The Promises and Pitfalls of Machine Learning for Predicting Cross-Sectional Stock Returns[✧]

Edward Leung[*], Harald Lohre[†], David Mischlich[‡], Yifei Shea[+] and Maximilian Stroh[#]

March 1, 2020

Abstract

Recent research suggests that machine learning models dominate traditional linear models in predicting cross-sectional stock returns. Indeed, we confirm this finding when predicting one-month forward looking returns based on a set of common equity factors, including predictors such as short-term reversal. Despite this statistical advantage of machine learning model predictions, we demonstrate economic gains to be more limited and critically dependent on the ability to take risk and implement trades efficiently. Unlike traditional models, machine-learning models have struggled less in discerning valuable predictions from cross-sectional equity characteristics.

JEL classification: G11, G12, G14, G15, G17

Keywords: Machine Learning, Data Science, Interpretable Machine Learning, Return Prediction, Cross-Section of Stock Returns, Gradient Boosting, Factor Investing

[✧] We'd very much like to acknowledge the feedback and comments by Livia Amato, Carlos Carvalho, Ronald Hochreiter, Eyal Kenig, Daniel Philps and participants at the "International Conference on Data Science in Finance with R" at WU Vienna and the Cambridge Judge Business School conference "AI/ML in Finance: Is It Just Another Bubble?". Note that this paper expresses the authors' views which do not necessarily coincide with those of Invesco or Quoniam.

[*] Invesco Quantitative Strategies, 1166 Avenue of the Americas, New York, NY 10036, USA

[†] Invesco Quantitative Strategies, An der Welle 5, 60322 Frankfurt/Main, Germany; Centre for Financial Econometrics, Asset Markets and Macroeconomic Policy (EMP), Lancaster University Management School, Bailrigg, Lancaster LA1 4YX, United Kingdom

[‡] Invesco Quantitative Strategies, An der Welle 5, 60322 Frankfurt/Main, Germany

[+] Invesco Quantitative Strategies, 100 Federal St, Boston, MA 02110, USA

[#] Quoniam Asset Management, Westhafenplatz 1, 60327 Frankfurt/Main, Germany

The last decade has seen machine learning (ML) methods automating complex tasks at an unprecedented pace that was unthinkable ten years ago. For instance, IBM's Watson impressively beat two of the best Jeopardy players in 2011. Five years later, DeepMind's AlphaGo came to the spotlight, beating Go champion Lee Sedol 4-1. More recently, in 2018, Alphabet's Waymo self-driving cars drove 1.2 million miles in California, with human intervention necessary only every 11 thousand miles.[1] Such advancements in ML are driven by the availability of data and computing power. A frequent statement is that 90% of today's data has been created in the last two years.[2] Similarly, computing power has been growing exponentially for a long time. To illustrate, Brynjolfsson and McAfee (2016) mention ASCI Red, which was the world's fastest supercomputer in 1996. It was operated by the U.S. government, costing $55 million and reaching 1.8 teraflops[3] in 1997. By 2006, Sony's PlayStation 3 was delivering 1.8 teraflops at a $500 price point. Currently, the fastest supercomputer is Summit at the Oak Ridge National Laboratory (USA), delivering 149 petaflops[4]. It is about 83,000 times faster than ASCI Red.

Machine learning applications in finance are nothing new. For instance, Hutchinson, Lo and Poggio (1994) show that learning networks can recover the Black-Scholes formula from a two-year training set of daily options prices. Levin (1995) uses multilayer feedforward neural networks to predict a stock's excess returns based on its exposure to a variety of factors. However, these early ML papers rarely stood their ground in real-world asset management applications.

Recent years have witnessed a resurgence of ML research for predicting the cross-section of stock returns. Huerta, Corbacho and Elkan (2013) train a support vector machine to classify stocks into future outperformers and underperformers. Moritz and Zimmermann (2016) use random forests to study how lagged monthly stock returns predict future stock returns. Their results are robust to adding 86 firm characteristics from the literature. Coqueret and Guida (2018) also rely on regression trees to predict returns over the following month, three months and twelve months. Again, past return-based predictors are deemed most important.

[1] State of California, Department of Motor Vehicles: Autonomous Vehicle Disengagement Reports 2018.
[2] This statement might have originated from a 2013 ScienceDaily report.
[3] A teraflop is a unit of computing speed equal to 10^{12} floating-point operations per second.
[4] A petaflop is a unit of computing speed equal to 1000 teraflops.

Similarly, Messmer (2017) investigates a set of 68 characteristics to predict U.S. stock returns, documenting short-term reversal and twelve-month momentum as the most important predictors. Instead of a regression tree-based method, he uses feedforward neural networks. Freyberger, Neuhierl and Weber (2020) use adaptive group LASSO to estimate expected stock returns based on a set of 62 characteristics. The authors find past return-based predictors to play a prominent role, but size is relevant as well. Gu, Kelly and Xiu (2020a) compare a multitude of ML methods for stock return prediction, ranging from partial least squares to feedforward neural networks. While the importance of predictors depends on the chosen ML method, short-term reversal is clearly the most relevant predictor, and further past return-based characteristics as well as size and dividend yield are the runner-ups. Rasekkschaffe and Jones (2019) train four different ML algorithms based on three different ways to construct training sets to arrive at a total of twelve ML models, ultimately observing that an ensemble of these twelve models improves stock return predictions. Chen, Pelger and Zhu (2019) integrate no-arbitrage conditions into the architecture of a combined feedforward network, recurrent Long-Short-Term-Memory (LSTM) network and generative adversarial network (GAN). Feng, Polson and Xu (2019) use deep learning to create a linear factor model directly from raw firm characteristics, whereas other articles rely on preprocessing characteristics before they feed them to supervised ML algorithms. Gu, Kelly and Xiu (2020b) develop a latent factor conditional asset pricing model using autoencoders.

