Journal of Financial Markets 56 (2021) 100588

Contents lists available at ScienceDirect

Journal of Financial Markets

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

Does it pay to follow anomalies research? Machine learning approach with international evidence☆

Ondrej Tobek [a,1], Martin Hronec [b,c,*,2]

[a] University of Cambridge, Department of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD, UK
[b] The Czech Academy of Sciences, Institute of Information Theory and Automation, Pod Vodarenskou Vezi 4, 182 00, Prague, Czech Republic
[c] Institute of Economic Studies, Charles University, Opletalova 26, 110 00, Prague, Czech Republic

ARTICLE INFO

Article history:
Received 9 September 2019
Revised 4 July 2020
Accepted 5 July 2020
Available online 5 August 2020

JEL classification:
G11
G12
G15

Keywords:
Anomalies
Machine learning
International finance

ABSTRACT

We study out-of-sample returns on 153 anomalies in equities documented in the academic literature. We show that machine learning techniques that aggregate all the anomalies into one mispricing signal are profitable around the globe and survive on a liquid universe of stocks. We investigate the value of international evidence for selection of quantitative strategies that outperform out-of-sample. Past performance of quantitative strategies in regions other than the United States does not help to pick out-of-sample winning strategies in the U.S. Past evidence from the U.S., however, captures most of the return predictability outside the U.S.

The empirical asset pricing literature is predominantly built on evidence from the financial markets in the United States. This phenomenon is termed "academic home bias puzzle" by Andrew Karolyi (2016), who documents that only 16% of the empirical papers published in the four leading finance journals investigate markets outside the United States. Yet essential research questions can be answered very differently when looking at the international evidence. One such question is the post-publication decline in the returns of anomalies. McLean and Pontiff (2016) document a 58% decrease in post-publication profitability relative to the in-sample profitability of portfolios based on the underlying anomalies. Jacobs and Müller (2020), however, show that the U.S. is the only country with a reliable post-publication decline in the returns of anomalies, emphasizing the importance of international evidence in asset pricing. On the other hand, in case of predictive regression for stock returns, evidence from the U.S. and around the globe is more similar. Green et al. (2017) in the U.S., and Jacobs and Müller (2018) internationally, find that combining anomalies into one mispricing signal using the least squares approach leads to superior out-of-sample risk-adjusted returns relative to focusing on individual anomalies. The benefit of combining individual anomalies through predictive regres-

☆ We have benefited greatly from comments from Andrew Harvey, Mark Salmon, Söhnke Bartram, Jozef Baruník and the participants at the FMND 2019 and ECOSTA 2019 conferences.

* Corresponding author. Institute of Economic Studies, Charles University, Opletalova 26, 110 00, Prague, Czech Republic.
E-mail addresses: ondrej.tobek@gmail.com (O. Tobek), martin.hronec@fsv.cuni.cz (M. Hronec).

1 Ondrej Tobek gratefully acknowledges support from the Economic and Social Research Council.

2 Martin Hronec gratefully acknowledges support from the Czech Science Foundation under the 19-28231X (EXPRO) project. This research was also supported by the Grant Agency of Charles University (grant no. 1270218) and Charles University Research Centre program No. UNCE/HUM/035.

https://doi.org/10.1016/j.finmar.2020.100588
1386-4181/

sions is further amplified by Gu et al. (2020), who conclude that more sophisticated machine learning methods offer higher out-of-sample predictability in the U.S. compared to the traditional methods in Jacobs and Müller (2018).

In this study, we examine international evidence in machine learning-based predictive regressions for stock returns using anomalies as predictors (mispricing strategy hereafter). We offer two main contributions. Firstly, using machine learning models instead of traditional linear models leads to substantial and superior out-of-sample long/short returns on the mispricing strategy both globally as well as in all the individual regions. This profitability is documented among large-cap stocks, after accounting for transaction costs, as well as short-selling constraints, and provides strong support to the findings of Gu et al. (2020). Secondly, unexpectedly, extending the estimation sample with observations from outside the U.S. does *not* benefit U.S. investor leveraging predictive-regression-based strategies using firm characteristics. The benefits of a larger estimation sample is offset by the region-specific differences. The differences across regions are visible in a heterogeneity of marginal importance of variables, which however does not translate into profitability differences in the mispricing strategy.

The mispricing strategy is based on the estimated historical relation between the past characteristics and future returns of individual stocks in the U.S., Japan, Europe, and Asia Pacific. We use 153 anomalies documented in the literature. Anomaly describes an individual stock characteristic that was shown to predict future returns in cross-section. These anomalies are, for example, earnings over price of Basu (1977), accruals of Sloan (1996), R&D over market equity of Chan et al. (2001), and composite equity issuance of Daniel and Titman (2006).

The historical relationships are typically linearly approximated using Fama and MacBeth (1973) least squares regressions in the literature, as in Lewellen et al. (2015), Green et al. (2017), or Jacobs and Müller (2018). The topic of this study is the closest to Jacobs and Müller (2018, 2020), who analyze returns on anomalies outside the U.S. Our study is, however, different in many aspects. Firstly, it focuses on a liquid universe of stocks that contains stocks with capitalization in the top 95% of the overall market's capitalization and dollar trading volume over the previous year in the top 95% of the overall market's volume in the individual regions. Only about 1000 of the most liquid stocks pass the criteria in the 2010s in a given month in the U.S. Excluding small-cap stocks leads to results more relevant to investors and limits the effect of microstructure noise.[3]

Secondly, we investigate the role of international evidence in the strategies. Jacobs and Müller (2018, 2020) focus solely on strategies that use data in the respective regions without evaluating the possible benefits of using global data to predict future returns.

Thirdly, the prediction methods differ among the studies. Our study is the closest in methodology and application of machine learning techniques to Gu et al. (2020), who show that machine learning methods significantly outperform the linear approximation in the U.S. We extend the use of machine learning methods from the U.S. to international markets. Specifically, we compare the baseline least squares regressions to more complex gradient boosting regression trees, random forests, and neural networks. We find that the machine learning methods lead to significant gains in performance of the mispricing strategy in all the regions, which provide support for the conclusions of Gu et al. (2020) in true out-of-sample fashion.

A large difference with respect to Gu et al. (2020) is that we allow only anomalies documented in the previously published studies to enter predictions in each year. That is, the information set of anomalies available to investors at the time they make an investment decision. Ignoring this assumption can lead to illusory profits that cannot be obtained in practice. Another difference is their focus on the full universe of stocks, which has profound effects on their conclusions. The most important anomalies in their estimations are liquidity, size, and return over the past month (short-term return reversal). Asparouhova et al. (2010) argue that these variables are connected to future returns mainly through microstructure biases and have nothing to do with the true predictability of stock returns that is of interest to investors.[4]

An alternative potential explanation for the profitability of the mispricing strategy could be the well-documented limits to arbitrage from the anomalies literature (e.g., Stambaugh et al., 2012; Avramov et al., 2013; Chordia et al., 2014; Hou et al., 2018).

The importance of imposing economic restrictions on the results of machine learning models in the asset pricing is emphasized by Avramov et al. (2019). Generative adversarial networks and recurrent neural network with long short-term memory architecture of Chen et al. (2019), instrumented principal component analysis of Kelly et al. (2019), conditional autoencoders of Gu et al. (2019), as well as Gu et al. (2020) offer weaker returns predictability under the economic restrictions. We address these concerns primarily by focusing on the large cap universe. Hou et al. (2018) show that most of the individual anomalies are much weaker when microcaps are eliminated from the U.S. investment universe. However, we find that the profitability of the mispricing strategy is substantial, even though it is studied on the liquid universe of stocks. Additionally, we examine short-selling constraints and transaction costs.

It is often impossible to short-sell certain stocks due to an insufficient supply of borrowable shares, but Stambaugh et al. (2012) describe how the profitability of anomalies often comes precisely from the short leg of the portfolios. Therefore, we decompose the returns on the long-short portfolios into long-only and short-only components. Short-selling constraints however cannot fully explain the profitability as both the long-only and short-only legs of the mispricing strategy offer a profitable investment opportunity with respect to market returns.

Moreover, we estimate the transaction costs associated with machine learning-based strategies leveraging predictive power of anomalies internationally for the first time. Novy-Marx and Velikov (2015) study transaction costs on a range of anomalies in the U.S. and conclude that the transaction costs are important mainly for high-turnover anomalies whose returns net of

[3] See Asparouhova et al. (2010) for a description of the effect of microstructure noise.

[4] See Roll (1984) for a simple model decomposing stock returns into microstructure noise and changes in true prices.

transaction costs often turn negative.[5] Our mispricing strategy remains profitable across the regions even after accounting for the transaction costs. The profitability of the strategy is therefore not illusory and can be capitalized by the investors.

Finally, we examine the value of international evidence for the prediction of out-of-sample returns on the anomalies. Harvey et al. (2016) and Hou et al. (2018) show that many anomalies cannot be replicated and many others are significant only due to the in-sample data snooping. New anomalies identified in the literature are based on the same historical datasets in the U.S., which can lead to false positive discoveries. International data provide new information with respect to the U.S. and it could, therefore, limit the number of false discoveries.[6] Using international data could also simply lead to different insights compared to the results from the U.S. as is the case with the stability of return predictability of anomalies over time, see Schwert (2003), Chordia et al. (2014), or McLean and Pontiff (2016) for the U.S. evidence and Lu et al. (2017) or Ilmanen et al. (2019) for international evidence.

Another reason we use international data is to increase sample size, which in turn leads to more precise estimates under certain conditions. Central limit theorem establishes that the confidence interval around a point estimate shrinks with the square root of a number of observations. In the simplest version of central limit theorem, it is assumed that the data are identically and independently distributed. There could be a problem in that some anomalies are specific to the U.S. as they depend on the local institutional setting. For example, accruals depend on country-specific accounting rules. The institutional uniqueness then limits the value of data outside the U.S. for predictions in the U.S. as it breaks the assumption on the identical distribution. In short, the usefulness of international data depends on the structure of drivers of stock returns around the globe. If the drivers are primarily *global* then the larger international sample should be beneficial. On the other hand, if the drivers are primarily *local* then the larger international sample should have little benefit. The role of international evidence for the mispricing signal is related to a variety of risk-factor structures outside the U.S. The international evidence is likely to add little value if there is no proximity of risk-factor structures across the regions. For investigations of the risk-factor structure of international stock returns, see Rouwenhorst (1999), Griffin (2002), Griffin et al. (2010), Hou et al. (2011), Fama and French (2012), Fama and French (2017), and Bartram and Grinblatt (2018).

We find that there is only a little gain in performance of the mispricing strategy in the U.S. when the strategy's estimation sample is extended from the U.S. stocks to international stocks. The profitability of the mispricing strategy in the other regions, however, improves when we extend the estimation sample from the U.S. stocks to stocks in the respective regions. The mispricing of stocks estimated on historical data in the U.S. captures most of the predictability of stock returns outside the U.S. Taken together, our results imply that both local and global drivers of stock returns are important.

Our paper is structured as follows. We start with the data and methodology description in Section 1. The profitability of the mispricing strategy estimated in the U.S. is examined in Section 2 and the value of international evidence in Section 3. Finally, in Section 4, we examine the effect of transaction costs on the profitability of the mispricing strategy.

1. Data and methodology

1.1. Data

Our source of the accounting and market data for the U.S. is the Merged CRSP/Compustat database from the Wharton Research Data Service (WRDS). The sample spans from 1963 to 2018 period and contains all NYSE, AMEX, and NASDAQ common stocks (CRSP share code 10 or 11). The returns are adjusted for delisting following guidance as in Hou et al. (2018).[7]

The international data are sourced from Reuters Datastream. They are filtered following Ince and Porter (2006), Griffin et al. (2010) and Lee (2011). We manually check the names of the shares in the database for over 100 expressions describing their share class. Only the primary quotes of the ordinary shares of the companies are retained, with few exceptions where fundamental data in Datastream are linked to other share classes.[8] Real estate investment trusts (REITs) are excluded from the sample. All the international returns and financial statements are converted to U.S. dollars. The daily returns are deleted for days when the stock market is closed in a given country. The quality of data is further improved with procedures described in Tobek and Hronec (2018) and covered in Appendix A. Tobek and Hronec (2018) study the implications of the choice of the fundamental database on the measurement of performance of individual fundamental anomalies. They show that the statistical significance of the individual anomalies varies across Datastream and Compustat. The research inference can therefore change when a different fundamental database is used. The differences across the databases are mainly due to imperfect historical

[5] Frazzini et al. (2012) demonstrated that real-life transaction costs for large portfolio managers are much lower than assumed by academics. The transaction costs can be further lowered by appropriately optimized portfolio rebalancing.

[6] Note that many anomalies have been individually studied in the international markets. For examples of studies investigating cross-sectional predictability of individual signals outside the U.S. see Rouwenhorst (1998), McLean et al. (2009), Chui et al. (2010), Lam and Wei (2011), Barber et al. (2013), Titman et al. (2013), and Watanabe et al. (2013). The goal here is not the study of performance of the anomalies outside of the U.S., but rather the use of international historical performance of the anomalies to better select anomalies that are likely to outperform in the future.

[7] If the delisting is on the last day of the month, returns over the month are used. The relevant delisting return is then added as a return over the next month. Delisting return (DLRET) from monthly file is used if it is not missing. $(1 + ret_{cum})^*(1 + DLRET_d) - 1$ is used if it is missing, where ret_{cum} is the cumulative return in the given month of delisting and $DLRET_d$ is delisting return from the daily file. Lastly, the gaps are filled with $(1 + ret_{cum})^*(1 + DLRET_{avg}) - 1$, where $DLRET_{avg}$ is the average delisting return for stocks with the same first digit of delisting code (DLSTCD).

[8] We follow the description in Griffin et al. (2010) on the classification of common shares.

Table 1
Number of stocks in cross-section.

	mean	min	max
U.S.	1100	647	1734
Japan	750	534	1079
Asia Pacific	431	226	712
Europe	691	413	1044
Global	2069	647	4058

Regional decomposition the number of stocks in the sample from January 1995 to December 2018. The sample consists of stocks within the top 95% of the overall capitalization of all stocks in each region at the end of the previous month and within the top 95% of the overall dollar trading volume over the previous 12 months of all stocks in each region. All the stocks are further required to have a price larger than $1 ($0.1 for Asia Pacific) at the end of the previous month.

fundamental coverage. Studies of the aggregated performance of the anomalies, however, do not suffer from these problems. Therefore, our analysis is not impacted.

The sample includes 23 developed countries. The countries are sorted into four regions: the U.S.; Europe (E) — Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and the United Kingdom; Japan (J); and Asia Pacific (AP) - Australia, New Zealand, Hong Kong, and Singapore.

Another important source of our data for the anomalies is the Institutional Brokers' Estimate System (I/B/E/S), which is obtained from WRDS. I/B/E/S is merged on Datastream directly as it is one of the databases provided by Thompson Reuters and Datastream includes the respective tickers in its static file. The merger with CRSP is done indirectly through CUSIPs. The databases are merged on eight-digit CUSIP and then on six-digit CUSIP if unsuccessful. The success of the merger is checked manually by comparing the quoted tickers on the exchanges with the actual names of the companies. All the variables in I/B/E/S are transformed to U.S. dollars with original Reuters exchange rates, which are provided by WRDS.

We focus on a very liquid universe of stocks. This universe covers only stocks that are both (a) within the top 95% of the overall capitalization of all stocks in each region at the end of previous month and (b) within the top 95% of the overall dollar trading volume over the previous 12 months of all stocks in each region.[9] All the stocks are required to have a price larger than $1 ($0.1 for Asia Pacific) at the end of the previous month. We select the size of the stock universe to provide an as liquid universe as possible while retaining a reasonable number of stocks for the estimation.

Our focus on the liquid universe of stocks makes the findings more realistic. The stocks with small capitalization (micro-caps) account for only a small fraction of the overall capitalization of the market and often cannot be traded at significant volumes due to their high illiquidity.

Table 1 shows the average, minimum, and the maximum number of stocks in the cross-section of the individual regions. There are on average about 1100 stocks in the U.S. that satisfy the inclusion criteria for the liquid universe. The average number of stocks satisfying the criteria is even smaller in the other regions. The average capitalization of stocks in the liquid universe after July 1995 is $12 billion in the U.S., $12 billion in Europe, $5 billion in Japan, and $5 billion in Asia Pacific. The average size of the stocks in the sample is therefore quite balanced over the regions.

1.2. Anomalies

The sample includes 153 anomalies published in the literature. There are 93 fundamental, 11 I/B/E/S, and 49 market friction anomalies in the sample. The anomalies come almost exclusively from the top finance and accounting journals. Fig. 1 shows the number of the published anomalies over time. It also shows the number of anomalies whose in-sample period in their respective studies has ended. The number of anomalies gradually increases over time without any apparent jumps. The full list of the anomalies is provided in Appendix B. We primarily include anomalies described in Harvey et al. (2016), McLean and Pontiff (2016) and Hou et al. (2018). We focus on anomalies that are valid in the cross-section of stocks to be able to form long-short portfolios out of them. Any anomaly that is specific to the U.S., and which therefore cannot be constructed outside the U.S., are excluded.[10]

[9] Trading volume data are often not available before year 2000 for international stocks and before 1980 for NASDAQ stocks. A stricter condition on capitalization being within the top 90% of the overall capitalization of all stocks is used instead.

[10] Examples of excluded are anomalies are those that are.

- based on quarterly fundamental data (since there is only short and problematic coverage internationally)
- connected to hand-collected data in the U.S. such as IPOs, SPOs, and mergers.
- requiring segment information and NBER data.

