To get U.S. shares only, we filter based on 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 PRC variable is missing or zero if the stock was not traded. Absolute value is taken as CRSP reports 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 is not corresponding 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 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 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 of 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 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 (identificators 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. 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 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 delisting return because Datastream does not report separate delisting return as CRSP. Order of magnitude of our variables is adjusted so that they are the same as in 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 missing return index. We calculate the daily return from 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 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 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 daily return is over 200%, but 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 last available value for a given month is used or sum over the month in case of volume traded. We compare 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 market value reported by Datastream with a self-created market value that we calculate by multiplying 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 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 regions.

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 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 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 the U.S. stocks.

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

In Table C.1 are reported 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. Average monthly return in the U.S. is two times higher than in the international sample. U.S. dataset has, on average, 1100 firms at the end of the month. Including international dataset provides, on average, additional 1870 firms per month.

Table C.1: Descriptive statistics

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.

	US			International		
	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
Min	-100.00	27.05	647.00	-99.97	52.00	1297.00
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
Max	300.17	1099436.06	1734.00	1301.01	563055.56	2347.00

D Robustness - Gradient boosted regression trees

As a robustness check, we use gradient boosted regression trees[11] 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

11. See subsection B.1 for more details

optimal number of trees, we test 50, 100, 200, 300, 400, and 500. Maximum depth of each tree between one and nine is considered and learning rates 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 holding period is equal to forecasting horizon used, are presented in Table D.1. Results for one-month horizon have 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 long-short strategy. Short only component seems ineffective in this case. More performance metrics for portfolios are in Table D.2. 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, two-month portfolio is better.

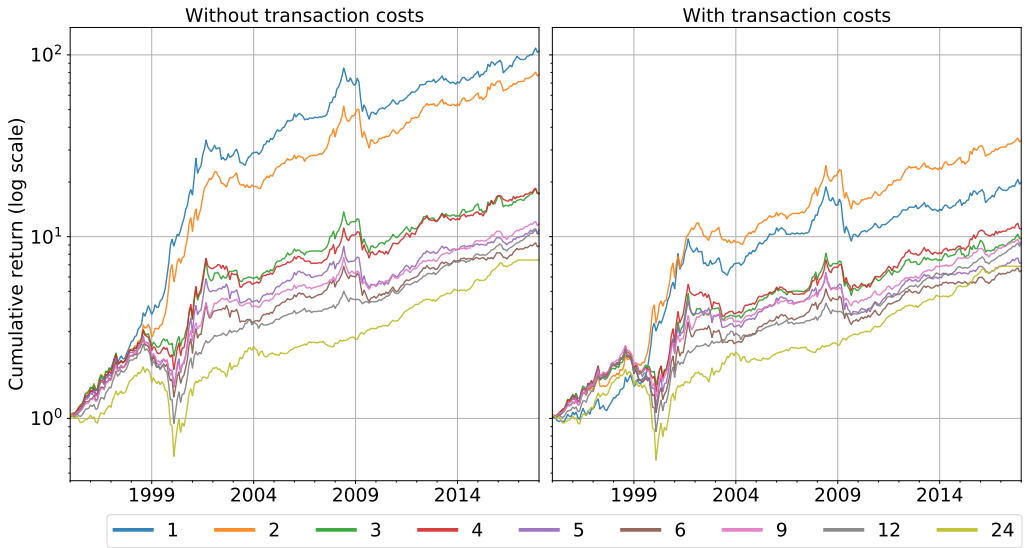

Figure D.1: Cumulative returns of long-short decile portfolios in the U.S.

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 D.1: Performance of long-short decile portfolios in the U.S.

The table shows the performance of long-short decile portfolios in the U.S. for the period between 1995 to 2018. Monthly mean returns, standard deviation, 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 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.

	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

Table D.2: Performance measures of long-short decile portfolios in the U.S.

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.

	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

Double-sorted portfolios were made with cutoffs of top 15% and bottom 15%. In Table D.3 is shown the performance of double-sorted long-short portfolios. Portfolio 1-2 has slightly higher Sharpe ratio than 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.4. Betas are more negative than in the case of decile portfolios.

Table D.3: Double-sorted portfolios performance in the U.S.

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, Sharpe ratio and maximum drawdown are reported. Reported values are in percentages with the exception of the Sharpe ratio.

	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

Table D.4: Double-sorted portfolios performance metrics - in the U.S.

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.

	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

Performance of long-short buy/hold spread of 10%/20% is presented in Table D.5. Portfolio 2-1 has the highest Sharpe ratio and mean, higher than one-month decile portfolio. Additional metrics for these portfolios are in Table D.6. 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).

Table D.5: Buy/hold spread portfolio performance in the U.S.

The profitability of long-short buy/hold spread portfolios in the U.S. for 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 show which horizons were used in the portfolio creation. Monthly mean returns, standard deviation, Sharpe ratio and maximum drawdown are reported. All values are reported in percentages except for the Sharpe ratio.