The general tenor of these articles is that ML models are superior to traditional linear factor models in predicting subsequent cross-sectional stock returns. Such findings can be rationalized in various ways: first, traditional linear factor models may not fully capture the cross-sectional risk and return tradeoffs, see Luo, Jussa, and Wang (2017). Second, measuring equity factor premia is ultimately about prediction, and ML methods are explicitly designed for prediction tasks. Third, the number of predictive stock characteristics reported in the literature has exploded, and predictors are often highly correlated, putting classical prediction methods at a disadvantage. ML techniques are explicitly designed to tackle such scenarios using variable selection and dimension reduction. Finally, ML is designed to handle "model uncertainty" – that is, to determine relevant inputs and an adequate functional form: Is the relationship between the dependent variable and

predictors linear? If the relationship is non-linear, what functional form is appropriate? Are there interactions among predictors?

On the other hand, there is an emerging literature that alludes to potential pitfalls in stock return predictability. Avramov, Chordia and Goyal (2006) show that a trading strategy based on short-term reversal is defeated by transaction costs. Such practical concerns are highly relevant with respect to machine learning-based strategies, as most ML models that predict stock returns one month ahead tend to rely largely on high turnover factors such as short-term reversal. Hou, Xue and Zhang (2020) replicate 452 cross-sectional anomalies from the finance and accounting literature and show that 65% of them are not significant when controlling for microcap stocks. The failure rate increases to 82% when multiple testing issues are also accounted for. Furthermore, the proper application of ML to the investment setting is no easy task. Lopez De Prado (2018) and Arnott, Harvey and Markowitz (2019) discuss how misapplying machine learning techniques can lead to promising in-sample results that are unlikely to translate to meaningful out-of-sample performance.

The open question and motivation for our paper is thus: How robust to the pitfalls described in the previous section are the promising results from the literature that applies machine learning in order to predict the cross-section of stock returns?

The contribution of this paper is twofold: First, we discuss how ML might improve predicting the cross-section of equity returns based on well-known equity factors. In particular, we demonstrate ways to rationalize the mechanics and outcome of a given ML model, thus alleviating its black-box character. Second, we investigate whether sizeable improvements in predictability exist and whether they translate into an improvement of associated real-world portfolio simulation performance.

To this end, we briefly introduce a machine learning method called Gradient Boosting Machine (GBM), which is based on regression trees and is used as the ML method of choice throughout the paper. To foster intuition, we first build a GBM model based on six common equity factors and compare it to a classic Ordinary Least Squares (OLS) regression model based on the same factors. The factors are: Size, Book-to-Market, Profitability and Asset Growth (Investment), from the Fama-French five factors model, as well as Momentum

and Short-Term Reversal. Resorting to methods from the Interpretable Machine Learning literature, we analyze the GBM model and document the following observations: First, the relative importance of predictors within each model is comparable across the OLS and the GBM models; for a one-month prediction horizon, the factors Momentum, Short-Term Reversal and Profitability tend to be more important than Asset Growth, Book-to-Market and Size. Second, the OLS-betas all have the "correct" signs according to economic theory and resonate with the sensitivities inferred from partial dependence plots for the GBM model. Third, there are still substantial non-linearities in the GBM model; for instance, stocks with a very low Short-Term Reversal score (a very low return in the previous month) tend to keep underperforming (and not reverse). Fourth, some predictors interact strongly within the fitted GBM model; consistent with the evidence presented in Zhu and Yung (2016), we find a positive interaction between Short-Term Reversal and Momentum.

Next, we assess the potential of machine learning to improve upon the predictive performance of simpler factor models. Using a set of up to twenty investable global equity factors, we compare four simple signals based on signal-averaging and linear regression to four GBM prediction models. We show that GBM signals indeed strongly outperform simpler linear signals in portfolio sorts before portfolio constraints or market frictions are considered. Obviously, accounting for the latter considerations may alter the assessment of the practical relevance of such improvements in prediction performance. In this vein, Avramov, Cheng and Metzker (2019) study the robustness of machine learning signals under real-world frictions and show that such signals extract profitability from difficult-to-arbitrage stocks as well as during alleviated limits-to-arbitrage market states. They compare equal-weighted to value-weighted long-short portfolios and subsamples of data with economic restrictions to study implementation hurdles. While the academic study of Avramov, Cheng and Metzker (2019) greatly improves the understanding of some pitfalls associated with ML-based cross-sectional equity predictions, we investigate whether machine learning techniques are truly attractive for practitioners seeking to transform a variety of factor scores into return predictions that can be used for portfolio optimization.

To this end, we perform portfolio simulations that account for realistic portfolio constraints and limited risk-taking ability as well as frictions, such as transaction costs, shorting costs and non-instantaneous

implementation of trades. We implement global portfolio simulations based on three different strategy setups: 1) market-neutral, 2) long-only 3% tracking error and 3) long-only 1% tracking error, with each strategy utilizing the eight signals as inputs. In such realistic portfolio simulations, the GBM signals do still outperform simpler linear benchmark signals. Yet the size and statistical significance of the economic benefits of machine learning relative to simpler linear benchmark signals crucially depend on the portfolio's ability to take risk and on efficient trade implementation. Especially, in case of the long-only 1% tracking error strategy, the performance difference between the best-performing linear and non-linear signal is considerably less compelling, and none of the four GBM-based signals has an information ratio significantly different from that of simpler models.