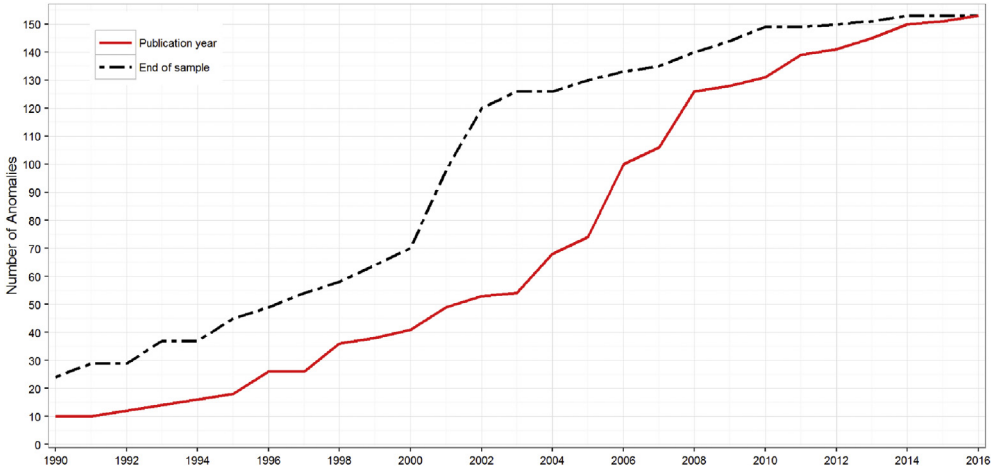

Fig. 1. Number of the published anomalies over time. The anomalies documented in Harvey et al. (2016), McLean and Pontiff (2016) and Hou et al. (2018) are primarily selected. The solid line shows the cumulative number of the published anomalies over time. The dashed line shows the number of anomalies whose in-sample period in their respective studies ends before the given date.

We update the fundamental signals for each month with financial statement information from financial years ending at least six months prior. For trading information-based signals, such as market cap, we use the latest data available.[11]

Some anomalies, such as the Herfindahl Index of Hou and Robinson (2006), require classification of industries for individual firms. The choice in the original papers is mostly with respect to the Standard Industrial Classification (SIC). We sort industries into 19 groups according to the third level Datastream classification. The broader industry groups should make the results more robust and consistent across the data vendors. The industry classification in Datastream is available only from the static file, which means that only the latest values are available. Data vendors may slightly differ in the classification of individual firms over time because the differences between individual SIC categories are often subtle.

1.3. Mispricing strategy

In this subsection, we discuss the strategy that shrinks all the anomalies into a single mispricing signal ("mispricing strategy"). Lewellen et al. (2015) define the prediction problem as follows: the goal is to devise a forecasting method that predicts which stocks are likely to have the highest returns in the next month and which have the lowest based on stock characteristics (the cross-sectional anomalies). To do this, we regress monthly returns on individual stocks on their past characteristics. The future returns are then predicted from the latest available characteristics. We estimate the regressions by pooling all the available stock returns up to the date of portfolio formation. The past characteristics have to be available before the start of the measurement period of the returns. The characteristics are normalized to their cross-sectional quantiles within each region to reduce problems with outliers.

To summarize, we estimate the following equation

$$r_{it} = f(x_{i,t-1,1}, x_{i,t-1,2}, \dots, x_{i,t-1,M}) + \epsilon_{it}, \tag{1}$$

where r_{it} is return on stock i in month t and $x_{i,t-1,1}$ is the cross-sectional quantile of a given anomaly (characteristic) for stock i available before the start of month t. The returns are demeaned by subtracting average cross-sectional returns in every region-month. We first discuss a simpler case with linear $f()$. It is then extended to a more general structure using machine learning. The machine learning exercise follows Gu et al. (2020), who apply a suite of standard machine learning algorithms and show that they outperform the linear models in the U.S. Readers are referred to Gu et al. (2020) or any advanced machine learning textbook for a detailed theoretical description of the machine learning methods and only basic definitions are covered here.[12]

The machine learning methods have some benefits and some negatives. They provide better out-of-sample forecasts through a limitation of in-sample over-fitting. They also allow for a very general interaction between the explanatory variables. This general form, however, makes the fitted models hard to estimate and the estimates hard to interpret due to the black-box

- institutionally specific, such as, share turnover or effective tax rate. Some fundamental anomalies could not be implemented in Datastream as the required items are missing.

[11] We test the impact of the monthly updating of fundamental signals. Annual updating of the fundamental signals leads to identical conclusions.

[12] See, for example, Friedman et al. (2001) for the textbook treatment.

approach. The intractability of the estimates is not a large concern in this study since even the linear method becomes intractable given the number of exogenous variables. Variable importance is examined in Subsections 3.3 and 3.4.

A crucial part of the application of the machine learning methods is hyperparameter optimization or "tuning." Standard k-fold cross-validation does not respect temporal ordering, i.e., in $k-1$ out of k folds predictability evaluation on validation sample is done using observations from the future in training. Therefore we split the historical sample into subsequent mutually exclusive training and validation samples each time models are retrained. The validation sample consists of the most recent 11-year period at the point-in-time when the model is retrained each December. The training sample always starts with the beginning of the dataset and ends one month before the validation sample starts. It is therefore expanding over time.

We estimate the model using the training sample with various hyperparameters, and we measure its predictive performance using the root-mean-square error via the hyperparameter grid search on the validation sample. Hyperparameters with the best predictive performance are then selected for the estimation. As an example, the initial historical sample period is January 1963 to December 1994 and it is split into a training sample of 21 years (1963–1983) and validation sample of 11 years (1984–1994). All models are re-estimated at the end of each year. The training sample is expanded by one year and the length of the validation sample is kept the same.[13] The set of possible hyperparameters for individual machine learning methods is provided at the end of the specific model descriptions in the following subsections. The evolution of selected hyperparameters for individual models is in Appendix C.

1.3.1. Weighted least squares

In the benchmark model, we use weighted least square estimation to determine a linear approximation of the relationship in equation (1). That is, we estimate a weighted least squares regressions of the stock returns on the rescaled characteristics,

$$r_{it} = \beta_0 + \beta_1 x_{i,t-1,1} + \beta_2 x_{i,t-1,2} + \cdots + \beta_M x_{i,t-1,M} + \epsilon_{it}, \tag{2}$$

where the weight on individual observations is the inverse of the number of stocks in each time period and region. The weights are introduced to give equal importance to each time period. The weighting makes the moment conditions equivalent to the Fama and MacBeth (1973) regressions in Lewellen et al. (2015). The linear specification has already been applied in an international context in Jacobs and Müller (2018, 2020). It is therefore selected as a benchmark for the more complicated machine learning methods.[14]

1.3.2. Penalized weighted least squares

The linear regression model with many explanatory variables can overfit the realization of past data since it has many degrees of freedom. One way to reduce the overfitting is to introduce L1 and L2 penalties on the coefficients during the estimation. There are two hyperparameters involved: α and λ. $\alpha \in [0, 1]$ is used to assign relative weights to the L1 and L2 penalties and λ controls the magnitude of penalization. The case with $\alpha = 1$, i.e., just L1 penalty, is denoted as the least absolute shrinkage and selection operator (LASSO) and was introduced in Tibshirani (1996). Tuned hyperparameters are α, considering values in {0, 0.1, …, 0.9, 1} and λ, considering values in {0.001, 0.0001}.

The regression tree family of methods is easy to estimate and requires a few specified hyperparameters. One such tree is depicted in Fig. 2. The decision tree consists of nodes (the round-edged boxes) and outcomes (sharp-edged boxes). The outcomes are in percentage return per month.[15] The tree starts with a decision on whether a given stock is within the smallest 40% of stocks in the cross-section. The decision can then continue to the split based on the book-to-market ratio. The depicted tree is of depth 3, which is the maximum number of nodes in the longest branch. The tree allows for arbitrary cross-effects between the variables up to the (depth - 1) degree. We deal mainly with relatively shallow trees. The shallow trees are nonetheless able to capture various important interactions between the explanatory variables. Random forest and gradient boosting regression trees are based on a combination of the individual trees. These methods cannot be easily visualized but they lead to better out-of-sample forecasting performance relative to simpler regression trees.

1.3.3. Random forest

Random forest is one of the most widely used ensemble tree methods. It combines forecasts from the individual decision trees that are based on subsamples of the training data. Explanatory variables are also subsampled in the individual trees to increase variety among the individual forecasts. Random forest is frequently among the top 10% of best-performing machine learning methods in various competitions and it is therefore a very robust method that is powerful in most of the settings. It requires only a few specified hyperparameters. The specification of the hyperparameters is not very important for its performance. It can therefore be used almost out-of-box. This is a large benefit with respect to neural networks where performance heavily depends on the model specification. The largest downside is that its estimations are time-consuming.

[13] The second historical sample period starts in January 1963, ends in December 1995 and it is split into a training sample of 22 years (1963–1984) and a validation sample of 11 years (1985–1995).

[14] Capitalization-weighted regressions as in Green et al. (2017) have also been tried. The capitalization-weighting puts lower weight on small cap stocks and is more suited for value-weighted portfolios. The weighting did not outperform the selected method and the results are therefore not reported here.

[15] The numbers are arbitrary and do not reflect real data.

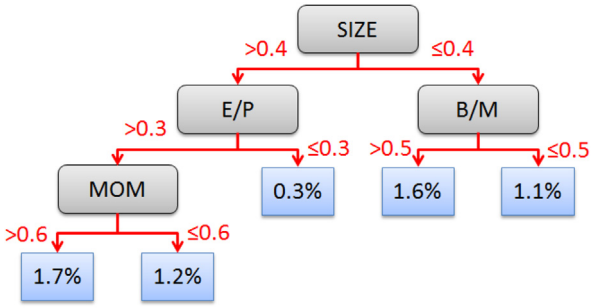

Fig. 2. Decision tree.

Tuned hyperparameters are *Number of trees*, considering values in {200, 300, 400, 500, 600} and *Max depth of the tree*, considering values in {1, …, 8, None}, where *None* means unlimited depth of the tree.

Other hyperparameters are fixed. The trees use randomly selected 50% of the overall training observations and the square root of the overall available explanatory variables. We employ a minimum node size of 0.1%, which is large enough to limit over-fitting but small enough to allow the method to approximate the true expected returns on stocks.[16]

1.3.4. Gradient boosting regression trees

The gradient boosting regression trees (GBRT) of Friedman (2001) rely on a different way of combining the regression trees than random forest. All the trees in a random forest are chosen independently, whereas they are selected in a dependent fashion in GBRT. The idea is to estimate a tree and use only a fraction of its fit for forecasts. The next iterations we conduct proceed on residuals of the dependent variable after we remove the fraction of the fitted values in the previous iteration. Shrinkage of the individual predictions guarantees that the learning can correct itself if the fitted values are selected suboptimally in some iterations. The fraction of individual predictions that is retained for the forecast is called a learning rate. The number of the learning iterations, given the learning rate, then determines how closely the particular realization of the sample from the whole population (the training sample) is over-fitted. A selection of fewer iterations reduces the risk of over-fitting (estimation error) but decreases the overall fit of the estimation (i.e., introduces an approximation error). It is therefore important to select the number of iterations with optimal estimation and approximation error trade-off. One way to do this is to rely on cross-validation. The method requires a specification of the learning rate, number of iterations (trees), and the maximum depth of the trees.

We conduct our analysis with a fast version of the gradient boosting — extreme gradient boosting (XGBOOST) of Chen and He (2017). The reason for this is that it is ten times faster to estimate and thus requires far less computational power. Gu et al. (2020) benchmark the different machine learning methods and only neural networks provide significantly better forecasts than GBRT. GBRT is therefore a good candidate for our empirical application and it captures most of the gains from the machine learning methods over the standard finance methods.

Tuned hyperparameters are *Number of trees*, considering values in {50, 100, 200, 300, 400, 500}, *Maximum depth of the trees*, considering values in {1, …, 9} and *learning rate*, considering values in {1%, 2.5%, 5%, 10%}.

1.3.5. Neural networks

Arguably the most powerful machine learning method of today is (deep) neural networks. Gu et al. (2020) show that they outperform any other method if they are optimally specified. The neural networks are a very flexible tool that encompasses many specifications.[17] The flexibility is also their largest disadvantage as it requires long experimentation and possible over-fitting of the sample.

Sequential neural networks consist of layers of neurons with information flowing between the layers in only one direction, from the input layer to the output layer. The information is fed in batches consisting of n sample points. Processing of the full training sample is called an epoch. The speed of change in the estimated parameters with new processed batches is determined through the learning rate. It is often an advantage to slow the learning rate over time to allow for the capture of finer details. We estimate the neural networks with back-propagation, along with a stochastic gradient descent version with adaptive moment estimation called Adam.[18]

In Fig. 3, we plot one of the neural network specifications corresponding to two architectures: *NN1_wide* or *NN1_narrow*. They are based on three layers. The initial layer has 150 neurons. The second hidden layer also has 150 neurons. Overall, six architectures are considered during the tuning of the hyperparameters and all of them have only one neuron in the last output layer.

[16] Ignoring this parameter completely, and leaving unlimited node size, leads to almost identical results. It is thus not an important assumption.

[17] A linear regression is the simplest specification.

[18] See Kingma and Ba (2014).

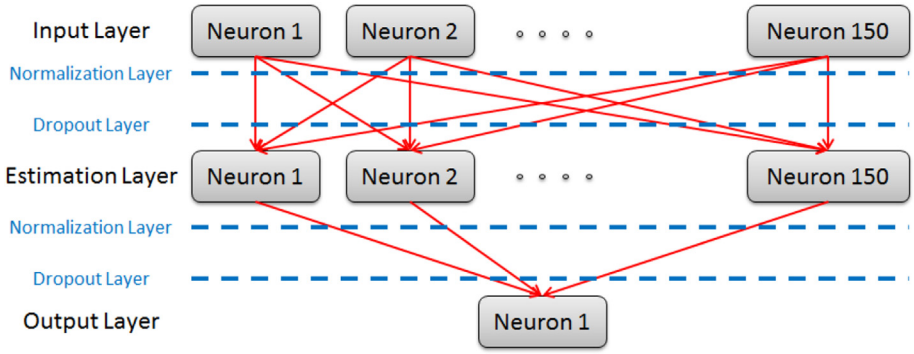

Fig. 3. Neural network.

- *NN1_wide*: One hidden layer with 150 neurons each.
- *NN1_narrow*: One hidden layer with 32 neurons respectively.
- *NN2_wide*: Two hidden layers with 150 neurons each
- *NN2_narrow*: Two hidden layers with 32 and 16 neurons respectively
- *NN3_wide*: Three hidden layers with 150, 150, and 100 neurons in the 1st, 2nd, and 3rd hidden layer.
- *NN3_narrow*: Three hidden layers with 32, 16, and 8 neurons in the 1st, 2nd, and 3rd hidden layer.

Narrow architectures are three out of the five neural network specifications used by Gu et al. (2020) and *wide* architectures are alternatives with a greater number of neurons offering higher model capacity. The input layer and all hidden layers use a rectified linear unit (ReLU) activation function while the last layer uses a linear activation. Input into each layer is batch normalized.

Additional fixed hyperparameters are a batch size of 256, maximum number of epochs equal to 25, and betas of 0.9 and 0.999 used in the Adam optimization.

Given the high model capacity of neural networks, regularization is of paramount importance. We employ dropout, early stopping, and ensembles for the purpose of regularization. For the *dropout* rate, we consider values of 0.001, 0.01, and 0.1. Meta-parameter in early stopping callback is called patience and it determines when to stop the learning process conditional on the lack of the validation-based mean squared loss improvement in consecutive epochs. It is fixed at four. Finally, another callback called reduce learning rate on plateau is used to reduce the learning rate by a factor of two when the validation-based mean squared loss stops decreasing from one epoch to another. We produce the final forecast from an ensemble of five estimated neural networks with different initial random seeds. The combination forecast leads to a great improvement in the performance of the mispricing strategy based on the neural networks.

1.4. Portfolio construction

The portfolios are constructed from the sorts of the predicted returns in the individual regions. They are constructed as long-short and self-financing. In the long leg of the strategy, stocks are purchased in the upper quintile of the predicted next month's returns. In the short leg of the strategy, stocks short sold in the bottom quintile of the predicted next month's returns. The quintile breakpoints are selected to provide more robust results relative to decile breakpoints. The portfolios are rebalanced every month based on signals from the end of the previous month. The portfolio returns correspond to an investable strategy that holds $1 in cash, invests $1 in the stocks that are likely to have the largest return in the next month, and shorts $1 worth of stocks that are likely to have the smallest return in the next month. The portfolios start in January 1995, unless stated otherwise.

A global strategy invests in stocks from all four regions. The global strategy is again based on stocks in the extreme quintiles of the predicted returns in the individual regions.

1.5. Liquidity measures

We use liquidity proxies to estimate transaction costs associated with investing into the mispricing strategy portfolios. The proxies are the Gibbs proxy of Hasbrouck (2009), the closing quoted spread proxy of Chung and Zhang (2014), and the VoV(% Spread) of Fong et al. (2017). They are defined in Appendix D.

We select these proxies to capture a fixed component of transaction costs and ignore the variable component that measures the price impact of larger orders. The variable component is very volatile and depends on the precise trade execution algorithm of each asset manager. The large capitalization universe of stocks reduces concerns about the variable component and it should be possible to avoid any execution costs altogether through the use of limit orders.

All of the proxies have some missing observations. The missing observations are backfilled from other proxies. The quoted spread is used first for the backfilling, followed by the VoV(% Spread), and the remaining missing observations are backfilled with the Gibbs proxy. Less than 0.02% of the observations are missing in all the three proxies and these observations are filled by 5% transaction costs.

2. Profitability of the mispricing strategy estimated in the U.S.

In this section, we examine the performance of the mispricing strategy worldwide. Jacobs and Müller (2018) show that the mispricing strategy estimated with least squares leads to higher returns in both absolute terms and on a risk-adjusted basis relative to the mixing of portfolios on individual anomalies. Gu et al. (2020) document that the more sophisticated machine learning methods provide higher out-of-sample predictability relative to the least squares method in the U.S. We extend the machine learning methods to the international sample to determine whether their benefits persist outside the U.S.

Table 2 presents the mean returns on portfolios created based on the mispricing strategy. The regressions of stock returns on their characteristics are fit on data available up to December each year and the future stock returns are then predicted with the latest available characteristics for each of the next 12 months. The regressions are estimated with weighted least squares (WLS), penalized least squares (PWLS), gradient boosting regression trees (GBRT), random forests (RF), and neural networks (NN). The estimates in Table 2 are based on the U.S. data going back to January 1963. Panel A in Table 2 shows the characteristics of returns on self-financing long-short portfolios. The long-short quintile portfolios invest in stocks in the top quintile of the predicted future returns and short-sell stocks in the bottom decile of the predicted returns. The reported portfolio returns are in percent per month and are from January 1995 to December 2018.

Table 2
Performance of the mispricing strategy estimated in the U.S.