		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

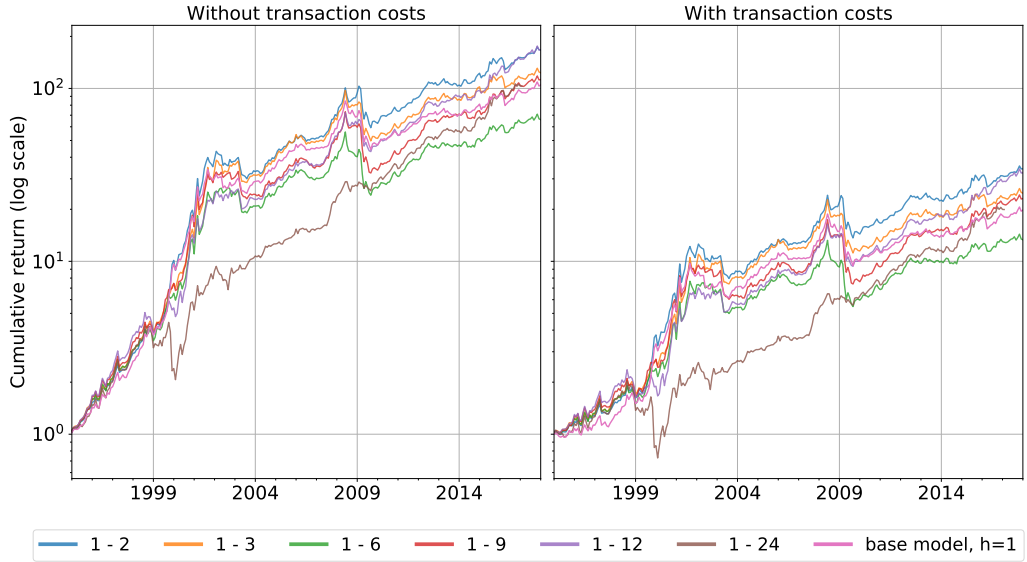

Figure D.2: Cumulative returns of long-short double sorting portfolios in the U.S.

The figure shows cumulative returns on a logarithmic scale of the double-sorting strategy and of 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.

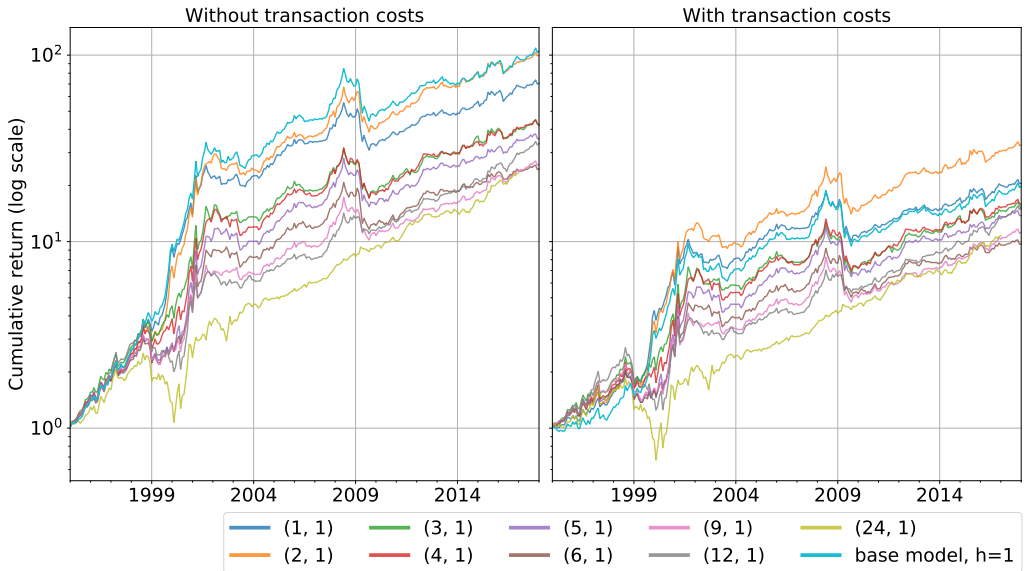

Figure D.3: Cumulative returns of buy/hold spread portfolios in the U.S.

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

Table D.6: Buy/hold spread portfolio performance metrics - U.S. sample

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. Portfolio label signifies horizon based on which we buy and horizon based on which we hold stocks respectively.

		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

IES Working Paper Series

2021

1. Mahir Suleymanov: *Foreign Direct Investment in Emerging Markets: Evidence from Russia since the 2000s*
2. Lenka Nechvátalová: *Multi-Horizon Equity Returns Predictability via Machine Learning*

All papers can be downloaded at: http://ies.fsv.cuni.cz

Univerzita Karlova v Praze, Fakulta sociálních věd
Institut ekonomických studií [UK FSV – IES] Praha 1, Opletalova 26
E-mail : ies@fsv.cuni.cz http://ies.fsv.cuni.cz