1. Machine learning with boosted regression trees

In this section, we briefly introduce two popular machine learning techniques, regression trees and gradient boosting, the latter of which will be used for predicting stock returns throughout the paper.[5]

1.1. Regression trees

The regression tree method uses decision trees to arrive at a prediction for a target variable y given a set of predictor variables x. To illustrate, we consider a regression tree predicting stock returns over the following month based on two stock characteristics, size and value, as predictors. Suppose the two characteristics' scores have been transformed to a uniform [0,1]-distribution. Figure 1 shows a hypothetical fitted regression tree. Stocks are sorted by size and assigned to bin 1 when smaller than 0.5. Larger cap stocks are then sorted by their value characteristic and assigned to bins 2 and 3 using a split value of 0.3. The right-hand side of Figure 1 illustrates how the predictor space is divided into disjoint bins (subsets). Stock return forecasts in each bin are defined as the simple average of the stocks' returns in subsequent months for bin 1, bin 2 and bin 3, respectively.[6] The determination of optimal splits depends on the specific algorithm and loss function. The basic idea is to split the predictor space such that the residuals of the fitted model are minimized.[7] Such simple regression trees form an important ingredient to the more powerful ML technique of gradient boosting.

[5] See Hastie, Tibshirani, and Friedman (2017) and Efron and Hastie (2017) for a broader introduction to ML methods.
[6] For a more detailed introduction to regression trees, see sections 9.2.1 and 9.2.2 in Hastie, Tibshirani, Friedman (2017).
[7] We use the open-source machine learning platform H2O-3. Its GBM algorithm uses different forms of histogram aggregation to speed up the splitting process instead of going through all possible split values.

1.2. Gradient Boosting Machine

Gradient Boosting Machine (GBM) is a popular ensemble method which combines multiple weak learners to create a strong learner. We consider a common setup of GBM where weak learners are formed from shallow regression trees. GBM sequentially adds regression trees to an ensemble, each one correcting its predecessor. This boosting method aims to fit a new regression tree to the residual errors made by the previous tree.

Almost every machine learning algorithm comes with some parameters that need to be specified to adjust model complexity. For instance, the regression tree algorithm requires the number of splits. Such parameters are called hyperparameters, distinguishing them from model parameters (e.g. split values) that are estimated on the training dataset. For GBM, the most important hyperparameters are the number and depth of the trees, the learn rate and the minimum number of observations for a leaf. We use a specific variation of GBM, called stochastic gradient boosting[8], further growing the list of relevant hyperparameters by the number of observations and columns to sample for each residual tree.

In machine learning, "tuning" a model describes the process of searching for hyperparameters that maximize prediction performance outside the training set. For tuning purposes, we use random search as introduced by Bergstra and Bengio (2012). For each hyperparameter deemed relevant, we define a probability distribution from which to sample its values. Next, we form various models, each defined by drawing a value for each of the hyperparameters. In our case, each model will compete based on mean absolute error (MAE) using validation sets. Training and validation of models are done sequentially, i.e. validation sets always come after training sets. The ultimate choice of the tuned model depends on the average validation score over a series of validation sets, based on a series of models estimated on the associated series of training sets. Our method thus replicates the process of how models are trained on expanding windows of pooled data and are then used to make out-of-sample predictions. Based on this tuning method, we aim to determine a vector of hyperparameters that achieves the latter objective reasonably well. A more detailed description of the hyperparameter tuning process is given in the Appendix.

[8] See Friedman (2002).

2. A simple non-linear ML model laid bare

2.1. Data and methodology

We test the efficacy of machine learning to predict stock returns in a global developed market universe of large and mid-cap stocks.[9] The dataset spans from January 1991 to December 2018 and considers 1,109,595 stock-months that satisfy the above constraints. Of these, 31% are from the U.S., 20% are from Europe ex U.K., 17% are from Japan, 14% from the U.K., 10% from Australia and 8% from the rest of the world. Given that we load data from 336 months, we consider an average of 3302 stocks per date, and this number increases over time, from 2401 stocks in January 2001 to 3671 stocks in December 2018. We exclude stocks with very small market capitalization. For example, this gives us a minimum market capitalization of $166 million for U.S. stocks and a time-average for the 1% percentile of $530 million. The time-average of the median U.S. market cap for stocks included in our sample is $3,334 million.

To predict cross-sectional stock returns, we generally build on a set of twenty equity factors, each belonging to one of four concepts: Momentum, Quality, Size and Value. Table 1 lists all factors and the concept to which each of them belongs. For the time being, our focus is on rationalizing the outcome of GBM models, and we thus only consider the six factors: Size, Book-to-Market, Profitability and Asset Growth (Investment), from the Fama-French five factors model (FF5), as well as Momentum and Short-Term Reversal.[10]

Note that we do not directly predict a stock's total return over the following month, but instead predict its outperformance or underperformance relative to its peers in the same region and industry. To this end, we apply a ranking-based standardization procedure to monthly total returns, standardizing by date, industry and region. This procedure, which is also applied to the twenty predictive factors, is beneficial for three reasons. First, one avoids portfolios with static industry tilts over time. While this requirement can be enforced by setting portfolio constraints in the portfolio optimization, we recommend already creating signals accordingly. Second, standardizing by each date neutralizes market shocks that affect all stocks, which we do not expect to be

[9] We restrict ourselves to non-financial stocks, as some of the used fundamental factors are less applicable to financial stocks.

[10] We abstract from the market factor, because we are predicting excess returns against peers. As for momentum, we use the 12M total return lagged by one month. The short-term reversal factor is based on the total return of a stock over the previous month.

predictable based on equity factor scores alone. Third, using ranks reduces the effects of outliers and conforms with sorting portfolios into quantiles.[11]

We seek to predict for stock i the standardized future total return $r_{i,t\to t+h}$ between date t and date $t+h$ based on the standardized factor scores $s_{i,t}^{(j)}$ of a given factor *j*. Hence, we are looking for a function

$$r_{i,t\to t+h} = \hat{f}\left(s_{i,t}^{(1)}, \dots, s_{i,t}^{(p)}\right) + \varepsilon_{i,t\to t+h} \quad (1)$$

that minimizes the error term $\varepsilon_{i,t\to t+h}$, and thus effectively explains a stock's future outperformance or underperformance relative to its peers. Note that observations are pooled across time and the cross-section of stocks in our estimation of $\hat{f}$. Furthermore, this function must not overfit the training data. The promise of ML is to help in determining a function $\hat{f}$ that works well on new data not used in the estimation of $\hat{f}$.