	Equal-weighted					Value-weighted				
	U.S.	Europe	Japan	AP	Global	U.S.	Europe	Japan	AP	Global
Panel A: Long-short Portfolios										
Weighted Least Squares										
Mean	0.513	0.694	0.563	0.434	0.560	0.328	0.728	0.518	0.515	0.476
t-stat	1.868	2.680	2.855	1.998	2.487	1.530	3.408	2.818	1.824	2.427
Sharpe Ratio	0.429	0.814	0.724	0.431	0.725	0.314	0.821	0.526	0.441	0.573
Max Drawdown	−33.84	−29.30	−22.61	−30.96	−23.21	−34.81	−24.16	−25.74	−54.95	−22.73
Penalized Weighted Least Squares										
Mean	0.651	0.798	0.809	1.048	0.807	0.496	0.630	0.614	0.761	0.623
t-stat	2.538	2.751	3.632	4.483	3.652	2.253	2.646	2.452	2.455	2.957
Sharpe Ratio	0.495	0.788	0.877	0.932	0.929	0.397	0.629	0.553	0.581	0.679
Max Drawdown	−36.13	−31.91	−25.33	−36.27	−26.05	−30.09	−28.56	−34.03	−60.19	−26.28
t-stat wrt WLS	1.104	0.785	1.317	2.888	2.184	1.283	−0.713	0.495	1.394	1.147
Gradient Boosting Regression Trees										
Mean	1.074	1.024	0.950	0.907	1.020	0.920	0.766	0.882	0.170	0.739
t-stat	3.728	3.616	4.542	4.126	4.657	3.201	3.100	3.336	0.837	3.732
Sharpe Ratio	0.870	1.100	1.108	0.974	1.415	0.728	0.749	0.754	0.175	0.954
Max Drawdown	−42.42	−31.07	−18.11	−19.11	−25.72	−45.22	−32.55	−21.42	−42.23	−25.61
t-stat wrt WLS	2.021	1.931	2.013	2.258	2.710	1.979	0.198	1.341	−1.197	1.305
t-stat wrt PWLS	1.663	1.818	0.718	−0.656	1.419	1.458	0.872	0.879	−2.004	0.636
Random Forest										
Mean	1.034	1.103	0.994	1.036	1.049	0.995	0.629	0.888	0.270	0.791
t-stat	3.387	4.017	4.096	4.534	4.590	3.349	2.374	3.455	1.126	3.663
Sharpe Ratio	0.875	1.247	1.182	1.109	1.440	0.894	0.632	0.773	0.252	1.030
Max Drawdown	−36.77	−27.01	−32.20	−20.63	−20.34	−39.88	−35.09	−21.57	−49.01	−19.23
t-stat wrt WLS	2.566	2.417	1.880	2.601	3.405	2.563	−0.459	1.292	−0.805	1.768
t-stat wrt PWLS	2.031	2.728	0.620	−0.048	1.803	1.917	−0.006	0.737	−1.454	0.944
t-stat wrt GBRT	−0.253	1.193	0.262	1.522	0.355	0.463	−1.608	0.030	0.742	0.610
Random Forest										
Mean	1.327	1.207	0.667	1.100	1.079	1.315	0.897	0.740	0.873	1.000
t-stat	3.900	2.877	1.510	2.709	3.309	3.674	2.095	1.528	2.493	3.008
Sharpe Ratio	0.820	0.731	0.354	0.574	0.757	0.858	0.548	0.403	0.506	0.713
Max Drawdown	−53.40	−68.86	−59.98	−59.46	−56.01	−57.71	−72.33	−59.92	−57.25	−60.16
t-stat wrt Mkt	2.668	3.308	2.796	2.449	3.748	2.087	1.059	2.601	0.540	2.399
t-stat wrt WLS	2.012	1.151	0.631	1.184	2.055	1.827	−1.157	0.835	−2.006	1.093
t-stat wrt PWLS	1.711	1.316	0.154	−1.075	0.973	1.747	−0.698	0.880	−2.394	0.837
t-stat wrt GBRT	−0.888	1.628	−0.420	0.099	−0.616	−0.104	−1.841	0.176	−0.479	0.162
Neural Networks										
Mean	1.361	1.150	0.733	1.278	1.124	1.397	1.005	0.624	1.287	1.138
t-stat	4.285	2.941	1.767	2.669	3.409	4.644	2.594	1.354	3.069	3.670

(*continued on next page*)

Table 2 *(continued)*

	Equal-weighted					Value-weighted				
	U.S.	Europe	Japan	AP	Global	U.S.	Europe	Japan	AP	Global
Sharpe Ratio	0.867	0.703	0.424	0.557	0.786	0.962	0.637	0.360	0.618	0.835
Max Drawdown	−51.78	−65.05	−56.26	−68.94	−56.40	−47.44	−62.53	−56.90	−63.85	−52.61
t-stat wrt Mkt	3.325	3.004	4.051	3.529	4.440	3.134	1.783	2.273	2.832	3.820
t-stat wrt WLS	4.014	0.768	1.804	2.859	3.881	3.599	−0.446	0.376	0.435	2.708
t-stat wrt PWLS	3.837	0.919	1.458	0.106	2.830	3.025	0.135	0.609	−0.162	2.293
t-stat wrt GBRT	−0.357	0.190	0.186	1.481	0.255	0.430	−0.384	−0.460	2.045	1.728
t-stat wrt RF	0.377	−1.080	0.400	1.347	0.809	0.655	0.927	−0.566	2.231	1.823
Panel C: Short-only Component of the Mispricing Strategy										
Weighted Least Squares										
Mean	0.574	0.404	0.002	0.489	0.349	0.677	0.327	0.028	0.720	0.394
t-stat	1.365	0.898	0.004	0.940	0.881	1.766	0.776	0.067	1.497	1.097
Sharpe Ratio	0.298	0.227	0.001	0.210	0.216	0.427	0.194	0.017	0.339	0.271
Max Drawdown	−79.10	−76.00	−80.26	−74.83	−71.01	−73.01	−71.02	−75.00	−65.55	−69.68
Penalized Weighted Least Squares										
Mean	0.474	0.308	−0.169	0.220	0.199	0.503	0.359	−0.115	0.553	0.266
t-stat	1.143	0.687	−0.367	0.399	0.500	1.316	0.904	−0.268	1.141	0.731
Sharpe Ratio	0.234	0.165	−0.094	0.095	0.119	0.304	0.211	−0.070	0.260	0.177
Max Drawdown	−77.60	−77.13	−84.23	−72.45	−69.58	−75.86	−71.94	−73.83	−64.68	−71.39
t-stat wrt WLS	−1.320	−1.204	−1.552	−2.346	−2.478	−2.214	0.354	−1.410	−1.754	−1.746
Gradient Boosting Regression Trees										
Mean	0.331	0.113	−0.238	0.189	0.088	0.408	0.278	−0.168	0.722	0.252
t-stat	0.812	0.238	−0.519	0.347	0.218	1.152	0.692	−0.401	1.561	0.725
Sharpe Ratio	0.168	0.059	−0.132	0.083	0.053	0.259	0.161	−0.105	0.365	0.177
Max Drawdown	−76.71	−79.98	−87.35	−80.18	−71.48	−71.54	−73.19	−81.03	−64.23	−67.36
t-stat wrt WLS	−1.757	−3.043	−2.028	−2.872	−3.064	−2.181	−0.475	−1.633	0.015	−1.509
t-stat wrt PWLS	−1.254	−2.625	−0.677	−0.275	−1.518	−0.841	−1.209	−0.540	1.034	−0.218
Random Forest										
Mean	0.292	0.104	−0.327	0.064	0.030	0.320	0.268	−0.148	0.603	0.209
t-stat	0.656	0.221	−0.724	0.116	0.071	0.874	0.664	−0.372	1.252	0.585
Sharpe Ratio	0.144	0.055	−0.190	0.028	0.018	0.200	0.158	−0.093	0.295	0.143
Max Drawdown	−82.92	−77.15	−87.93	−81.12	−73.80	−77.32	−72.92	−80.36	−69.32	−70.95
t-stat wrt WLS	−2.598	−2.867	−2.949	−3.772	−4.013	−2.974	−0.559	−1.599	−0.590	−2.213
t-stat wrt PWLS	−1.899	−2.602	−1.153	−1.397	−2.364	−1.538	−1.199	−0.259	0.259	−0.937
t-stat wrt GBRT	−0.459	−0.181	−1.165	−2.008	−1.324	−1.177	−0.179	0.232	−1.024	−0.904
Neural Networks										
Mean	0.233	0.156	−0.373	−0.089	−0.025	0.212	0.271	−0.078	0.459	0.126
t-stat	0.515	0.328	−0.840	−0.154	−0.059	0.558	0.676	−0.190	0.939	0.355
Sharpe Ratio	0.109	0.083	−0.202	−0.037	−0.014	0.126	0.158	−0.046	0.218	0.084
Max Drawdown	−85.74	−79.71	−89.61	−82.61	−78.68	−80.52	−72.92	−73.10	−68.13	−71.24
t-stat wrt WLS	−4.445	−2.926	−2.965	−4.773	−5.782	−4.477	−0.607	−0.979	−1.753	−3.458
t-stat wrt PWLS	−4.082	−2.992	−3.571	−3.141	−5.723	−3.753	−1.901	0.666	−0.765	−3.551
t-stat wrt GBRT	−0.830	0.776	−1.681	−3.867	−1.991	−1.439	−0.100	1.081	−1.524	−1.925
t-stat wrt RF	−0.760	0.884	−0.378	−2.060	−1.137	−0.806	0.047	0.606	−0.744	−1.289
Panel D: Performance on Risk-adjusted Basis										
Weighted Least Squares										
Mean Return	0.513	0.694	0.563	0.434	0.560	0.328	0.728	0.518	0.515	0.476
	1.868	2.680	2.855	1.998	2.487	1.530	3.408	2.818	1.824	2.427
CAPM Alpha	0.716	0.764	0.569	0.505	0.662	0.466	0.782	0.523	0.617	0.538
	2.794	3.615	3.394	2.378	3.634	2.105	4.139	2.791	2.709	2.950
FF5 Alpha	0.256	0.180	0.448	0.380	0.166	0.120	0.294	0.418	0.299	0.133
	1.303	1.278	2.877	2.178	1.278	0.701	1.861	2.267	1.340	0.898
Penalized Weighted Least Squares										
Mean Return	0.651	0.798	0.809	1.048	0.807	0.496	0.630	0.614	0.761	0.623
	2.538	2.751	3.632	4.483	3.652	2.253	2.646	2.452	2.455	2.957

(*continued on next page*)

Panel A in Table 2 documents the striking profitability of the mispricing strategy across all regions. Diversification over the four regions (in the global columns) further reduces the maximum drawdowns and increases the Sharpe ratios for the mispricing strategy.

The *t-stat wrt WLS*, *t-stat wrt PWLS*, *t-stat wrt GBRT*, and *t-stat wrt RF* rows provide t-statistics for difference in mean returns for the given two estimation methods. Both the tree-based methods and neural networks outperform simple least squares. In particular, neural networks outperform least squares in all the regions for both mean returns and risk-adjusted Sharpe ratios. Neural networks also have the smallest maximum drawdowns and investing in them is therefore the least risky. Neural networks and random forest also significantly outperform penalized least squares.

Table 2 *(continued)*

	Equal-weighted					Value-weighted				
	U.S.	Europe	Japan	AP	Global	U.S.	Europe	Japan	AP	Global
CAPM Alpha	0.894	0.909	0.813	1.083	0.922	0.651	0.721	0.620	0.853	0.726
	3.476	3.768	3.916	4.550	4.683	2.524	3.462	2.632	3.422	3.878
FF5 Alpha	0.489	0.266	0.603	0.873	0.458	0.370	0.185	0.401	0.531	0.330
	2.465	1.660	3.110	3.820	3.042	1.961	1.179	1.806	1.900	2.095
Gradient Boosting Regression Trees										
Mean Return	1.074	1.024	0.950	0.907	1.020	0.920	0.766	0.882	0.170	0.739
	3.728	3.616	4.542	4.126	4.657	3.201	3.100	3.336	0.837	3.732
CAPM Alpha	1.194	1.105	0.949	0.988	1.094	0.917	0.820	0.877	0.204	0.748
	4.504	4.600	5.058	4.973	5.970	3.394	3.559	3.685	0.983	4.186
FF5 Alpha	1.158	0.490	0.754	0.690	0.848	1.018	0.386	0.708	0.024	0.602
	4.930	2.634	4.057	4.050	5.595	4.103	1.847	2.990	0.127	3.881
Random Forest										
Mean Return	1.034	1.103	0.994	1.036	1.049	0.995	0.629	0.888	0.270	0.791
	3.387	4.017	4.096	4.534	4.590	3.349	2.374	3.455	1.126	3.663
CAPM Alpha	1.199	1.180	0.993	1.119	1.136	1.034	0.645	0.883	0.307	0.810
	4.288	5.229	5.119	5.688	6.096	3.724	2.669	4.034	1.316	4.243
FF5 Alpha	0.757	0.673	0.846	0.747	0.690	0.806	0.225	0.725	0.016	0.487
	3.601	4.088	4.475	4.813	5.253	3.539	1.187	3.596	0.075	3.172
Neural Networks										
Mean Return	1.128	0.994	1.107	1.367	1.149	1.185	0.734	0.702	0.828	1.012
	4.345	3.529	5.687	6.160	5.383	5.136	3.294	3.411	3.296	5.697
CAPM Alpha	1.376	1.078	1.111	1.402	1.259	1.277	0.775	0.703	0.847	1.068
	5.680	4.631	6.121	6.591	6.777	5.378	3.416	3.247	3.586	6.285
FF5 Alpha	0.907	0.547	0.913	1.201	0.802	1.106	0.405	0.531	0.842	0.843
	5.305	3.620	5.657	6.597	6.323	5.941	2.062	2.812	3.124	5.668

The table shows returns on quintile portfolios from mispricing strategy described in Subection 1.3. The estimation methods are weighted least squares (WLS), penalized weighted least squares (PWLS), gradient boosting regression trees (GBRT), random forests (RF), or neural networks (NN). The regressions are rerun at the end of each December using only anomalies that have been documented by that time and using hyperparameters selected based on the most recent 11-year validation sample available. The regressions are estimated only on the past U.S. data and the future returns are predicted in all the regions. Long portfolios are constructed by buying stocks in the top quintile of the predicted next month returns and short portfolios are constructed by short-selling stocks in the bottom quintile of the predicted next month returns. The portfolios are either value-weighted or equal-weighted. Panel A presents returns on long-short portfolios. The returns on the long-short portfolios are decomposed to long-only component in Panel B and short-only component in Panel C. Panel D presents the performance of the long-short portfolio adjusted for capital asset pricing model (CAPM) market risk factor and five Fama-French factors (FF5). The out-of-sample performance is observed in the U.S., Europe, Japan, and Asia Pacific. Estimation samples are expanding annually and correspond to periods (Jan 1963–Dec 1994) up to the period of (Jan 1963–Dec 2018). The reported returns are for January 1995 to December 2018 period and are in percentage points. The standard errors in t-statistics are adjusted for heteroskedasticity and autocorrelation with Newey-West adjustment for up to 12 lags. *t-stat wrt WLS, PWLS, GBRT, RF* rows provide t-statistics for difference in mean returns between the two corresponding estimation methods.

To understand the source of outperformance it is important to focus on the differences between these methods. The key differences between the most successful methods and more traditional methods are penalization, the interaction of predictive variables, and non-linearity. Penalization is used to solve multi-collinearity and overfitting problems. The benefit of penalization is best observed when comparing results from penalized least squares compared to the simple least squares. Gradient boosting regression trees, random forests, and neural networks further benefit from the non-linearity and possible interaction of predictive variables. The results in Panel A in Table 2 provide some support for the benefits of adding penalization to least squares and introducing non-linearity and variable interactions in the more complex methods.

The anomalies we use are different from anomalies used by Gu et al. (2020). Furthermore, there are no macroeconomic predictors, industry dummies, or explicit interactions between stock-level characteristics and factors. The maximum number of variables[19] we us is 153, which is smaller than the 920 variables used by Gu et al. (2020). Their variables are not filtered over time based on the publication date. Even though there is a different setup in both of these studies, the profitability of the mispricing strategy in the U.S. region is comparable with the profitability of the machine learning portfolios constructed by Gu et al. (2020). Our results, which are obtained in different empirical setting provide support for Gu et al.'s (2020) findings.

In Fig. 4, we plot the cumulative returns on the neural network estimation method of the mispricing strategy in Panel A in Table 2. We find that there is a small drop in return profitability around 2003 in the U.S. The mispricing strategy is the least profitable in the European region on an equal-weighted basis.

2.1. Long-only and short-only components of the strategy

Short-selling can be connected to large costs and sometimes even outright impossible. That is why it might not be possible

[19] Only anomalies published before the time of estimation are considered.

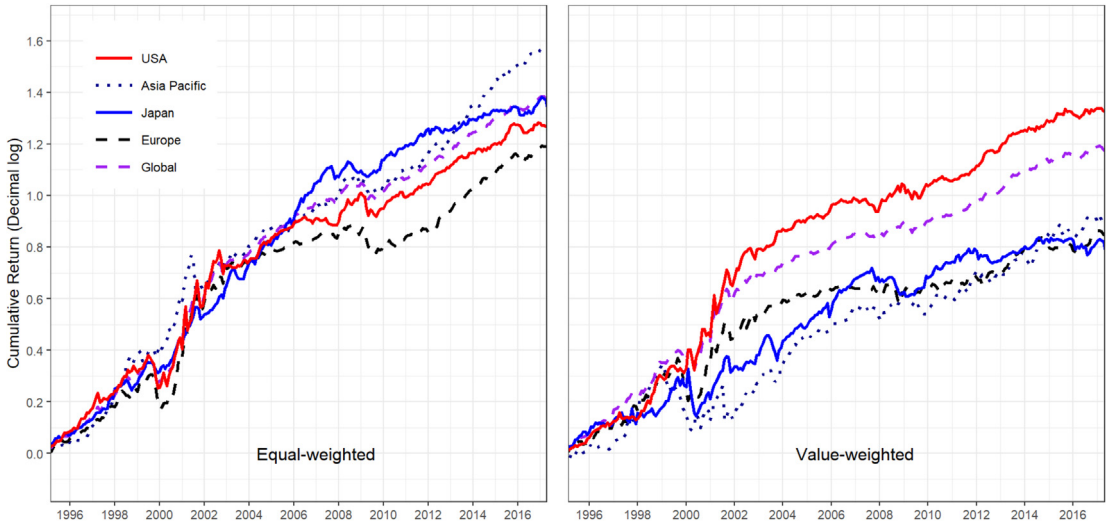

Fig. 4. Cumulative returns on the mispricing strategy using neural networks. The figure shows cumulative returns for the mispricing strategy as shown in Panel A in Table 2 that is estimated on the individual stocks from the U.S. using neural networks. The returns are presented in decimal logarithms. One on the vertical scale therefore corresponds to 1000% return on the initial investment.

to replicate the returns on the mispricing strategy in practice.[20] In order to determine the role of short-selling for the strategy's y, we decompose the results of the long-short portfolios in Panel A in Table 2 into the long-only and short-only components in Panel B and C. We also decompose the long-short returns separately for the individual machine learning methods. The long-only component can be compared to equal-weighted and value-weighted returns on the whole market comprised of the liquid stocks, as described in Subsection 1.1.