2.2. Dismantling a non-linear ML model

An increase in the number of parameters in a given ML model can quickly render the ensuing prediction models highly complex. To foster the intuition, we first consider parsimonious ML models and calibrate the GBM method with a limited set of six factors and compare this model to a classic OLS regression model based on the same factors.

2.2.1. OLS-betas versus variable importance in ML models

We consider the six factors: Size, Book-to-Market, Profitability and Asset Growth (Investment), from the Fama-French five factors model (FF5), as well as Momentum and Short-Term Reversal. For the OLS regression, $\hat{f}$ has the familiar form

$$r_{i,t\to t+h} = \alpha + \beta_1 s_{i,t}^{(1)} + \cdots + \beta_6 s_{i,t}^{(6)} + \varepsilon_{i,t+1} \quad (2)$$

[11] Rasekkschaffe and Jones (2019) also advocate industry or sector neutralization and normalize the factors; however, they prefer forecasting categories instead of returns. In addition, Avramov et al. (2019) show that an intra-industry strategy delivers substantially higher returns than an inter-industry strategy. Finally, in unreported results we find that using ranks gives similar results to a Huber-weighted loss function, if the rest of the standardization procedure is unchanged.

For GBM, the resulting function $\hat{f}$ is too complex to spell out. We estimate a series of OLS and GBM models based on expanding data windows.[12] For OLS, we estimate a linear regression model using data from January 1991 to December 2000 in order to predict returns one month ahead from December 2000 to January 2001, plugging in factor scores as of December 2000. We continue estimating linear regression models using expanding windows of data, i.e. using data from January 1991 to January 2001, then January 1991 to February 2001 and so forth. We obtain a time series of betas for each of the six factors, see the upper chart in Figure 2.

Stocks with high Momentum, high Profitability and a high Book-to-Market ratio relative to their peers tend to outperform in the following months, while stocks with high returns in the previous month, high Asset Growth, and large Size tend to underperform their peers over the following month.[13] To get a sense of the factors' relative importance, we also plot the absolute values of the factor betas, normalized by the sum of absolute betas, see the middle chart in Figure 2. Momentum is the most important factor in relative terms, and its importance has increased in recent years given that other factors have worked less well and their betas have decreased towards zero. The two next most important factors are Short-Term Reversal and Profitability. Asset Growth and Book-to-Market are of less importance; however, Size is the least important factor in the linear six-factor OLS model. Next, we compare these results to those emerging from a series of GBM models estimated using the same data and expanding windows. Given the complex functional form of a GBM model, there is no direct analogue to linear regression betas. Still, a meaningful remedy is to look at variable importance[14] over time. Similar to the analysis for the linear model, the lower chart in Figure 2 documents that Momentum is the most important factor in the GBM model. Moreover, its relative importance increased over the second half of the studied period. Again, Short-Term Reversal is of high importance. There is a noticeable difference with respect to Profitability, which the GBM model deems close to Asset Growth and Book-to-

[12] In this section, we do not tune the hyper-parameters of the GBM model, as it is fitted just for illustrative purposes.
[13] Even though momentum and short-term reversal factors have been added to the FF-set and factors and returns are standardized using ranks, the signs of the remaining four factors are consistent with Fama and French (2015).
[14] H2O-3 calculates variable importance in the following way: For every split in building a regression tree, it measures which variable was selected and by what amount the squared error was reduced as an effect of the split (http://docs.h2o.ai/h2o/latest-stable/h2o-docs/variable-importance.html#variable-importance-calculation-gbm-drf).

Market in terms of importance. The remaining three factors in the GBM model are similar to what we observed for the OLS model, with Size being roughly half as important as Asset Growth and Book-to-Market.

2.2.2. Partial dependence plots

Another route to studying the GBM model corresponds to the notion that linear regression betas can be interpreted as sensitivities. A predictor's beta informs about the prediction change when the corresponding predictor changes, keeping all other predictors constant. Such sensitivity analysis can even be performed for more complex models: One simply needs to vary a given input predictor while keeping the others fixed and observe the ensuing change in the model prediction. We follow Friedman and Popescu (2008) in formalizing this concept. Given any fitted model $\hat{f}$ and any vector of features $s_j = (s^{(j_1)}, \dots, s^{(j_k)})$, we define the partial dependence function $\hat{f}_j$ by

$$\hat{f}_j(s_j) = \frac{1}{N}\sum_{i=1}^{N} \hat{f}(s_j, s_{i\setminus j}) \tag{3}$$

where $s_{i\setminus j}$ is the value of the i-th observation in the training dataset of all other features on which the model has been fitted that are not part of s_j, i.e. $s_{.\setminus j} = (s^{(j_{k+1})}, \dots, s^{(j_p)})$ where $\{j_1, \dots, j_k\} \cap \{j_{k+1}, \dots, j_p\} = \emptyset$ and $\{j_1, \dots, j_k\} \cup \{j_{k+1}, \dots, j_p\} = \{1, \dots, p\}$. Hence, for every value of s_j, we calculate the average prediction over all "observations" $(s_j, s_{i\setminus j})$. Note that s_j can take any value, while $s_{.\setminus j}$ takes only values given in the training set. For example, the following function describes the average effect of the first (standardized) factor score, accounting for the effect of all other factors.

$$x \mapsto \frac{1}{N}\sum_{i,t} \hat{f}\left(x, s_{i,t}^{(2)}, \dots, s_{i,t}^{(p)}\right) \tag{4}$$

A chart of a partial dependence function is called a partial dependence plot,[15] and we make use of such a plot to contrast the sensitivities of the fitted GBM model to the fitted OLS model, see Figure 3. By and large, the direction of the effects in the GBM model is in line with that of the OLS model. High Momentum, Book-

[15] See section 10.13.2 in Hastie, Tibshirani, and Friedman (2017) and Greenwell, Boehmke, and McCarthy (2018).

to-Market and Profitability all lead to higher expected returns (compared to peers), while Short-Term Reversal, Asset Growth (Investment) and Size lead to lower return expectations. Yet there are also some non-linearities detected by the GBM model that a linear model cannot pick up: Stocks with a very low Short-Term Reversal score, i.e. a very low return in the previous month, tend to keep underperforming (and not reverse). High Asset Growth seems to be relatively more detrimental to future returns than low Asset Growth is beneficial. Finally, profitability exhibits a distinct non-linear pattern characterized by a dichotomous nature of sensitivities: Companies whose profitability score is below the median experience a fairly constant drag on future returns. Conversely, highly profitable companies should expect to earn the same return premium as those slightly above the median profitability in the sample.

2.2.3. Interactions

Beside the ability to capture non-linearities, ML models account for potential interaction effects of factors. In the linear regression model, factors are additive and do not influence one another. Yet the six-factor GBM model can contain interactions. For example, whether Momentum is high or low may impact the efficacy of Short-Term Reversal (or any other factor). Friedman and Popescu (2008) quantify the interaction of two features by comparing their two-dimensional partial dependence function with the sum of their univariate partial dependence functions.[16] To tease out relevant pairwise interactions, we take the absolute value of this difference and average across a grid of possible values. The results are displayed in Figure 4, documenting the strongest pairwise interaction for Short-Term Reversal and Momentum.

To further gauge the strength and direction of these two interactions, we can plot the difference between the partial dependence function of Short-Term Reversal and Momentum and their respective univariate partial dependence functions:

$$\hat{f}_{STR,MOM}(x,y) - \hat{f}_{STR}(x) - \hat{f}_{MOM}(y) \tag{5}$$

Across the various Short-Term Reversal and Momentum realizations, this results in a heatmap as portrayed in Figure 5, which illustrates the difference between the two-dimensional partial dependence function

[16] In fact, their results also hold for higher-dimensional interactions, where one looks for interactions between two sets of features. Note that due to the pre-processing, our y-variable is demeaned.

and the sum of the two one-dimensional partial dependence functions on a product rule grid consisting of 121 points (in each dimension we use 11 different score values to create the grid).

To foster intuition, consider a stock with very high Momentum and Short-Term Reversal scores, i.e. a stock that outperformed its peers over the previous year as well as over the last month. Judging by the univariate partial dependence plots in Figure 3 alone, one might infer that the stock's expected outperformance is slightly negative (since the expected outperformance based on Short-Term Reversal is strongly negative, whereas the one based on Momentum is only slightly positive). However, it turns out that there is a strong positive interaction between the two factors, and the expected outperformance given high Short-Term Reversal and high Momentum is positive. Overall, Figure 5 shows that the efficacy of the Momentum factor is much higher for stocks that strongly outperformed in the last month than for those that underperformed substantially, consistent with Zhu and Yung (2016).[17]

3. A horse race of ML models to predict the cross-section of equity returns

Having become familiar with the nature and mechanics of GBM models using a parsimonious set of common equity factors, we move to the broader set of twenty factors, allowing us to better assess the potential of machine learning to improve upon the prediction performance of simpler factor models. In particular, we consider gradient boosting applied to four different prediction contexts:

- GBM-1M-6F: We predict stock returns one month ahead, focusing on the predictors Size, Book-to-Market, Investment, Profitability, Momentum and Short-Term Reversal.
- GBM-1M-20F: We predict stock returns one month ahead using twenty factors as predictors.
- GBM-6M-6F: We predict stock returns six months ahead, focusing on the predictors Size, Book-to-Market, Investment, Profitability, Momentum and Short-Term Reversal.
- GBM-6M-20F: We predict stock returns six months ahead using twenty factors as predictors.

[17] Running the same analysis for the six-factor OLS model would result in a table with every entry being equal to zero, as there are no interactions by construction. Of course, it would be possible to manually include interaction terms in a linear regression, which then could show up in an interaction effect heatmap.

Note that we include the two models from the previous section in the analyses. Yet, unlike in Section 2, each of the above four signals is estimated with hyperparameter tuning based on the random search methodology detailed in the Appendix. For benchmarking of these ML models, we resort to four simpler signals:

- 1/N: Equal-weighted average of all twenty factors.
- Concepts: First, take the equal-weighted average of factors within the broader concepts[18] Momentum, Quality and Value. Second, take the equal-weighted average of concept scores.
- OLS-1M-6F: We predict stock returns one month ahead using a linear regression based on the predictors Size, Book-to-Market, Investment, Profitability, Momentum and Short-Term Reversal.
- OLS-6M-6F: We predict stock returns six months ahead using a linear regression based on the predictors Size, Book-to-Market, Investment, Profitability, Momentum and Short-Term Reversal.

Being more flexible, the GBM model might discern more structure in the data than a simple OLS model. Still, we must examine whether there is any predictive value-added in fitting more complex models to the data. In fact, given the low signal-to-noise ratio in financial markets, more complex models are prone to overfitting and therefore may deliver worse out-of-sample performance than simpler models. In this regard, Suhonen et al. (2017) show that the reduction in live versus backtested Sharpe of the most complex alternative beta strategies is over 30% higher than that of the simplest strategies.

3.1. Prediction performance

For each of the above modeling approaches, we estimate a series of models, generating predictions of a stock's outperformance or underperformance relative to its region and industry peers over the following month (or the following six months). This estimation is performed monthly from December 2000 to November 2018, while the hyperparameter tuning procedure for the GBM models is updated every 12 months. The

[18] A concept is a group of similar equity factors, in our case the groups are Momentum, Quality, and Value, see **Table 1** for the specific factor groupings.

hyperparameter vector is therefore kept constant within a given year, whereas the model is retrained monthly to reflect the most recent data.