Panel B in Table 2 documents that the mispricing strategy is more profitable than the whole market in all regions. The long-only component is responsible for most of the returns on the mispricing strategy. The short-only component then mainly serves as a hedge that increases the Sharpe ratio and lowers the maximum drawdown. The annualized returns on the long-only component of the mispricing strategy employing neural networks are about 5% a year larger than returns on the market. The other machine learning methods also outperform the market.

The more advanced machine learning methods outperform simple least squares both on the long-side of the portfolio and on the short-side. To conclude, the out-performance of the mispricing strategy is robust to short-selling constraints. Even short-selling-constrained investors can therefore benefit from the strategy.

2.2. Risk-adjusted performance of the strategy

Thus far, our focus has been on the raw returns on the mispricing strategy without accounting for any risk factors. Panel D in Table 2 presents the performance of the long-short strategy as presented in Panel A in Table 2 after accounting for market returns and five Fama-French factors. The results in Panel D in Table 2 show that accounting for market returns have little impact on the performance of the strategy as the capital asset pricing model (CAPM) alpha is close to the mean returns for all the estimation methods. This is expected, since the portfolios are long-short and thus close to market neutral by construction.

The results are, however, very different when adjusting for the five Fama-French factors. Across the regions, the alphas are mostly insignificant for the weighted least squares portfolio but become significant at 5% level for penalized weighted least squares portfolio for all regions except Europe for the equal-weighted portfolios. Since penalization is the only difference, it clearly shows one part of the difference with respect to the traditional risk factors.

The difference between the mean returns and alphas is substantially smaller for the machine learning methods than for the linear estimation methods. This is true for the U.S. as well as internationally. The linear estimation methods therefore lead to the mispricing signal that is close to the traditional risk factors. On the other hand, due to the non-linear interactions between the predictive variables, the gradient boosting regression trees, random forests and neural networks are able to capture the predictive relationships that linear methods are not able to.

To conclude, the profitability of the mispricing strategy is significant even on a risk-factor-adjusted basis and this holds true in all the regions. Looking at the risk-adjusted performance of the mispricing strategy when employing different underlying models

[20] Short-selling constrains should not be a large issue on our liquid universe of stocks. Andrikopoulos et al. (2013) show that although some stocks cannot be short-sold in practice, focusing only on those that can be short-sold does not statistically diminish returns on eight quantitative strategies in the United Kingdom. They also show that short-selling costs are small at about 1% annually in the United Kingdom.

further clarifies the reasons for the superiority of the machine learning methods. Penalization and the non-linear interaction of the variables lead to returns that are unrelated to the traditional risk factors.

3. The role of international evidence

The evidence thus far documents that the mispricing strategy trained on the past data in the U.S. is profitable out-of-sample in all the regions. Can international data outside the U.S. be used to better train the strategy?

There are some arguments for the usefulness of international data. The international data increases sample size and therefore limits the possibility for data-mining and in-sample overfitting. The larger sample size also generally provides larger power to statistical tests, which should lead to the more precise selection of truly significant strategies. One crucial requirement for the tangible benefit of the new observations is that they are independent from the original observations. The international evidence extends the sample size mainly in the year 2000. The most recent data are also the most useful as financial markets change rapidly and the older data may not be relevant anymore.

There are, however, also some problems with the suitability of international evidence. The individual global regions have very different institutional settings. Bankruptcy laws, tax laws, investor protection, and accounting standards vary widely across the regions. The institutional differences can lower the usefulness of historical data outside the respective regions. The larger estimation sample improves forecasts through consistency. The consistency, however, works only if the underlying true drivers of stock returns are uniform over the regions, which is in no way guaranteed.

3.1. Locally-trained mispricing strategy

Previous machine learning evidence is based on predictive regressions estimated solely on data from the U.S. In this subsection, we first investigate whether estimating the predictive regressions in the respective regions is more suitable than estimating them only on data from the U.S. In the next subsection, we explore whether combining estimation samples from the individual regions can improve the profitability of the mispricing strategy.

Panel in Table 3 presents statistics on the portfolio returns of mispricing strategy introduced in Subsection 1.3 that is estimated on local samples of stocks. That is, for example, the portfolios in Japan are formed based on predicted next month returns from predictive regressions fitted on historical data from Japan. Table 3 also shows the differences in portfolio returns when using the predictive regressions estimated on data from the U.S. versus locally.

There is surprisingly only a small difference between the returns on strategies that are estimated using data from the U.S. in Table 2 and those that are estimated using data in the respective regions in Panel A in Table 3. One explanation for the similarity is that the sample size in the U.S. is already large enough to capture the true drivers of stock returns that are globally valid.

The performance of the mispricing strategy in Asia Pacific improves when the predictive regressions are estimated in Asia Pacific for all the machine learning methods apart from neural networks. However, there are only a few liquid stocks in Asia Pacific region historically. The local estimation sample is therefore composed of stocks with relatively a smaller market cap than in the U.S. sample. This different stock universe composition might help in tailoring the fitted relationships.

The performance of the mispricing strategy in Japan is notably worse than when estimated using the U.S. data. The explanation is again simple. Japan experienced a slow eruption of an asset price bubble at the beginning of the estimation sample, in the early 1990s. The estimated relationships that are valid for this specific period fare badly out-of-sample where the stock market dynamics go back to their normal state.

3.2. Globally-trained mispricing strategy

Panel B in Table 3 shows the mean returns and other performance statistics for the mispricing strategy when the future individual stock returns are predicted from regressions estimated on historical data that are not solely from the U.S. but from the global training sample comprised of stocks from the U.S., Japan, Europe, and Asia Pacific.

We find that there is no gain from adding international stocks to the local training sample in the U.S. Historical data in the U.S. is therefore completely sufficient for future predictions in the U.S. Profitability of the mispricing strategy in Europe improves with predictions based on the global training sample relative to the training sample from the U.S. The profitability in Japan also improves with the global training sample instead of from the U.S. only sample. The largest gains in profitability from extending the training sample outside the U.S. are in the Asia Pacific region.

To conclude, Table 3 provides mixed results on the value of international evidence. The region-specific settings are indeed an important determinant of stock return drivers. There is no gain for the U.S. investor seeking international evidence for the quantitative strategy. The larger statistical power, due to a larger sample, seems to be completely offset by the region-specific differences.

3.3. Variable importance

One of the disadvantages of the more complex machine learning methods is the difficulty in interpreting the resulting models due to the potentially high-dimensional and nonlinear interactions among variables. Our main goal is the superior out-of-sample

Table 3
Performance of the mispricing strategy estimated on the international data.

	Equal-weighted					Value-weighted				
	U.S.	Europe	Japan	AP	Global	U.S.	Europe	Japan	AP	Global
Panel A: Locally-trained Strategy										
Weighted Least Squares										
Mean	0.513	0.818	0.063	1.616	0.893	0.328	0.653	−0.113	0.714	0.584
t-stat	2.214	3.432	0.263	5.222	4.579	1.647	2.829	−0.403	2.473	3.111
Sharpe Ratio	0.429	0.787	0.060	1.073	1.024	0.314	0.573	−0.091	0.480	0.660
Max Drawdown	−33.84	−35.60	−62.92	−55.74	−26.74	−34.81	−33.01	−77.18	−45.19	−32.89
Diff mean wrt U.S.	0.000	0.125	−0.500	1.182	0.333	0.000	−0.074	−0.630	0.199	0.107
Diff t-stat wrt U.S.		0.724	−1.687	3.710	2.626		−0.358	−1.979	0.598	0.683
Penalized Weighted Least Squares										
Mean	0.651	0.819	0.466	1.640	0.889	0.496	0.651	0.135	0.880	0.658
t-stat	2.572	3.286	1.689	5.174	4.175	2.134	2.752	0.419	2.985	3.373
Sharpe Ratio	0.495	0.747	0.389	1.058	0.951	0.397	0.544	0.095	0.559	0.696
Max Drawdown	−36.13	−37.03	−62.61	−55.91	−27.62	−30.09	−32.61	−79.34	−47.59	−27.53
Diff mean wrt U.S.	0.000	0.021	−0.342	0.593	0.082	0.000	0.021	−0.479	0.119	0.035
Diff t-stat wrt U.S.		0.180	−1.102	2.169	1.463		0.175	−1.304	0.422	0.484
Gradient Boosting Regression Trees										
Mean	1.074	0.555	0.497	1.647	1.114	0.920	0.200	0.242	1.145	0.689
t-stat	4.464	2.429	1.830	7.497	6.523	3.731	0.934	0.822	4.289	3.890
Sharpe Ratio	0.870	0.489	0.420	1.535	1.404	0.728	0.174	0.176	0.837	0.772
Max Drawdown	−42.42	−48.38	−66.08	−29.00	−22.07	−45.22	−36.71	−74.31	−37.48	−26.04
Diff mean wrt U.S.	0.000	−0.468	−0.453	0.740	0.094	0.000	−0.566	−0.640	0.976	−0.050
Diff t-stat wrt U.S.		−1.697	−1.056	3.529	1.118		−2.364	−1.287	3.660	−0.404
Random Forest										
Mean	1.034	0.668	0.367	1.581	1.206	0.995	0.396	0.166	1.169	0.850
t-stat	4.173	2.717	1.241	8.258	7.020	4.066	1.675	0.571	4.242	4.877
Sharpe Ratio	0.875	0.563	0.299	1.675	1.584	0.894	0.342	0.124	0.884	1.029
Max Drawdown	−36.77	−47.62	−71.74	−21.65	−18.45	−39.88	−37.27	−72.45	−28.31	−21.92
Diff mean wrt USA	0.000	−0.435	−0.627	0.544	0.157	0.000	−0.234	−0.722	0.899	0.058
Diff t-stat wrt U.S.		−1.735	−1.286	2.790	2.349		−0.919	−1.619	3.107	0.536
Neural Networks										
Mean	1.128	0.925	0.755	1.315	1.321	1.185	0.700	0.418	0.711	1.101
t-stat	4.747	3.841	3.238	4.587	7.431	5.671	3.016	1.643	2.607	5.785
Sharpe Ratio	0.952	0.885	0.691	1.060	1.686	1.093	0.630	0.326	0.555	1.216
Max Drawdown	−25.69	−39.26	−47.98	−45.66	−18.19	−16.50	−35.36	−61.58	−50.49	−18.63
Diff mean wrt USA	0.000	−0.069	−0.352	−0.051	0.172	0.000	−0.034	−0.284	−0.117	0.090
Diff t-stat wrt U.S.		−0.479	−1.424	−0.151	3.674		−0.192	−1.008	−0.385	1.171
Panel B: Globally-trained Strategy										
Weighted Least Squares										
Mean	0.677	0.870	0.794	1.386	0.893	0.425	0.683	0.424	0.753	0.584
t-stat	2.446	3.165	3.454	6.278	3.944	1.907	3.017	1.922	2.935	2.767
Sharpe Ratio	0.514	0.978	0.791	1.185	1.024	0.357	0.695	0.359	0.546	0.660
Max Drawdown	−37.35	−26.27	−25.84	−26.54	−26.74	−40.62	−25.41	−48.21	−42.77	−32.89
Diff mean wrt U.S.	0.164	0.177	0.231	0.952	0.333	0.097	−0.045	−0.094	0.238	0.107
Diff t-stat wrt U.S.	1.225	1.319	1.040	4.187	2.626	0.525	−0.289	−0.417	1.124	0.683
Penalized Weighted Least Squares										
Mean	0.623	0.937	0.799	1.377	0.889	0.521	0.752	0.379	0.886	0.658
t-stat	2.116	3.130	3.385	5.637	3.643	2.255	3.360	1.484	3.458	3.152
Sharpe Ratio	0.438	0.981	0.765	1.187	0.951	0.399	0.750	0.308	0.672	0.696
Max Drawdown	−38.45	−29.32	−23.34	−27.87	−27.62	−31.57	−21.08	−49.26	−45.90	−27.53
Diff mean wrt USA	−0.028	0.139	−0.009	0.330	0.082	0.025	0.122	−0.235	0.126	0.035
Diff t-stat wrt U.S.	−0.376	2.189	−0.086	3.047	1.463	0.239	1.408	−1.655	0.740	0.484
Gradient Boosting Regression Trees										
Mean	0.922	1.134	0.885	1.772	1.114	0.646	0.779	0.608	1.173	0.689
t-stat	3.224	4.027	4.625	7.879	5.326	2.470	3.336	2.477	3.776	3.723
Sharpe Ratio	0.722	1.201	1.054	1.944	1.404	0.478	0.772	0.529	0.919	0.772
Max Drawdown	−36.12	−20.18	−23.82	−9.389	−22.07	−38.94	−28.14	−36.18	−40.84	−26.04
Diff mean wrt USA	−0.152	0.110	−0.065	0.865	0.094	−0.275	0.013	−0.274	1.004	−0.050
Diff t-stat wrt U.S.	−1.260	1.061	−0.419	5.048	1.118	−1.466	0.114	−1.239	3.678	−0.404

(*continued on next page*)

Table 3 *(continued)*

	Equal-weighted					Value-weighted				
	U.S.	Europe	Japan	AP	Global	U.S.	Europe	Japan	AP	Global
Random Forest										
Mean	1.032	1.127	1.027	1.825	1.206	0.795	0.799	0.587	1.266	0.850
t-stat	3.321	4.446	5.375	9.468	5.508	2.840	3.231	2.593	4.840	4.126
Sharpe Ratio	0.889	1.212	1.217	1.925	1.584	0.699	0.787	0.514	0.969	1.029
Max Drawdown	−30.95	−19.19	−21.56	−21.68	−18.45	−39.01	−23.14	−35.41	−31.47	−21.92
Diff mean wrt U.S.	−0.002	0.024	0.032	0.788	0.157	−0.200	0.170	−0.301	0.996	0.058
Diff t-stat wrt U.S.	−0.024	0.233	0.147	5.081	2.349	−1.735	1.264	−1.119	5.211	0.536
Neural Networks										
Mean	1.123	1.205	1.272	1.848	1.321	1.161	0.835	0.818	1.405	1.101
t-stat	4.037	4.563	6.777	7.964	6.083	3.913	3.409	3.998	5.196	4.971
Sharpe Ratio	0.944	1.345	1.488	1.779	1.686	0.969	0.774	0.663	1.048	1.216
Max Drawdown	−23.70	−23.12	−16.81	−23.71	−18.19	−23.71	28.61	−38.28	−33.07	−18.63
Diff mean wrt U.S.	−0.005	0.211	0.165	0.482	0.172	−0.024	0.102	0.116	0.577	0.090
Diff t-stat wrt U.S.	−0.069	2.973	2.017	5.717	3.674	−0.217	1.242	0.804	3.268	1.171

The table shows returns on quintile long-short portfolios from mispricing strategy described in Subsection 1.3. The estimation methods are least squares, penalized least squares, random forests, gradient boosting regression trees, or neural networks. The regressions are rerun at the end of each January using only anomalies that have been published by that time and using hyperparameters selected based on the most recent 11-year validation sample available. The historical predictive regressions are estimated on individual stocks from all the four covered regions: the U.S., Japan, Europe, and Asia Pacific. Predictive regressions in Panel A are estimated on data from region where they are used for prediction. Predictive regressions in Panel B are estimated globally on data from all the regions. The long-short portfolios are constructed by buying stocks in the top quintile of the predicted next month returns and short-selling stocks in the bottom quintile of the predicted next month returns. Estimation samples are expanding annually and correspond to periods (Jan 1963–Dec 1994) in the U.S., or (Jan 1990–Dec 1994) in other regions, up to the period of Jan 1963–Dec 2018). The reported returns are for January 1995 to December 2018 period and are in percentage points. The standard errors in t-statistics are adjusted for heteroskedasticity and autocorrelation with Newey-West adjustment for up to 12 lags. Diff mean rows present difference of average monthly long-short portfolio returns between strategy estimated on the global data minus mean returns when it is estimated only on the U.S. data as in Table 2. Diff t-stat presents t-statistic for the significance of the difference in mean returns.

Table 4
Spearman's correlation matrices for region-specific variable importance.

	Weighted Least Squares					Gradient Boosting Regression Trees					Neural Networks				
	Global	U.S.	Europe	Japan	AP	Global	U.S.	Europe	Japan	AP	Global	U.S.	Europe	Japan	AP
Global	1.000	0.552	0.610	0.564	0.576	1.000	0.737	0.712	0.634	0.705	1.000	0.516	0.314	0.385	0.290
U.S.	0.552	1.000	0.406	0.363	0.397	0.737	1.000	0.611	0.490	0.512	0.516	1.000	0.378	0.278	0.309
Europe	0.610	0.406	1.000	0.482	0.443	0.712	0.611	1.000	0.548	0.603	0.314	0.378	1.000	0.222	0.299
Japan	0.564	0.363	0.482	1.000	0.541	0.634	0.490	0.548	1.000	0.590	0.385	0.278	0.222	1.000	0.171
AP	0.576	0.397	0.443	0.541	1.000	0.705	0.512	0.603	0.590	1.000	0.290	0.309	0.299	0.171	1.000

The table shows Spearman correlation matrices between region-specific ranks of variable importance as in equation (3) for the mispricing strategy. The mispricing strategy is described in Subsection 1.3. The strategy combines signals through predictive regressions of individual stock returns on transformed characteristics. The historical predictive regressions are estimated on data from the individual regions using weighted least squares, gradient boosting regression trees, or neural networks. Estimation sample consists of period from January 1963 to December 2018 in the U.S. and January 1990 to December 2018 in other regions.

performance even at the cost of the inability to fully interpret all the variable interactions in the resulting models. That being said, inspired by Sirignano et al. (2016), Chen et al. (2019), and Horel and Giesecke (2019), we next examine the importance of individual variables. We define the variable importance VI_j for variable j as the elasticity of predicted (region-wise demeaned) returns to changes in the individual characteristics used as predictors, as follows:

$$VI_j = \sum_{t\in T}\sum_{i\in N_t}\left|\frac{\partial \hat{r}_{it}(x_{i,t-1,1}, x_{i,t-1,2}, \dots, x_{i,t-1,M})}{\partial x_j}\right|, \tag{3}$$

We calculate the variable importance VI_j for each characteristic in various settings. Fig. 5 shows the variable importance across regions for the twenty five most important variables globally for the mispricing strategy as described in Subsection 1.3. The most important predictors are all connected to high portfolio rebalancing costs. The most important fundamental signal is sales over price, which is a measure of the value of the stocks. The second most important fundamental signal, a measure of R&D spendings, is important only in the U.S. where it was identified.