Table 2 shows pairwise correlations between the generated signals. We observe that each of the two six-factor OLS signals exhibits a correlation of around 0.65 with the corresponding GBM signal when keeping the prediction horizon fixed. Coincidentally, the six-factor GBM signals exhibit a correlation similar to the 20-factor counterpart (0.68). Of course, bringing in more factors renders GBM model predictions less correlated with the OLS benchmark models (around 0.5). Similar observations apply to the simple equal-weighted benchmarks.

To evaluate prediction performance, we sort stocks into deciles according to the predicted outperformance or underperformance: Each month, we assign the 10% of stocks with the lowest predictions to an equal-weighted portfolio called D1. Likewise, stocks with the highest predictions go into the top decile portfolio called D10. Next, we evaluate the average monthly performance of top and bottom decile portfolios, as well as the decile spread of the long-short portfolio D10-D1, see Table 3.[19]

Consistent with Gu, Kelly and Xiu (2020a), Messmer (2017), Coqueret and Guida (2018), Rasekkschaffe and Jones (2019) and Chen, Pelger and Zhu (2019), the GBM models substantially outperform simpler models when a larger set of factors is considered. The best performing model is GBM-1M-20F. It displays an average monthly return in the long-short portfolio D10-D1 of 1.34%, of which 0.94% can be attributed to the short-leg and 0.40% to the long-leg. Together with a monthly volatility of 1.86%, this amounts to an (annualized) information ratio of 2.50 and a t-stat of 10.63, which compares favorably with benchmark information ratios between 0.97 for the OLS-6M-6F signal and 1.99 for the 1/N signal.

Admittedly, signal decay is clearly high for the GBM-1M-20F model. The information ratio drops from 2.50 to 1.65 when we lag the signal by two months. Similarly, the OLS-1M-6F and GBM-1M-6F models both display substantial signal decay. One way to mitigate this is to increase the prediction horizon of the models.

[19] Returns are relative to the median stock return in each month, i.e. before calculating the equal-weighted decile portfolio returns for a month, we subtract the median return across all stocks in that month from every stock return.

For instance, the information ratio of the GBM-6M-20F model only declines from 1.90 to 1.77 when the signal is lagged by two months, and the six-factor OLS and GBM models also exhibit more stable signal efficacy. While the raw performance of the GBM-6M-20F model is not completely on par with the GBM-1M-20F model, this stability could prove valuable when transaction costs are considered.

3.2. A practitioner's view

While the out-of-sample prediction performance of the 20-factor GBM models is impressive, it is still unclear whether this predictive power can be turned into economic gains in the context of realistic portfolio simulations. Therefore, we next account for various issues that could lead to muted performance in actual portfolio implementation.

First, at the one-month forecast horizon, the 20-factor GBM model could be unduly driven by the hard-to-implement Short-Term Reversal factor. Table 4 documents the time average percentage variable importance of the two GBM models based on twenty factors. Indeed, Short-Term Reversal is the most influential factor in the one-month prediction horizon model, while it is among the least important factors for predicting six months ahead. Of course, this finding resonates with the significant differences in signal decay documented earlier. To execute trades for a large fund can take days, and the ideal of a contemporaneous calculation and implementation of a given signal is therefore hardly realistic. For slow moving factors, such as the Fama-French factors, ignoring implementation lag is not a substantive problem. But, for the Short-Term Reversal factor, an implementation lag of a few days can make a big difference.

Second, while reversal factors have been known for many years (Jegadeesh, 1990), such factors also induce high turnover and are thus difficult to profitably implement when transaction costs and other frictions are considered, see Avramov, Chordia and Goyal (2006). Moreover, one needs to consider the lending fees to be paid in the case of short positions.

Third, we wonder how much value-added can be enjoyed in long-only mandates. Common prediction performance metrics, such as R-Squared, information coefficient (IC) or decile spread returns, relate to long and short positions. While such practice allows an evaluation of a signal's predictive power, one has to

acknowledge that not all stocks are available for shorting at all times, and that the costs of shorting can be significant. Especially in the case of long-only strategies, there is a zero lower limit for underweight positions. Furthermore, a real-world portfolio is typically subject to constraints on active risk and often has limited capacity for country, industry or other top-down exposures.

To address the above concerns, we compare the eight signals' performance in portfolio simulations where portfolios are rebalanced monthly using Axioma's portfolio optimizer and global risk model. For each signal, we simulate three different strategies: A market-neutral strategy at 5% active risk and two long-only strategies with MSCI World ex Financials as the benchmark index (1% and 3% active risk, respectively). Each strategy is restricted to have minimal active sector, industry and country bets.[20]

3.2.1. Market-neutral

We consider a market-neutral strategy that can take long and short positions in global developed market stocks, excluding all financial stocks. Portfolios are rebalanced monthly based on mean-variance optimization where the risk aversion parameter is set to give an ex-post tracking error around 5% p.a.;[21] the strategy is dollar and beta-neutral. Sector, industry, region, country and currency exposures are also kept close to zero. The strategy is calculated including a 100% cash investment, and hence we compare its returns against the three months USD-LIBOR rate.

Panel A of Table 5 presents portfolio simulation results for the market-neutral strategy from December 2000 to November 2018. All strategies have ex-post active risk between 4.35% and 5.53% p.a., consistent with expectations. In terms of active returns, simple signal averaging methods (1/N and Concepts) have lower gross returns compared to other signals, leading also to lower net active returns (around 2% and less than 1%, respectively) despite their relatively low annual turnover (one-way 130% and 187%, respectively). As a result, these two simple signals also exhibit the lowest net information ratios among all signals (0.37 and 0.15, respectively).

[20] To mitigate the implementation lag issue, we lag the Short-Term Reversal factor by five business days between factor computation and actual implementation.

[21] Note that we use the same risk aversion parameter for all signals. That is, we do not adjust the risk aversion parameter to have every strategy exhibiting a 5% ex-post tracking error.

Next, we discuss the six regression-based models, focusing first on turnover: signals based on a one-month prediction horizon tend to produce higher turnover than those based on a six-month prediction horizon. For example, the OLS-1M-6F signal implies an annual turnover of 224%, while the turnover of the OLS-6M-6F signal is only 119%. Keeping the forecast horizon fixed, GBM signals have higher turnover than linear signals. With one-way annual turnover of 321% for the GBM-1M-6F signal and 309% for the GBM-1M-20F signal, both are close to the imposed upper turnover limit of 27.5% per month.[22]

The differences in turnover and transaction costs have direct implications for the strategies' net returns and net information ratios. For instance, the GBM-1M-6F signal has some 4% annual return advantage over OLS-1M-6F before costs. However, the annual net return difference is just around 2%. Before costs, GBM-1M-6F exhibits an information ratio of 2.32, which drops to 1.06 after costs. Nevertheless, it is still superior to that of the OLS-1M-6F signal (0.56).

Overall, the market-neutral portfolio simulation setting sees GBM-based signals clearly outperforming simpler approaches such as signal averaging or OLS-based linear factor models, see Figure 6. Given conservative transaction cost and implementation lag assumptions, the benefits of GBM models could be even larger for investors who can trade at low cost and without much delay. To gauge the statistical relevance of the documented value-added of the GBM models, Panel B of Table 5 gives test statistics for pairwise hypothesis tests assuming equal information ratios. In almost all cases, we can reject the null hypothesis of equal information ratios between a GBM-based signal and one of the benchmark signals (5% level). In many cases, significance even holds at the 1% level; this finding applies, for instance, to all pairings of the GBM-6M-20F signal with the four benchmark signals.

3.2.2. Long-only 3% tracking error

Next, we study a long-only strategy with 3% tracking error relative to the MSCI World ex Financials. The strategy keeps portfolio betas as well as sector, industry, region, country and currency exposures close to that of the benchmark. A one-way turnover constraint of 7% is imposed for monthly rebalancing.

[22] In unreported results, we find that gross information ratios for the GBM-1M-20F signal do not suffer under the turnover constraints for any of the three strategies, as simulations without turnover constraints yield similar gross IRs.

The results from the corresponding portfolio simulations for the eight signals can be found in Panel A of Table 6. Similar to the market-neutral strategy setting, the simple signal averaging methods (1/N and Concepts) have relatively low gross average returns. In a similar vein, one-month prediction horizon signals tend to produce higher turnover than the six-month prediction horizon signals. In the case of GBM signals, both the 1M-6F and 6M-6F models are very close to the annual turnover limit of 84%. Comparing GBM-6M-6F versus OLS-6M-6F, the return advantage of GBM shrinks from 1% (gross) to just over 30 bps (net) because of the associated high turnover; the net information ratio of GBM-6M-6F is therefore slightly higher compared to OLS-6M-6F (as shown in Figure 6). Note that the simulated 3% long-only strategy based on the GBM-6M-20F signal still has some 1% return advantage compared to the strategy based on the OLS-6M-6F signal, combined with slightly lower active risk ex post, as shown in Panel A of Table 6. GBM-6M-20F still has the highest net information ratio among all (Figure 6).

By and large, the observations from the market-neutral simulations are confirmed in the long-only simulations: GBM outperforms simpler approaches such as signal averaging or OLS regressions. However, Panel B of Table 6 documents that the statistical significance of the differences in information ratios is less compelling than in the market-neutral setting. None of the GBM signals has an information ratio significantly different from the OLS-6M-6F signal (5% level); only the GBM-6M-20F signal is significantly different at the 10% level. In this statistical sense, the advantages of GBM-based signals are considerably reduced in the long-only setting relative to the market-neutral setting.

3.2.3. Long-only 1% tracking error

Finally, we study a long-only strategy against the MSCI World ex Financials, but with a 1% tracking error. Portfolios are rebalanced monthly using mean-variance optimization. Again, we keep portfolio betas as well as sector, industry, region, country and currency exposures close to the benchmark. A monthly one-way turnover constraint of 5% is enforced when rebalancing the portfolio.

Panel A of Table 7 gives the results when implementing the long-only 1% TE strategy using the eight signals. Compared to the market-neutral and the long-only 3% TE strategies, the signal averaging methods (1/N and Concepts) again have the lowest gross active, net active and net information ratios. In a similar vein,

GBM-6M-20F has the highest net information ratio of 1.40. But the difference compared to OLS-6M-6F further decreases, with a net information ratio of 1.08 (Figure 6). All in all, more rigid portfolio constraints (long-only constraint, limited active risk, low turnover) further reduce the performance difference between the best performing linear and non-linear signal. Panel B of Table 7 confirms that the OLS-6M-6F signal is also competitive in a statistical sense, evidencing that none of the four GBM-based signals has a significantly higher information ratio, even at the 10% level.

3.3. Robustness checking: performance through time

Many well-known equity factors and linear models combining these factors have struggled over the last decade. To study whether machine learning-based signals would have performed better over that time period, we investigate simulated active portfolio performance (for each of the three strategies and eight signals) in the first and second half of the sample period, i.e. from January 2001 to December 2009 and from January 2010 to December 2018. For the market-neutral strategy, all eight signals generated positive and statistically significant returns in the first period, see Figure 7. Over the second period, all eight signals declined in efficacy. Yet it is only the four benchmark signals that fail to exhibit net returns significantly different from zero. The four GBM-based signals still display t-statistics between 2.33 to 3.29. However, the two long-only strategies perform less convincingly in the second period. While the GBM signals do hold up better than the benchmark signals, most signals are no longer significant at the 5% level. For the 3% tracking error strategy, only the GBM-1M-6F signal is significant, with a t-statistic of 2.15. Reducing active risk to the 1% level, the GBM-6M-20F signal emerges as the only signal with a t-statistic over 2.