Table 4 shows the Spearman's rank correlation of variable importance scores across the regions under various forecasting methods. We find that there is great heterogeneity in the ranks of variable importance across the regions. This results provide insight on the limited value of extending the estimation sample from the U.S. to international stocks. More importantly, the predictions from the U.S. perform as well as the predictions from the other regions despite having only loosely connected variable importance.

We also find that there are pronounced differences in variable importance under different forecasting methods. The Spearman's rank correlation coefficient between variable importance scores using neural networks and gradient boosting regression

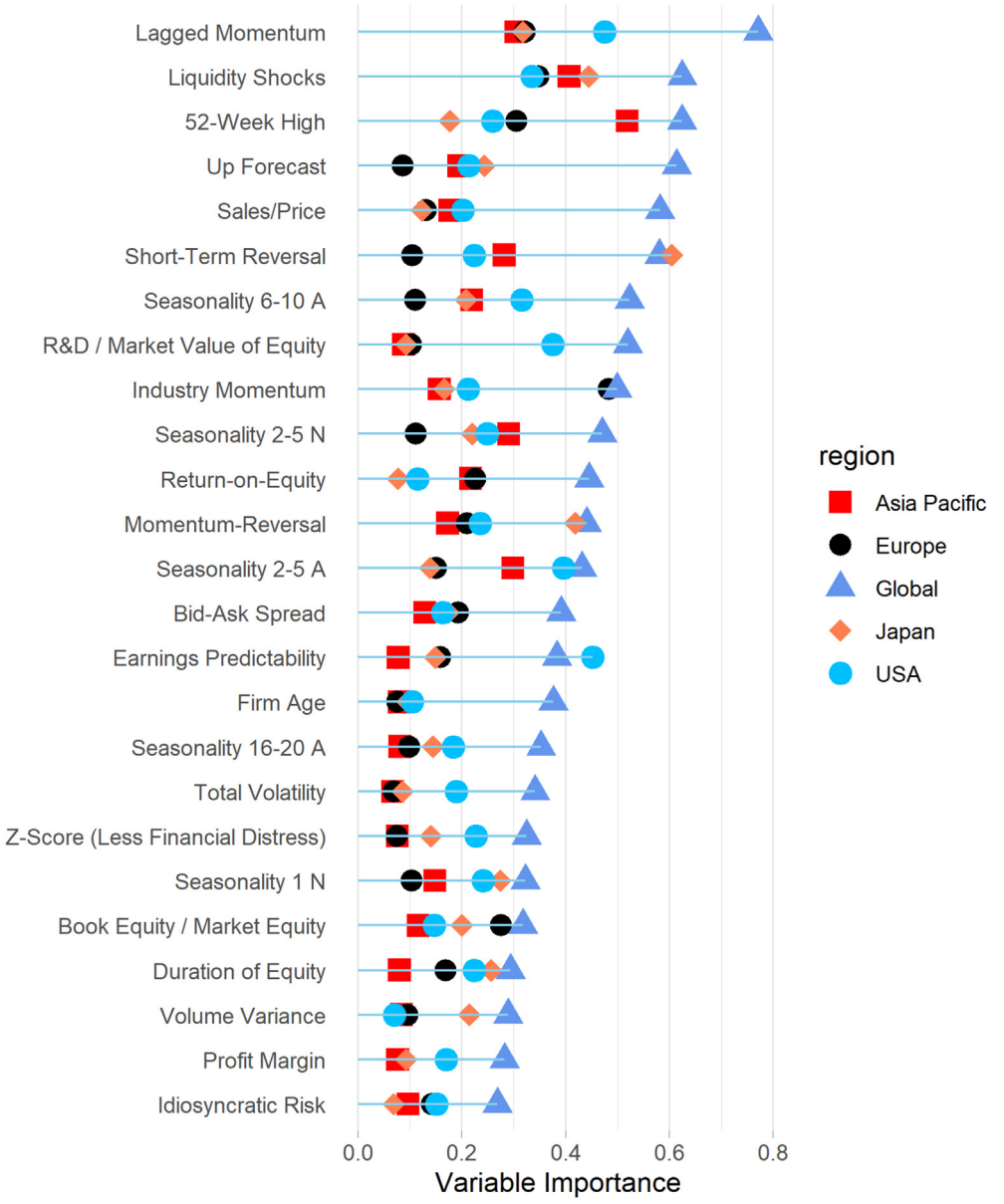

Fig. 5. Variable importance. The figure shows variable importance as in equation (3) for the 25 globally most important variables in the mispricing strategy. The mispricing strategy is described in detail in Subsection 1.3. The strategy combines signals through predictive regressions of individual stock returns on transformed characteristics. The historical predictive regressions are estimated on data from the individual regions using neural networks. Estimation sample spans 1963 to 2018 in the U.S. and 1990 to 2018 in the other regions.

trees is only 0.498. The dispersion of variable importance across the estimation methods does not translate to out-of-sample profitability. This suggests that there is a lot of noise behind the estimation.

3.4. Marginal predictive contributions of signals

Another option when interpreting results from the machine learning models is to look at the marginal relationships. The marginal predictive relationship *MRj* between individual signal *j* and (region-wise demeaned) expected returns is determined as follows:

$$MR_j(x_j) = \frac{1}{T}\sum_{t\in T}\frac{\sum_{i\in N_t}\widehat{f}_{i,t}(x_{i,t-1,1},\dots,x_{i,t-1,j},\dots,x_{i,t-1,M})}{N_t} \quad \text{for } x_j \in (0,1), \tag{4}$$

where $\widehat{f}_{i,t}(x_{i,t-1,1}, \dots, x_{i,t-1,j}, \dots, x_{i,t-1,M})$ is a model prediction. The value of signal x_j is evaluated on a grid ranging from 0 to 1 with a step of 0.01. The values for x_j range from 0 to 1 since variables are region-month cross-sectional quantiles of an underlying characteristic. Values of all signals other than x_j are taken at their true historical realizations. Models are trained on the same regions where forecasts are generated. A set of 100 forecasts corresponding to different values of x_j is obtained by averaging over the cross-section and over time.

Marginal predictive relationships for *size*, *momentum*, *book-to-market*, *asset growth*, *operating profitability*, and *short-term reversal* characteristics can be seen in Figs. 6 and 7.

They show that machine learning models capture the well known empirical asset pricing relationships. Our results are similar to the variable importance results of Gu et al. (2020). We find that the monotonicity of the relationship varies from model to model but the marginal relationships are similar to the relationships documented in the literature for all six reported characteristics but *asset growth* (*AGr*). In case of asset growth, as shown in Fig. 7, we find its association with expected returns is missing for more complex models as can be seen by comparing the slopes for weighted least squares and penalized least squares. The zero slope for penalized least squares is due to the penalization, which is one of the two core features responsible for the superiority of machine learning methods. The second feature is non-linearity. It can be seen by comparing the marginal relationship for penalized weighted least squares model with those for gradient boosted trees, random forests or neural networks. Differences in forecasting performance after introduction of penalization and non-linearity answers how machine learning techniques outperform more standard approaches.

Figs. 6 and 7 also show the effects of non-linearity and penalization internationally as well as in individual regions. These effects are however more diverse across the regions. We also document some of the well known empirical observations like the lack of a momentum anomaly or presence of book-to-market anomaly in Japan (e.g.,Fama and French, 2012). Other relationships are missing at the marginal level, like the momentum anomaly in Europe (e.g., Asness et al., 2013). One of the traditionally strongest signals, short-term momentum, shows a strong relationship in the U.S. as well as in all the other regions except for Asia Pacific. Despite the international diversity of the marginal relationships for anomalies, out-of-sample profitability is not fundamentally impacted. This supports the notion of a high level of noise in estimation as mentioned at the end of Section 2.

4. Transaction costs

In this section, we describe the out-of-sample performance of the strategies after accounting for their transaction costs.

4.1. Transaction costs on the strategy

Panel A in Table 5 presents the average transaction costs on the mispricing strategy introduced in Subsection 1.3 using neural networks. We estimate transaction costs using the three liquidity proxies introduced in Subsection 1.5. All the proxies provide similar estimates of the transaction costs outside the US. Estimates from the Gibbs proxy are significantly higher in the U.S. than for the two other proxies. The Gibbs proxy is, however, also the noisiest proxy since it is constructed at an annual frequency. Thus, it is not very suitable for measuring the transaction costs for the most liquid stocks due to its construction relying on the auto-correlation of daily stock returns.

Panel A in Table 5 also shows the turnover of the mispricing strategy. The turnover is given as:

$$Turnover_t = \sum_i \text{abs}(w_{i,t} - w_{i,t-1} r_{i,t-1})/2, \tag{5}$$

where $w_{i,t}$ is the weight of stock i in the investment portfolio at the start of period $t-1$ and $r_{i,t-1}$ is stock return over period $t-1$ to t. The sum of all the absolute weights $w_{i,t}$ is equal to 2 since the portfolio is long-short. We find that the turnover is close to 100% monthly in all the regions, which means that over 50% of all the held stocks have to be sold and new purchased for both the short and long leg of the strategy. The turnover can be easily reduced by staggered portfolio rebalancing but it is not a source of serious worries here due to the small average transaction costs on the liquid universe of stocks.

We select the sample of stocks to be liquid ex ante. Only about 1000 of the most liquid U.S. stocks fulfil this criterion. These stocks should command virtually no fixed transaction costs after year 2010. The depicted costs therefore correspond to unfavourable trade executions through aggressive marketable orders. Sophisticated trade execution systems using limit orders are able to execute the strategies with much smaller transaction costs.

In Fig. 8, we map transaction costs on the mispricing strategy we estimate using neural networks in the U.S. The transaction costs are measured by using VoV(% Spread) proxy introduced in Fong et al. (2017). The figure shows that the trading costs are similar across the regions. The highest transaction costs tend to be in the Asia Pacific region and the lowest tend to be in the U.S. The transaction costs decrease significantly over time due to the advent of an electronic trading in 2000s. There are several historical episodes where the costs are heavily elevated. Two such major episodes are the Global Financial Crisis of 2007–2009 and dot-com bubble of the early 1995–2001. The costs are smaller on value-weighted portfolios relative to equal-weighted portfolios. The difference is expected because value-weighting puts a larger emphasis on more liquid stocks.

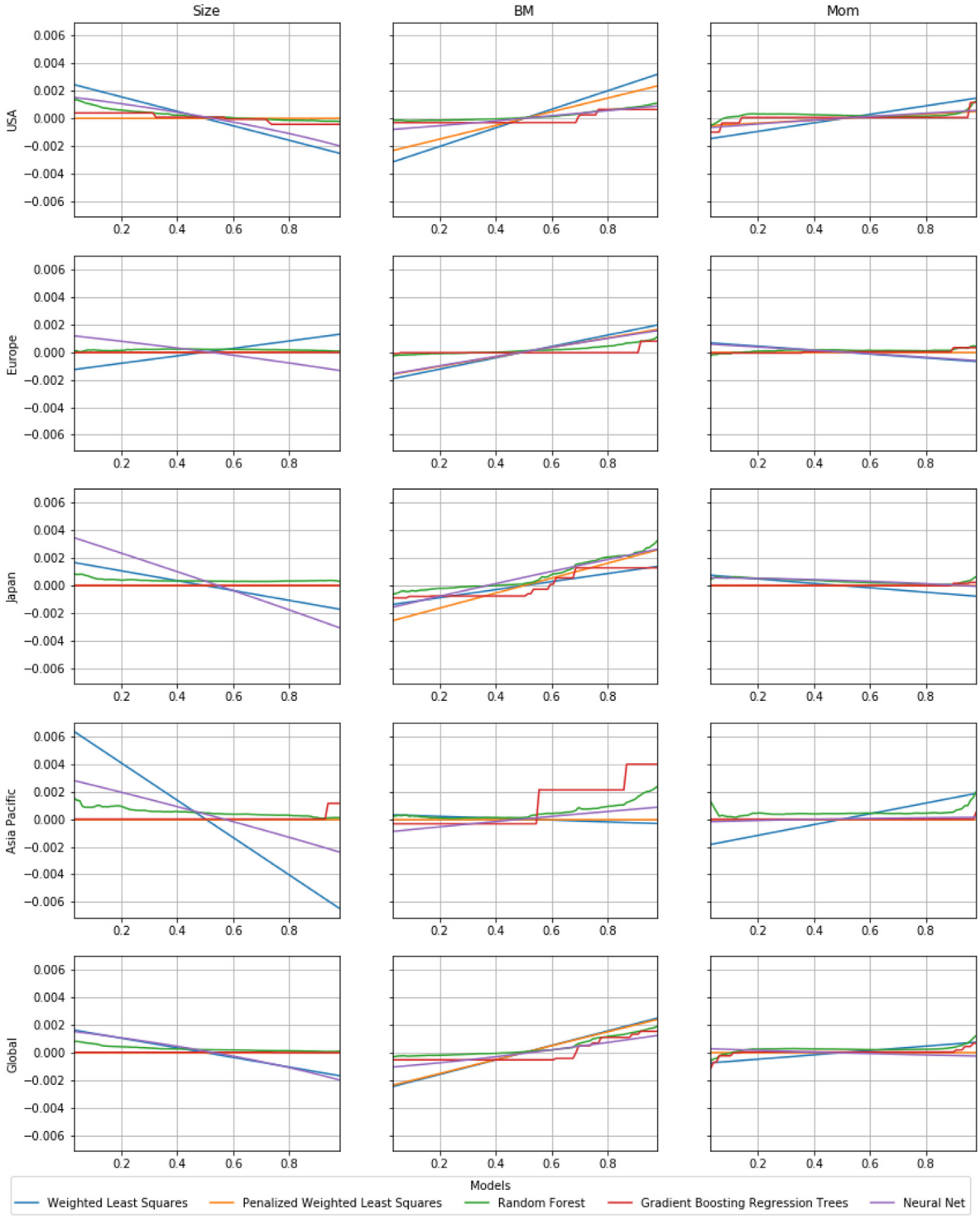

Fig. 6. Marginal predictive contributions of Size, BM, and Mom. The figure shows the marginal predictive relationship for the mispricing strategy between size, book-to-market (BM), momentum (Mom) and (region-wise demeaned) expected returns as defined in Equation (4). The mispricing strategy is described in detail in Subsection 1.3. The methods used in the mispricing strategy are weighted least squares, penalized weighted least squares, gradient boosting regression trees, random forests, and neural networks. Predictions for individual regions are obtained using models estimated on the respective regions only. Estimation sample spans January 1963 to December 2018 in the U.S. and January 1990 to December 2018 in the other regions.

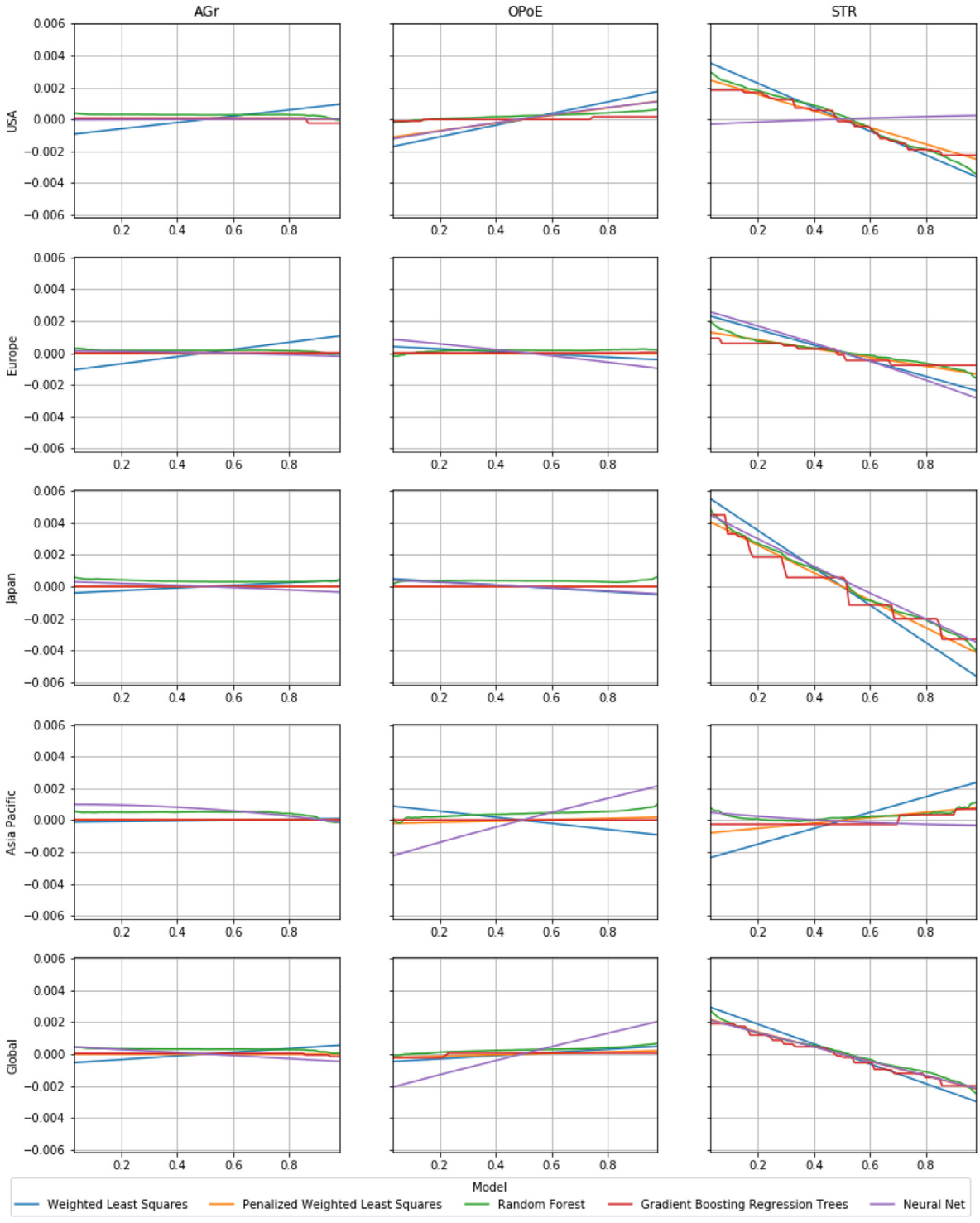

Fig. 7. Marginal predictive contributions of AGr, OPoE, and STR. The figure shows the marginal predictive relationship for the mispricing strategy between asset growth (AGr), operating profitability (OPoE), short-term reversal (STR) and (region-wise demeaned) expected returns as defined in Equation (4). The mispricing strategy is described in detail in Subsection 1.3. The methods used in the mispricing strategy are weighted least squares, penalized weighted least squares, gradient boosting regression trees, random forests, and neural networks. Predictions for individual regions are obtained using models estimated on the respective regions only. Estimation sample spans January 1963 to December 2018 in the U.S. and January 1990 to December 2018 in the other regions.

Table 5
Performance of the strategies after transaction costs.

	Equal-weighted					Value-weighted				
	U.S.	Europe	Japan	AP	Global	U.S.	Europe	Japan	AP	Global
Panel A: Transaction Costs on the Portfolios Using Neural Networks										
VoV	0.248	0.346	0.468	0.590	0.413	0.163	0.259	0.314	0.365	0.275
Gibbs	0.663	0.529	0.694	0.725	0.653	0.595	0.481	0.628	0.553	0.564
Quoted Spread	0.495	0.548	0.498	0.859	0.600	0.424	0.390	0.405	0.578	0.449
Turnover	103.90	104.20	102.70	104.10	103.70	118.00	110.00	111.10	112.90	113.00
Panel B: Net Returns on the Portfolios										
Weighted Least Squares										
Mean	0.339	0.471	0.249	0.048	0.286	0.226	0.564	0.332	0.302	0.311
t-stat	1.296	1.905	1.508	0.260	1.407	1.077	2.741	1.988	1.124	1.680
Sharpe Ratio	0.284	0.556	0.326	0.049	0.376	0.216	0.637	0.339	0.260	0.375
Max Drawdown	−39.42	−31.62	−26.14	−44.42	−29.86	−40.58	−25.45	−34.09	−59.97	−30.95
Penalized Weighted Least Squares										
Mean	0.392	0.447	0.326	0.450	0.385	0.338	0.382	0.321	0.417	0.362
t-stat	1.549	1.546	1.535	1.856	1.767	1.541	1.637	1.329	1.337	1.749
Sharpe Ratio	0.298	0.440	0.354	0.399	0.442	0.271	0.382	0.290	0.317	0.394
Max Drawdown	−41.57	−34.63	−31.94	−49.51	−34.01	−35.58	−31.52	−40.08	−64.30	−34.50
Gradient Boosting Regression Trees										
Mean	0.809	0.636	0.426	0.475	0.618	0.751	0.473	0.543	−0.050	0.484
t-stat	2.890	2.249	2.368	2.291	2.999	2.646	1.939	2.174	−0.252	2.533
Sharpe Ratio	0.657	0.682	0.502	0.514	0.864	0.594	0.462	0.466	−0.052	0.626
Max Drawdown	−45.11	−38.11	−29.50	−23.96	−30.16	−47.00	−37.68	−29.94	−50.99	−28.08
Random Forest										
Mean	0.763	0.710	0.475	0.587	0.641	0.822	0.328	0.562	0.036	0.533
t-stat	2.558	2.592	2.239	2.684	2.972	2.809	1.245	2.294	0.157	2.576
Sharpe Ratio	0.647	0.800	0.574	0.631	0.887	0.740	0.330	0.490	0.033	0.697
Max Drawdown	−39.67	−30.90	−39.47	−23.73	−25.08	−41.70	−40.69	−23.01	−56.66	−22.41
Neural Networks										
Mean	0.880	0.648	0.639	0.776	0.736	1.022	0.475	0.388	0.463	0.736
t-stat	3.473	2.330	3.495	3.732	3.597	4.550	2.234	1.946	1.868	4.466
Sharpe Ratio	0.744	0.706	0.796	0.774	0.952	0.944	0.479	0.359	0.381	0.935
Max Drawdown	−26.97	−28.83	−18.15	−29.90	−20.34	−17.17	−32.92	−35.65	−49.58	−18.93

The table shows returns after transaction costs, along with transaction costs on quintile long-short portfolios from the mispricing strategy as described in Subsection 1.3. The estimation methods are least squares, penalized least squares, random forests, gradient boosting regression trees, or neural networks. The regressions are rerun at the end of each January using only observations from the past and only those anomalies that have been published by that time. Hyperparameters in the estimation are also selected yearly and always based on the most recent 11-year validation sample available. The historical predictive regressions are estimated on individual stocks from the U.S. The long-short portfolios are constructed by buying stocks in the top quintile of the predicted next month returns and short-selling stocks in the bottom quintile of the predicted next month returns. Estimation sample is expanding annually and spans January 1963 to December 1994 up to January 1963 to December 2017 in the U.S. The reported returns are for January 1995 to December 2018 period and are in percentage points. The standard errors in t-statistics are adjusted for heteroskedasticity and autocorrelation with Newey-West adjustment for up to 12 lags. Panel A describes transaction costs and turnover on the mispricing strategy estimated using neural networks. The transaction costs are measured either with VoV(% Spread) proxy of Fong et al. (2017), average daily closing quoted spread, or Gibbs proxy of Hasbrouck (2009). The proxies are further described in Subsection 1.5 The transaction costs and turnover are in percentage points per month. Panel B shows portfolio returns after transaction costs on the mispricing strategy. The transaction costs in Panel B are estimated with VoV(% Spread) proxy of Fong et al. (2017). The returns are reported in percentage points per month.

4.2. Performance of the strategy after transaction costs

Panel B in Table 5 presents the transaction costs adjusted performance of the mispricing strategy. We find that the mean returns on the strategy remain significantly positive at the 5% level. The net mean annualized returns in the U.S. are around 10% for the machine learning strategies. Sharpe ratios remain high, especially for the global strategy using neural networks, where they are close to one.

The mean returns after transaction costs for the weighted least squares method are again smaller than for the more advanced machine learning methods. The difference is even larger on a risk-adjusted basis. This difference in performance documents that the choice of appropriate forecasting method is very important for the success of investing based on the anomalies.

To conclude, the mispricing strategy remains profitable even after accounting for the transaction costs. The profitability of the strategy can therefore be capitalized by investors.

5. Conclusion

We study the profitability of quantitative strategies based on previously documented anomalies around the globe. We show that synthesizing anomalies into one mispricing signal using machine learning leads to profitable investment strategy which survives on the liquid universe of stocks and after accounting for the transaction costs. The machine learning methods lead to

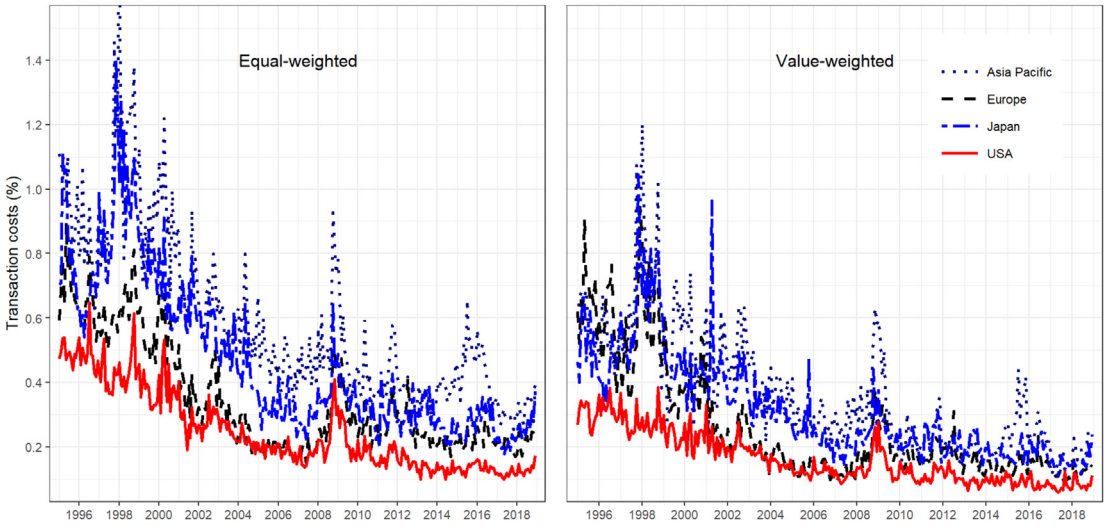

Fig. 8. Monthly transaction costs. The figure visualizes transaction costs for the mispricing strategy using neural networks from the Panel A of Table 5. The transaction costs are estimated with VoV(% Spread) proxy of Fong et al. (2017).

higher (risk-adjusted) returns relative to standard methods applied in the finance literature.

We examine the role of international evidence on the precision of predictions of future stock returns. We find that out-of-sample performance in the U.S. is not improved by the inclusion of international evidence in the training sample for the mispricing strategy. Most of the predictability of the expected stock returns in all the regions can be captured solely with the U.S. training sample as the benefits of a larger estimation sample is offset by the region-specific differences.

Appendix A. Adjustments of returns in datastream

We apply a series of adjustments is applied on the raw returns to improve their quality. The return index (RI) is required to be larger than 0.001 on the first day of the month for precision reasons. RI is set to missing if the daily return is larger than 500% or if the price on the first day of the month is larger than $1 million. Any monthly return larger than 2000% is also set to missing. Datastream provides stale prices when there is no trade during the day or when the stock is no longer traded so that the price of the last trade is repeated until new information arrives. We therefore delete the latest observations of price with no trading. Daily returns are fixed following Tobek and Hronec (2018) when there are stale price quotes around corporate events. Monthly returns larger than 300% that revert back over the next month are set to missing following Ince and Porter (2006).[21] We winsorize 0.01% of returns in each region and years before 2000 to limit the role of outliers in returns.

[21] Specifically, returns in two consecutive months are set as missing if the return in the first month is larger than 300% and the overall return over the two months is lower than 50%.

Appendix B. List of the anomalies

Table 6
List of anomalies

Fundamental	
Accruals	
Accruals	Sloan (1996)
Change in Common Equity	Richardson et al. (2006)
Change in Current Operating Assets	Richardson et al. (2006)
Change in Current Operating Liabilities	Richardson et al. (2006)
Change in Financial Liabilities	Richardson et al. (2006)
Change in Long-Term Investments	Richardson et al. (2006)
Change in Net Financial Assets	Richardson et al. (2006)
Change in Net Non-Cash Working Capital	Richardson et al. (2006)
Change in Net Non-Current Operating Assets	Richardson et al. (2006)
Change in Non-Current Operating Assets	Richardson et al. (2006)
Change in Non-Current Operating Liabilities	Richardson et al. (2006)
Change in Short-Term Investments	Richardson et al. (2006)
Discretionary Accruals	Dechow et al. (1995)
Growth in Inventory	Thomas and Zhang (2002)
Inventory Change	Thomas and Zhang (2002)
Inventory Growth	Belo and Lin (2011)
M/B and Accruals	Bartov and Kim (2004)
Net Working Capital Changes	Soliman (2008)
Percent Operating Accrual	Hafzalla et al. (2011)
Percent Total Accrual	Hafzalla et al. (2011)
Total Accruals	Richardson et al. (2006)
Intangibles	
△ Gross Margin - △ Sales	Abarbanell and Bushee (1998)
△ Sales - △ Accounts Receivable	Abarbanell and Bushee (1998)
△ Sales - △ Inventory	Abarbanell and Bushee (1998)
△ Sales - △ SG and A	Abarbanell and Bushee (1998)
Asset Liquidity	Ortiz-Molina and Phillips (2014)
Asset Liquidity II	Ortiz-Molina and Phillips (2014)
Cash-to-assets	Palazzo (2012)
Earnings Conservatism	Francis et al. (2004)
Earnings Persistence	Francis et al. (2004)
Earnings Predictability	Francis et al. (2004)
Earnings Smoothness	Francis et al. (2004)
Earnings Timeliness	Francis et al. (2004)
Herfindahl Index	Hou and Robinson (2006)
Hiring rate	Belo et al. (2014)
Industry Concentration Assets	Hou and Robinson (2006)
Industry Concentration Book Equity	Hou and Robinson (2006)
Industry-adjusted Organizational Capital-to-Assets	Eisfeldt and Papanikolaou (2013)
Industry-adjusted Real Estate Ratio	Tuzel (2010)
Org. Capital	Eisfeldt and Papanikolaou (2013)
RD/Market Equity	Chan et al. (2001)
RD Capital-to-assets	Li (2011)
RD Expenses-to-sales	Chan et al. (2001)
Tangibility	Hahn and Lee (2009)
Unexpected RD Increases	Eberhart et al. (2004)
Whited-Wu Index	Whited and Wu (2006)
Investment	
△ CAPEX - △ Industry CAPEX	Abarbanell and Bushee (1998)
Asset Growth	Cooper et al. (2008)
Change Net Operating Assets	Hirshleifer et al. (2004)
Changes in PPE and Inventory-to-Assets	Lyandres et al. (2007)
Composite Debt Issuance	Lyandres et al. (2007)
Composite Equity Issuance (5-Year)	Daniel and Titman (2006)
Debt Issuance	Spiess and Affleck-Graves (1995)
Growth in LTNOA	Fairfield et al. (2003)
Investment	Titman et al. (2004)
Net Debt Finance	Bradshaw et al. (2006)
Net Equity Finance	Bradshaw et al. (2006)
Net Operating Assets	Hirshleifer et al. (2004)
Noncurrent Operating Assets Changes	Soliman (2008)

(*continued on next page*)

Table 6 *(continued)*

Fundamental	
Share Repurchases	Ikenberry et al. (1995)
Total XFIN	Bradshaw et al. (2006)
Profitability	
Asset Turnover	Soliman (2008)
Capital Turnover	Haugen and Baker (1996)
Cash-based Operating Profitability	Ball et al. (2016)
Change in Asset Turnover	Soliman (2008)
Change in Profit Margin	Soliman (2008)
Earnings/Price	Basu (1977)
Earnings Consistency	Alwathainani (2009)
F-Score	Piotroski (2000)
Gross Profitability	Novy-Marx (2013)
Labor Force Efficiency	Abarbanell and Bushee (1998)
Leverage	Bhandari (1988)
O-Score (More Financial Distress)	Dichev (1998)
Operating Profits to Assets	Ball et al. (2016)
Operating Profits to Equity	Fama and French (2015)
Profit Margin	Soliman (2008)
Return on Net Operating Assets	Soliman (2008)
Return-on-Equity	Haugen and Baker (1996)
Z-Score (Less Financial Distress)	Dichev (1998)
Value	
Assets-to-Market	Fama and French (1992)
Book Equity/Market Equity	Fama and French (1992)
Cash Flow/Market Equity	Lakonishok et al. (1994)
Duration of Equity	Dechow et al. (2004)
Enterprise Component of Book/Price	Penman et al. (2007)
Enterprise Multiple	Loughran and Wellman (2011)
Intangible Return	Daniel and Titman (2006)
Leverage Component of Book/Price	Penman et al. (2007)
Net Payout Yield	Boudoukh et al. (2007)
Operating Leverage	Novy-Marx (2010)
Payout Yield	Boudoukh et al. (2007)
Sales Growth	Lakonishok et al. (1994)
Sales/Price	Barbee Jr et al. (1996)
Sustainable Growth	Lockwood and Prombutr (2010)
Market Friction	
11-Month Residual Momentum	Blitz et al. (2011)
52-Week High	George and Hwang (2004)
Amihud's Measure (Illiquidity)	Amihud (2002)
Beta	Fama and MacBeth (1973)
Betting against Beta	Frazzini and Pedersen (2014)
Bid-Ask Spread	Amihud and Mendelson (1986)
Cash Flow Variance	Haugen and Baker (1996)
Coefficient of Variation of Share Turnover	Chordia et al. (2001)
Coskewness	Harvey and Siddique (2000)
Downside Beta	Ang et al. (2006)
Earnings Forecast-to-Price	Elgers et al. (2001)
Firm Age	Barry and Brown (1984)
Firm Age-Momentum	Zhang (2006)
Idiosyncratic Risk	Ang et al. (2006)
Industry Momentum	Moskowitz and Grinblatt (1999)
Lagged Momentum	Novy-Marx (2012)
Liquidity Beta 1	Acharya and Pedersen (2005)
Liquidity Beta 2	Acharya and Pedersen (2005)
Liquidity Beta 3	Acharya and Pedersen (2005)
Liquidity Beta 4	Acharya and Pedersen (2005)
Liquidity Beta 5	Acharya and Pedersen (2005)
Liquidity Shocks	Bali et al. (2013)
Long-Term Reversal	Bondt and Thaler (1985)
Max	Bali et al. (2011)
Momentum	Jegadeesh and Titman (1993)
Momentum and LT Reversal	Kot and Chan (2006)

(*continued on next page*)

Table 6 *(continued)*

Fundamental	
Momentum-Reversal	Jegadeesh and Titman (1993)
Momentum-Volume	Lee and Swaminathan (2000)
Price	Blume and Husic (1973)
Seasonality	Heston and Sadka (2008)
Seasonality 1 A	Heston and Sadka (2008)
Seasonality 1 N	Heston and Sadka (2008)
Seasonality 11–15 A	Heston and Sadka (2008)
Seasonality 11–15 N	Heston and Sadka (2008)
Seasonality 16–20 A	Heston and Sadka (2008)
Seasonality 16–20 N	Heston and Sadka (2008)
Seasonality 2–5 A	Heston and Sadka (2008)
Seasonality 2–5 N	Heston and Sadka (2008)
Seasonality 6–10 A	Heston and Sadka (2008)
Seasonality 6–10 N	Heston and Sadka (2008)
Share Issuance (1-Year)	Pontiff and Woodgate (2008)
Share Turnover	Datar et al. (1998)
Short-Term Reversal	Jegadeesh (1990)
Size	Banz (1981)
Tail Risk	Kelly and Jiang (2014)
Total Volatility	Ang et al. (2006)
Volume/Market Value of Equity	Haugen and Baker (1996)
Volume Trend	Haugen and Baker (1996)
Volume Variance	Chordia et al. (2001)
I/B/E/S	
Analyst Value	Frankel and Lee (1998)
Analysts Coverage	Elgers et al. (2001)
Change in Forecast + Accrual	Barth and Hutton (2004)
Change in Recommendation	Jegadeesh et al. (2004)
Changes in Analyst Earnings Forecasts	Hawkins et al. (1984)
Disparity between LT and ST Earnings Growth Forecasts	Da and Warachka (2011)
Dispersion in Analyst LT Growth Forecasts	Anderson et al. (2005)
Down Forecast	Barber et al. (2001)
Forecast Dispersion	Diether et al. (2002)
Long-Term Growth Forecasts	La Porta (1996)
Up Forecast	Barber et al. (2001)

Appendix C. Optimal hyperparameters

Every time a model is trained, we perform hyperparameter optimization as described in Subsection 1.3. In the models we employ for the mispricing strategy, we combine signals through predictive regressions of individual stock returns on transformed characteristics. The mispricing strategy is described in detail in Subsection 1.3. Underlying predictive regressions are estimated using penalized weighted least squares, gradient boosted trees, random forests and neural networks. Optimal hyperparameters vary over time and their comparison for models trained and validated on the U.S. data only versus globally,[22] can be seen in Figures C.1, C.2, C.3, and C.4.

Figure C.2 shows the evolution of selected number of trees, as one of the hyperparameters for gradient boosted trees. The selected number of trees is very similar for the U.S. only sample and the global sample.

Optimal hyperparameters in case of penalized weighted least squares and random forests, i.e. l1 and l2 mixing parameter in Figure C.1 and the number of tree in Figure C.3, are also similar between the U.S. only sample and the global sample. This holds strongly especially before 2010. Looking at the architecture of the neural networks and dropout shown in Figure C.4, selected hyperparameters differ substantially across the regions and over the time.

[22] Includes the U.S. data as well.

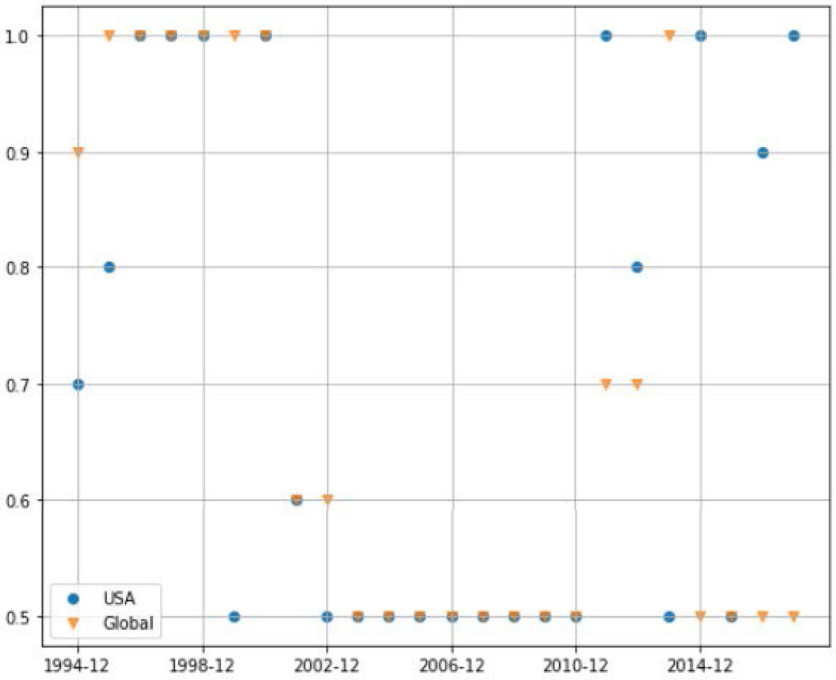

Fig. C.1 L1 and L2 mixing parameter for penalized weighted least square. The figure shows the evolution of optimal selected L1 and L2 mixing parameter as a hyperparameter for penalized weighted least squares during each tuning phase as described in Subsection 1.3. Penalized weighted least square are trained and cross-validated either only in the U.S. or in all the regions, i.e. globally. Training-validation-test splits are described in Subsection 1.3. The estimation sample is from January 1963 to December 2018 in the U.S. and from January 1990 to December 2018 elsewhere.

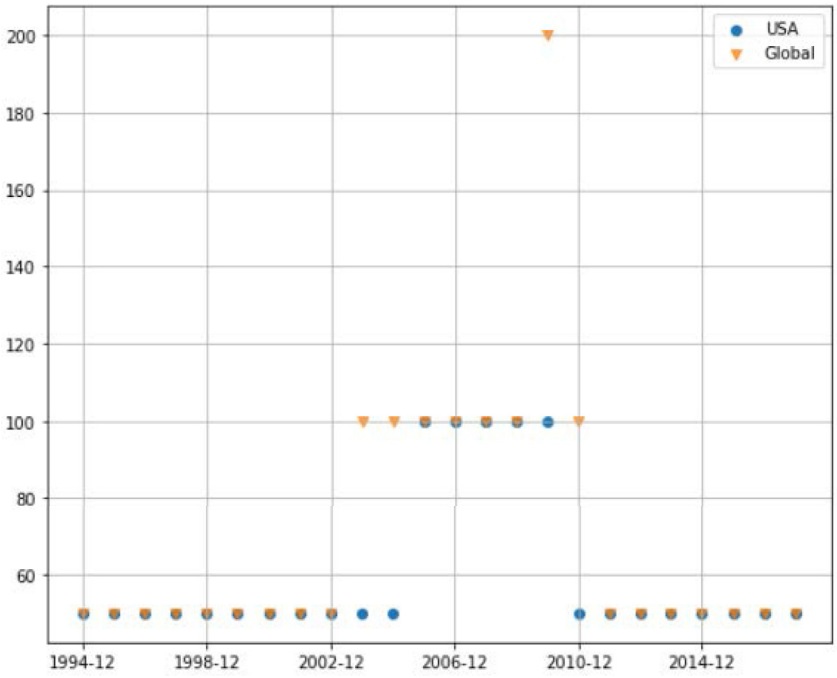

Fig. C.2 Number of trees for gradient boosting trees. The figure shows the evolution of selected optimal number of trees as a hyperparameter for gradient boosting trees during each tuning phase as described in Subsection 1.3. Gradient boosting trees are trained and cross-validated either only in the U.S. or in all the regions, i.e. globally. Training-validation-test splits are described in Subsection 1.3. The estimation sample is from January 1963 to December 2018 in the U.S. and from January 1990 to December 2018 elsewhere.

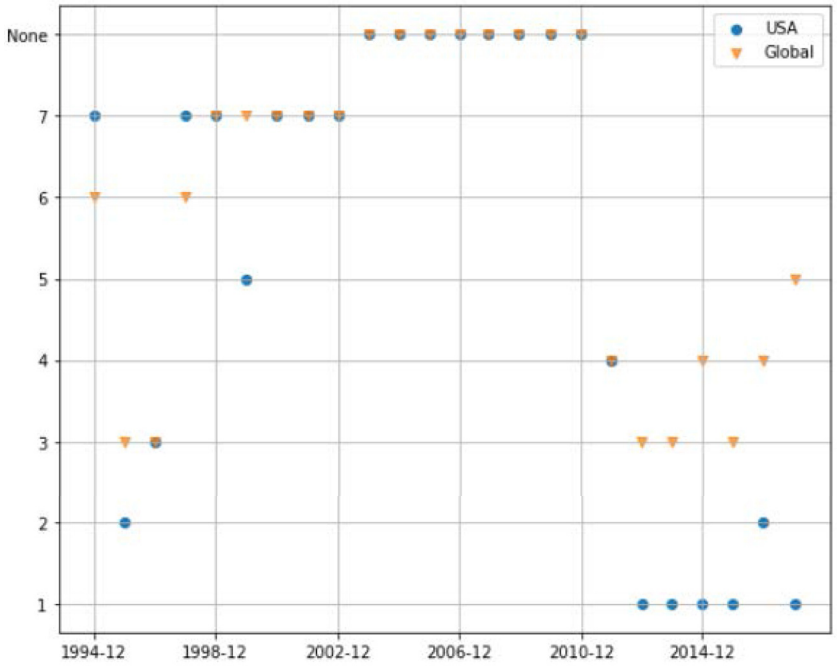

Fig. C.3 Maximum depth of the tree for random forests. The figure shows the evolution of selected optimal maximum depth of the tree as a hyperparameter for random forests during each tuning phase as described in Subsection 1.3. Random forests are trained and cross-validated either only in the U.S. or in all the regions, i.e. globally. Training-validation-test splits are described in Subsection 1.3. The estimation sample is from January 1963 to December 2018 in the U.S. and from January 1990 to December 2018 elsewhere.

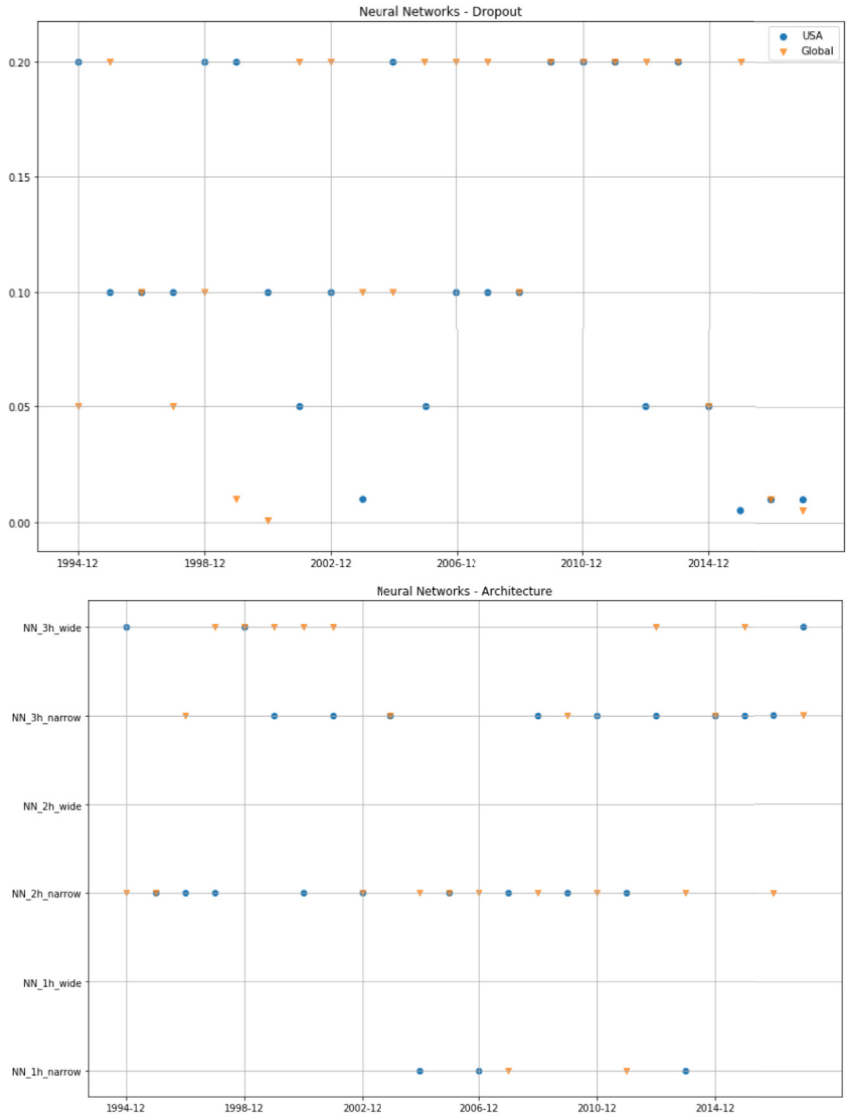

Fig. C.4 Dropout and the architecture for neural networkds. The figure shows the evolution of selected optimal dropout and the architecture as hyperparameters for neural networks during each tuning phase as described in Subsection 1.3. Neural networks are trained and cross-validated either only in the U.S. or in all the regions, i.e. globally. Training-validation-test splits are described in Subsection 1.3. The estimation sample is from January 1963 to December 2018 in the U.S. and from January 1990 to December 2018 elsewhere.

Appendix D. Definition of Liquidity Proxies

D.1. VoV(% Spread) Proxy

The fixed transaction costs are approximated with the VoV(% Spread) proxy introduced in Fong et al. (2017). It is defined as:

$$8\,\frac{\sigma^{2/3}}{avg\ vol^{1/3}}, \tag{6}$$

where σ is the standard deviation of daily returns and *avg vol* is the average daily trading volume in USD within a given month. The trading volume is in USD and deflated to 2000 prices. The proxy roughly measures the fixed component of the trading costs and excludes the price impact. Including the price impact would further increase the transaction costs. Fong et al. (2017) show that the price impact component is very hard to measure. It is volatile over regions, and therefore, very dependent on

the execution strategy of individual asset managers. The focus is therefore solely on the fixed component of transaction costs (effective spread).

Kyle and Obizhaeva (2016) estimate a relationship between transaction costs and the size of large institutional portfolio transfers depending on the average daily trading volume and the volatility of the stocks. Their analysis is conducted on a proprietary dataset covering the 2002–2005 period. VoV(% Spread) roughly corresponds to the fixed component of their estimated transaction cost function.

Fong et al. (2017) benchmark the proxy to other proxies and find that it can be outperformed only by the closing quoted spread. The quoted spread is, however, not available for all the regions over the whole sample period.

D.2. Closing Quoted Spread

The closing quoted spread for a given month is as follows:

$$QS = \frac{1}{T}\sum_{t=1}^{T}\frac{2(ask - bid)}{ask + bid}, \tag{7}$$

where ask and bid are observed at the end of the trading day on each stock exchange and T is the number of days in the given month. Observations with missing or negative daily values of QS are excluded from the average. CRSP lists the best quote of bid and ask for NASDAQ stocks and the last representative quotes before the market close for NYSE and AMEX stocks. The precise definition of QS can therefore vary over the exchanges.

Chung and Zhang (2014) first benchmark the QS by comparing it to high frequency effective spread estimates from the Trade and Quote (TAQ) database. They show that QS has about a 95% average cross-sectional correlation with the TAQ effective spread over the 1998 to 2009 period. Fong et al. (2017) document that it is also the best spread proxy in an international setting. One problem with QS is that it is often missing in earlier periods and therefore has to be backfilled with other proxies.

D.3. Gibbs Proxy

Roll (1984) introduces one of the first spread proxies in the literature. He assumes that the true price of a stock follows a random walk with bid-ask jumps. That is,

$$P_t^A = P_{t-1}^A + u_t, \quad P_t^O = P_t^A + sq_t \tag{8}$$

$$\triangle P_t^o = s \triangle q_t + u_t,\ u_t \% N(0, \sigma_u^2), \tag{9}$$

where P_t^o is the observed log price, P_t^A is the price of the underlying Brownian motion, and s is a half spread. Indicator q_t is equal to one if the last trade in the day is a buy, minus one if it is a sell, and zero if no prices are available during the day. Serial correlation of the price changes $\triangle P_t^o$ should be negative and related to the spread through the following relationship:

$$S_{roll} = 2\sqrt{-\text{cov}(\triangle P_t^o, \triangle P_{t+1}^o)}. \tag{10}$$

This can be contributed to the fact that

$$\text{cov}(\triangle P_t^o, \triangle P_{t+1}^o) = \text{cov}(s(q_t - q_{t-1}) + u_t, s(q_{t+1} - q_t) + u_{t+1}) = \mathbb{E}[-s^2 q_t^2] = -s^2. \tag{11}$$

The covariance can be positive in practice. In which case the estimate of spread is set equal to zero.

Hasbrouck (2009) proposes to extend the Roll model by estimating it with the Gibbs sampler. The idea is to estimate equation (9) augmented with another dependent variable (market return) via Bayesian regression. The variables q_t are generated from the data using a Gibbs sampler.[23]

We estimate the proxy at annual frequency for each stock and calendar year. Lower frequency than annual leads to severe deterioration of the proxy's performance.

References

Abarbanell, J.S., Bushee, B.J., 1998. Abnormal returns to a fundamental analysis strategy. Account. Rev. 19–45.
Acharya, V.V., Pedersen, L.H., 2005. Asset pricing with liquidity risk. J. Financ. Econ. 77 (2), 375–410.
Alwathainani, A.M., 2009. Consistency of firms' past financial performance measures and future returns. Br. Account. Rev. 41 (3), 184–196.
Amihud, Y., 2002. Illiquidity and stock returns: cross-section and time-series effects. J. Financ. Mark. 5 (1), 31–56.
Amihud, Y., Mendelson, H., 1986. Asset pricing and the bid-ask spread. J. Financ. Econ. 17 (2), 223–249.

[23] Note that there is an error in the original paper in the *Journal of Finance*. The correct posterior distribution for σ_u^2 is $IG(\alpha_{prior} + \frac{n}{2}, \beta_{prior} + \frac{\sum u_t^2}{2})$.

Anderson, E.W., Ghysels, E., Juergens, J.L., 2005. Do heterogeneous beliefs matter for asset pricing? Rev. Financ. Stud. 18 (3), 875–924.
03 Andrew Karolyi, G., 2016. Home bias, an academic puzzle. Rev. Finance 20 (6), 2049–2078.
Andrikopoulos, P., Clunie, J., Siganos, A., 2013. Short-selling constraints and 'quantitative'investment strategies. Eur. J. Finance 19 (1), 19–35.
Ang, A., Chen, J., Xing, Y., 2006a. Downside risk. Rev. Financ. Stud. 19 (4), 1191–1239.
Ang, A., Hodrick, R.J., Xing, Y., Zhang, X., 2006b. The cross-section of volatility and expected returns. J. Finance 61 (1), 259–299.
Asness, C.S., Moskowitz, T.J., Pedersen, L.H., 2013. Value and momentum everywhere. J. Finance 68 (3), 929–985.
Asparouhova, E., Bessembinder, H., Kalcheva, I., 2010. Liquidity biases in asset pricing tests. J. Financ. Econ. 96 (2), 215–237.
Avramov, D., Cheng, S., Metzker, L., 2019. Machine Learning versus Economic Restrictions: Evidence from Stock Return Predictability. Available at: SSRN 3450322.
Avramov, D., Chordia, T., Jostova, G., Philipov, A., 2013. Anomalies and financial distress. J. Financ. Econ. 108 (1), 139–159.
Bali, T.G., Cakici, N., Whitelaw, R.F., 2011. Maxing out: stocks as lotteries and the cross-section of expected returns. J. Financ. Econ. 99 (2), 427–446.
Bali, T.G., Peng, L., Shen, Y., Tang, Y., 2013. Liquidity shocks and stock market reactions. Rev. Financ. Stud. 27 (5), 1434–1485.
Ball, R., Gerakos, J., Linnainmaa, J.T., Nikolaev, V., 2016. Accruals, cash flows, and operating profitability in the cross section of stock returns. J. Financ. Econ. 121 (1), 28–45.
Banz, R.W., 1981. The relationship between return and market value of common stocks. J. Financ. Econ. 9 (1), 3–18.
Barbee Jr., W.C., Mukherji, S., Raines, G.A., 1996. Do sales–price and debt–equity explain stock returns better than book–market and firm size? Financ. Anal. J. 52 (2), 56–60.
Barber, B., Lehavy, R., McNichols, M., Trueman, B., 2001. Can investors profit from the prophets? security analyst recommendations and stock returns. J. Finance 56 (2), 531–563.
Barber, B.M., De George, E.T., Lehavy, R., Trueman, B., 2013. The earnings announcement premium around the globe. J. Financ. Econ. 108 (1), 118–138.
Barry, C.B., Brown, S.J., 1984. Differential information and the small firm effect. J. Financ. Econ. 13 (2), 283–294.
Barth, M.E., Hutton, A.P., 2004. Analyst earnings forecast revisions and the pricing of accruals. Rev. Account. Stud. 9 (1), 59–96.
Bartov, E., Kim, M., 2004. Risk, mispricing, and value investing. Rev. Quant. Finance Account. 23 (4), 353–376.
Bartram, S.M., Grinblatt, M., 2018. Global Market Inefficiencies.
Basu, S., 1977. Investment performance of common stocks in relation to their price-earnings ratios: a test of the efficient market hypothesis. J. Finance 32 (3), 663–682.
Belo, F., Lin, X., 2011. The inventory growth spread. Rev. Financ. Stud. 25 (1), 278–313.
Belo, F., Lin, X., Bazdresch, S., 2014. Labor hiring, investment, and stock return predictability in the cross section. J. Polit. Econ. 122 (1), 129–177.
Bhandari, L.C., 1988. Debt/equity ratio and expected common stock returns: empirical evidence. J. Finance 43 (2), 507–528.
Blitz, D., Huij, J., Martens, M., 2011. Residual momentum. J. Empir. Finance 18 (3), 506–521.
Blume, M.E., Husic, F., 1973. Price, beta, and exchange listing. J. Finance 28 (2), 283–299.
Bondt, W.F., Thaler, R., 1985. Does the stock market overreact? J. Finance 40 (3), 793–805.
Boudoukh, J., Michaely, R., Richardson, M., Roberts, M.R., 2007. On the importance of measuring payout yield: implications for empirical asset pricing. J. Finance 62 (2), 877–915.
Bradshaw, M.T., Richardson, S.A., Sloan, R.G., 2006. The relation between corporate financing activities, analysts' forecasts and stock returns. J. Account. Econ. 42 (1), 53–85.
Chan, L.K., Lakonishok, J., Sougiannis, T., 2001. The stock market valuation of research and development expenditures. J. Finance 56 (6), 2431–2456.
Chen, L., Pelger, M., Zhu, J., 2019. Deep Learning in Asset Pricing. Available at: SSRN 3350138.
Chen, T., He, T., 2017. Xgboost: Extreme Gradient Boosting.
Chordia, T., Subrahmanyam, A., Anshuman, V.R., 2001. Trading activity and expected stock returns. J. Financ. Econ. 59 (1), 3–32.
Chordia, T., Subrahmanyam, A., Tong, Q., 2014. Have capital market anomalies attenuated in the recent era of high liquidity and trading activity? J. Account. Econ. 58 (1), 41–58.
Chui, A.C., Titman, S., Wei, K.J., 2010. Individualism and momentum around the world. J. Finance 65 (1), 361–392.
Chung, K.H., Zhang, H., 2014. A simple approximation of intraday spreads using daily data. J. Financ. Mark. 17, 94–120.
Cooper, M.J., Gulen, H., Schill, M.J., 2008. Asset growth and the cross-section of stock returns. J. Finance 63 (4), 1609–1651.
Da, Z., Warachka, M., 2011. The disparity between long-term and short-term forecasted earnings growth. J. Financ. Econ. 100 (2), 424–442.
Daniel, K., Titman, S., 2006. Market reactions to tangible and intangible information. J. Finance 61 (4), 1605–1643.
Datar, V.T., Naik, N.Y., Radcliffe, R., 1998. Liquidity and stock returns: an alternative test. J. Financ. Mark. 1 (2), 203–219.
Dechow, P.M., Sloan, R.G., Soliman, M.T., 2004. Implied equity duration: a new measure of equity risk. Rev. Account. Stud. 9 (2–3), 197–228.
Dechow, P.M., Sloan, R.G., Sweeney, A.P., 1995. Detecting Earnings Management. *Accounting Review*, pp. 193–225.
Dichev, I.D., 1998. Is the risk of bankruptcy a systematic risk? J. Finance 53 (3), 1131–1147.
Diether, K.B., Malloy, C.J., Scherbina, A., 2002. Differences of opinion and the cross section of stock returns. J. Finance 57 (5), 2113–2141.
Eberhart, A.C., Maxwell, W.F., Siddique, A.R., 2004. An examination of long-term abnormal stock returns and operating performance following r&d increases. J. Finance 59 (2), 623–650.
Eisfeldt, A.L., Papanikolaou, D., 2013. Organization capital and the cross-section of expected returns. J. Finance 68 (4), 1365–1406.
Elgers, P.T., Lo, M.H., Pfeiffer Jr., R.J., 2001. Delayed security price adjustments to financial analysts' forecasts of annual earnings. Account. Rev. 76 (4), 613–632.
Fairfield, P.M., Whisenant, J.S., Yohn, T.L., 2003. Accrued earnings and growth: implications for future profitability and market mispricing. Account. Rev. 78 (1), 353–371.
Fama, E.F., French, K.R., 1992. The cross-section of expected stock returns. J. Finance 47 (2), 427–465.
Fama, E.F., French, K.R., 2012. Size, value, and momentum in international stock returns. J. Financ. Econ. 105 (3), 457–472.
Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. J. Financ. Econ. 116 (1), 1–22.
Fama, E.F., French, K.R., 2017. International tests of a five-factor asset pricing model. J. Financ. Econ. 123 (3), 441–463.
Fama, E.F., MacBeth, J.D., 1973. Risk, return, and equilibrium: empirical tests. J. Polit. Econ. 81 (3), 607–636.
Fong, K., Holden, C., Tobek, O., 2017. Are Volatility over Volume Liquidity Proxies Useful for Global or Us Research? .
Francis, J., LaFond, R., Olsson, P.M., Schipper, K., 2004. Costs of equity and earnings attributes. Account. Rev. 79 (4), 967–1010.
Frankel, R., Lee, C.M., 1998. Accounting valuation, market expectation, and cross-sectional stock returns. J. Account. Econ. 25 (3), 283–319.
Frazzini, A., Israel, R., Moskowitz, T.J., 2012. Trading Costs of Asset Pricing Anomalies. Fama-Miller Working Paper. 14–05.
Frazzini, A., Pedersen, L.H., 2014. Betting against beta. J. Financ. Econ. 111 (1), 1–25.
Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning, ume 1. Springer series in statistics, New York, NY, USA.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.
George, T.J., Hwang, C.-Y., 2004. The 52-week high and momentum investing. J. Finance 59 (5), 2145–2176.
Green, J., Hand, J.R., Zhang, X.F., 2017. The characteristics that provide independent information about average us monthly stock returns. Rev. Financ. Stud..
Griffin, J.M., 2002. Are the fama and French factors global or country specific? Rev. Financ. Stud. 15 (3), 783–803.
Griffin, J.M., Kelly, P.J., Nardari, F., 2010. Do market efficiency measures yield correct inferences? a comparison of developed and emerging markets. Rev. Financ. Stud. 23 (8), 3225–3277.
02 Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. Rev. Financ. Stud. 33 (5), 2223–2273.
Gu, S., Kelly, B.T., Xiu, D., 2019. Autoencoder Asset Pricing Models. (Available at: SSRN).
Hafzalla, N., Lundholm, R., Matthew Van Winkle, E., 2011. Percent accruals. Account. Rev. 86 (1), 209–236.

Hahn, J., Lee, H., 2009. Financial constraints, debt capacity, and the cross-section of stock returns. J. Finance 64 (2), 891–921.
Harvey, C.R., Liu, Y., Zhu, H., 2016. And the cross-section of expected returns. Rev. Financ. Stud. 29 (1), 5–68.
Harvey, C.R., Siddique, A., 2000. Conditional skewness in asset pricing tests. J. Finance 55 (3), 1263–1295.
Hasbrouck, J., 2009. Trading costs and returns for us equities: estimating effective costs from daily data. J. Finance 64 (3), 1445–1477.
Haugen, R.A., Baker, N.L., 1996. Commonality in the determinants of expected stock returns. J. Financ. Econ. 41 (3), 401–439.
Hawkins, E.H., Chamberlin, S.C., Daniel, W.E., 1984. Earnings expectations and security prices. Financ. Anal. J. 40 (5), 24–38.
Heston, S.L., Sadka, R., 2008. Seasonality in the cross-section of stock returns. J. Financ. Econ. 87 (2), 418–445.
Hirshleifer, D., Hou, K., Teoh, S.H., Zhang, Y., 2004. Do investors overvalue firms with bloated balance sheets? J. Account. Econ. 38, 297–331.
Horel, E., Giesecke, K., 2019. Towards Explainable Ai: Significance Tests for Neural Networks. arXiv preprint arXiv:1902.06021.
Hou, K., Karolyi, G.A., Kho, B.-C., 2011. What factors drive global stock returns? Rev. Financ. Stud. 24 (8), 2527–2574.
Hou, K., Robinson, D.T., 2006. Industry concentration and average stock returns. J. Finance 61 (4), 1927–1956.
Hou, K., Xue, C., Zhang, L., 2018. Replicating anomalies. Rev. Financ. Stud. 12.
Ikenberry, D., Lakonishok, J., Vermaelen, T., 1995. Market underreaction to open market share repurchases. J. Financ. Econ. 39 (2), 181–208.
Ilmanen, A., Israel, R., Moskowitz, T., Thapar, A., Wang, F., 2019. Do Factor Premia Vary over Time? a Century of Evidence. (Technical report, Working Paper, AQR Capital Management).
Ince, O.S., Porter, R.B., 2006. Individual equity return data from thomson datastream: handle with care!. J. Financ. Res. 29 (4), 463–479.
Jacobs, H., Müller, S., 2018. And Nothing Else Matters? on the Dimensionality and Predictability of International Stock Returns.
Jacobs, H., Müller, S., 2020. Anomalies across the globe: once public, no longer existent? J. Financ. Econ. 135 (1), 213–230.
Jegadeesh, N., 1990. Evidence of predictable behavior of security returns. J. Finance 45 (3), 881–898.
Jegadeesh, N., Kim, J., Krische, S.D., Lee, C., 2004. Analyzing the analysts: when do recommendations add value? J. Finance 59 (3), 1083–1124.
Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: implications for stock market efficiency. J. Finance 48 (1), 65–91.
Kelly, B., Jiang, H., 2014. Tail risk and asset prices. Rev. Financ. Stud. 27 (10), 2841–2871.
Kelly, B.T., Pruitt, S., Su, Y., 2019. Characteristics are covariances: a unified model of risk and return. J. Financ. Econ. 134 (3), 501–524.
Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980.
Kot, H.W., Chan, K., 2006. Can contrarian strategies improve momentum profits. J. Invest. Manag. 4 (1).
Kyle, A.S., Obizhaeva, A.A., 2016. Market microstructure invariance: empirical hypotheses. Econometrica 84 (4), 1345–1404.
La Porta, R., 1996. Expectations and the cross-section of stock returns. J. Finance 51 (5), 1715–1742.
Lakonishok, J., Shleifer, A., Vishny, R.W., 1994. Contrarian investment, extrapolation, and risk. J. Finance 49 (5), 1541–1578.
Lam, F.E.C., Wei, K.J., 2011. Limits-to-arbitrage, investment frictions, and the asset growth anomaly. J. Financ. Econ. 102 (1), 127–149.
Lee, C., Swaminathan, B., 2000. Price momentum and trading volume. J. Finance 55 (5), 2017–2069.
Lee, K.-H., 2011. The world price of liquidity risk. J. Financ. Econ. 99 (1), 136–161.
Lewellen, J., et al., 2015. The cross-section of expected stock returns. Critic. Finan. Rev. 4 (1), 1–44.
Li, D., 2011. Financial constraints, r&d investment, and stock returns. Rev. Financ. Stud. 24 (9), 2974–3007.
Lockwood, L., Prombutr, W., 2010. Sustainable growth and stock returns. J. Financ. Res. 33 (4), 519–538.
Loughran, T., Wellman, J.W., 2011. New evidence on the relation between the enterprise multiple and average stock returns. J. Financ. Quant. Anal. 46 (6), 1629–1650.
Lu, X., Stambaugh, R.F., Yuan, Y., 2017. Anomalies Abroad: beyond Data Mining. Technical report. National Bureau of Economic Research.
Lyandres, E., Sun, L., Zhang, L., 2007. The new issues puzzle: testing the investment-based explanation. Rev. Financ. Stud. 21 (6), 2825–2855.
McLean, R.D., Pontiff, J., 2016. Does academic research destroy stock return predictability? J. Finance 71 (1), 5–32.
McLean, R.D., Pontiff, J., Watanabe, A., 2009. Share issuance and cross-sectional returns: international evidence. J. Financ. Econ. 94 (1), 1–17.
Moskowitz, T.J., Grinblatt, M., 1999. Do industries explain momentum? J. Finance 54 (4), 1249–1290.
Novy-Marx, R., 2010. Operating leverage. Rev. Finance 15 (1), 103–134.
Novy-Marx, R., 2012. Is momentum really momentum? J. Financ. Econ. 103 (3), 429–453.
Novy-Marx, R., 2013. The other side of value: the gross profitability premium. J. Financ. Econ. 108 (1), 1–28.
Novy-Marx, R., Velikov, M., 2015. A taxonomy of anomalies and their trading costs. Rev. Financ. Stud. 29 (1), 104–147.
Ortiz-Molina, H., Phillips, G.M., 2014. Real asset illiquidity and the cost of capital. J. Financ. Quant. Anal. 49 (1), 1–32.
Palazzo, B., 2012. Cash holdings, risk, and expected returns. J. Financ. Econ. 104 (1), 162–185.
Penman, S.H., Richardson, S.A., Tuna, I., 2007. The book-to-price effect in stock returns: accounting for leverage. J. Account. Res. 45 (2), 427–467.
Piotroski, J.D., 2000. Value investing: the use of historical financial statement information to separate winners from losers. J. Account. Res. 1–41.
Pontiff, J., Woodgate, A., 2008. Share issuance and cross-sectional returns. J. Finance 63 (2), 921–945.
Richardson, S.A., Sloan, R.G., Soliman, M.T., Tuna, I., 2006. The implications of accounting distortions and growth for accruals and profitability. Account. Rev. 81 (3), 713–743.
Roll, R., 1984. A simple implicit measure of the effective bid-ask spread in an efficient market. J. Finance 39 (4), 1127–1139.
Rouwenhorst, K.G., 1998. International momentum strategies. J. Finance 53 (1), 267–284.
Rouwenhorst, K.G., 1999. Local return factors and turnover in emerging stock markets. J. Finance 54 (4), 1439–1464.
Schwert, G.W., 2003. Anomalies and market efficiency. Handb. Econ. Finance 1, 939–974.
Sirignano, J., Sadhwani, A., Giesecke, K., 2016. Deep Learning for Mortgage Risk. arXiv preprint arXiv:1607.02470.
Sloan, R.G., 1996. Do stock prices fully reflect information in accruals and cash flows about future earnings? Account. Rev. 289–315.
Soliman, M.T., 2008. The use of dupont analysis by market participants. Account. Rev. 83 (3), 823–853.
Spiess, D.K., Affleck-Graves, J., 1995. Underperformance in long-run stock returns following seasoned equity offerings. J. Financ. Econ. 38 (3), 243–267.
Stambaugh, R.F., Yu, J., Yuan, Y., 2012. The short of it: investor sentiment and anomalies. J. Financ. Econ. 104 (2), 288–302.
Thomas, J.K., Zhang, H., 2002. Inventory changes and future returns. Rev. Account. Stud. 7 (2), 163–187.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. B 267–288.
Titman, S., Wei, K.J., Xie, F., 2004. Capital investments and stock returns. J. Financ. Quant. Anal. 39 (4), 677–700.
Titman, S., Wei, K.J., Xie, F., 2013. Market development and the asset growth effect: international evidence. J. Financ. Quant. Anal. 48 (5), 1405–1432.
Tobek, O., Hronec, M., 2018. Does source of fundamental data matter? .
Tuzel, S., 2010. Corporate real estate holdings and the cross-section of stock returns. Rev. Financ. Stud. 23 (6), 2268–2302.
Watanabe, A., Xu, Y., Yao, T., Yu, T., 2013. The asset growth effect: insights from international equity markets. J. Financ. Econ. 108 (2), 529–563.
Whited, T.M., Wu, G., 2006. Financial constraints risk. Rev. Financ. Stud. 19 (2), 531–559.
Zhang, X., 2006. Information uncertainty and stock returns. J. Finance 61 (1), 105–137.