In the previous sections 3.2.1 to 3.2.3, we tested whether the GBM-based signals are significantly different from the benchmark signals in terms of the net information ratios. Moving from a high active risk long-short strategy to a highly constrained long-only 1% tracking error strategy inhibits the GBM-based signals' ability to significantly outperform other models. Restricting this analysis to the second period, from January 2010 to December 2018, we still see an increase in p-values of the paired test for equality of information ratios. For example, the p-values for the comparison between the OLS-6M-6F signal and the GBM-6M-20F signal increase from 8.0% for the market-neutral strategy to 23.4% for the long-only 3% tracking error strategy; this figure

stands at 46.1% for the long-only 1% tracking error strategy. Yet the level of p-values is, of course, much higher. And we therefore no longer observe significantly different information ratios between these two signals.

4. Conclusion

A growing body of literature suggests that machine learning can be used to derive better return predictions from well-known equity factors than traditional linear factor models. On the other hand, academics and practitioners are becoming increasingly aware of the associated pitfalls which impede a successful real-world implementation of many signals that look promising on paper. While many equity factors and linear factor models have already been probed for these pitfalls, our paper bridges this gap with respect to the application of machine learning for predicting the cross-section of stock returns.

Avramov, Cheng, Metzker (2019) were the first to conduct a detailed academic study in this regard, pinpointing the potential pitfalls of machine learning-based signals for equity investors. Our study further expands the understanding of machine learning-based return predictions by performing real-world portfolio simulations that account for realistic portfolio constraints and the limited risk-taking abilities typical for equity investors. We also account for frictions, such as transaction costs, shorting costs and a non-instantaneous implementation of trades, in our analysis.

Building on an investable global equity universe, we implement market-neutral, long-only 3% tracking error and long-only 1% tracking error simulations to compare machine learning-based alphas against simpler alternatives. Our analyses confirm that machine learning models are superior to traditional linear models in predicting cross-sectional stock returns one-month ahead using a set of well-documented equity factors. However, the extent to which the statistical advantage of machine learning predictions can be translated into economic gains in portfolio simulations critically depends on the ability to take risk and to implement trades efficiently. Practitioners should be aware of their respective capabilities when considering the adoption of machine learning-based predictions in their investment process.

References

Arnott, R., C.R. Harvey, and H. Markowitz (2019) "A Backtesting Protocol in the Era of Machine Learning". *Journal of Financial Data Science* 1(1): 64–74.

Avramov, D., S. Cheng, and L. Metzker (2019) "Machine Learning versus Economic Restrictions: Evidence from Stock Return Predictability". Working Paper.

Avramov, D., T. Chordia, and A. Goyal (2006) "Liquidity and Autocorrelations in Individual Stock Returns". *Journal of Finance* 61 (5): 2365–2394.

Balakrishnan, K., E. Bartov, and L. Faurel (2010) "Post Loss/Profit Announcement Drift". *Journal of Accounting and Economics* 50 (1): 20–41.

Banz, R.W. (1981) "The Relationship between Return and Market Value of Common Stocks". *Journal of Financial Economics* 9 (1): 3–18.

Basu, S. (1977) "Investment Performance of Common Stocks in Relation to Their Price–Earnings Ratios: A Test of the Efficient Market Hypothesis". *Journal of Finance* 32 (3): 663–682.

Bergstra, J., and Y. Bengio (2012) "Random Search for Hyper-Parameter Optimization". *Journal of Machine Learning Research* 13 (Feb): 281–305.

Bhandari, L.C. (1988) "Debt/Equity Ratio and Expected Common Stock Returns: Empirical Evidence". *Journal of Finance* 43 (2): 507–528.

Blume, M.E. (1980) "Stock Returns and Dividend Yields: Some More Evidence". *Review of Economics and Statistics* 62 (4): 567–577.

Brynjolfsson, E., and A. McAfee (2016) "The Second Machine Age". W. W. Norton & Company.

Campbell, J.Y., and R.J. Shiller (1988) "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors". *Review of Financial Studies* 1 (3): 195–228.

Chan, L.K., Y. Hamao, and J. Lakonishok (1991) "Fundamentals and Stock Returns in Japan". *Journal of Finance 46* (5): 1739–1764.

Chandrashekar, S., and R.K. Rao (2009) "The Productivity of Corporate Cash Holdings and the Cross-Section of Expected Stock Returns". McCombs Research Paper Series No. FIN-03-09.

Chen, L., M. Pelger, and J. Zhu (2019) "Deep Learning in Asset Pricing". Working Paper.

Cohen, R.B., P.A. Gompers, and T. Vuolteenaho (2002) "Who Underreacts to Cash-Flow News? Evidence from Trading between Individuals and Institutions". *Journal of Financial Economics* 66 (2–3): 409–462.

Cooper, M.J., H. Gulen, and M.J. Schill (2008) "Asset Growth and the Cross-Section of Stock Returns". *Journal of Finance* 63 (4): 1609–1651.

Coqueret, G., and T. Guida (2018) "Stock Returns and the Cross-section of Characteristics: A Tree-based Approach". Working Paper.

Da, Z., and M.C. Warachka (2009) "Cashflow Risk, Systematic Earnings Revisions, and the Cross-Section of Stock Returns". *Journal of Financial Economics* 94 (3): 448–468.

Daniel, K., and S. Titman (2006) "Market Reactions to Tangible and Intangible Information". *Journal of Finance* 61 (4): 1605–1643.

De Bondt, W.F., and R. Thaler (1985) "Does the Stock Market Overreact?". *Journal of Finance* 40 (3): 793–805.

DeMiguel, V., A. Martin-Utrera, F.J. Nogales, and R. Uppal (2020) "A Transaction-Cost Perspective on the Multitude of Firm Characteristics". *Review of Financial Studies*, forthcoming.

Dichtl, H., W. Drobetz, H. Lohre, C. Rother, and P. Vosskamp (2019) "Optimal Timing and Tilting of Equity Factors". *Financial Analysts Journal* 75 (4): 84–102.

Efron, B., and T. Hastie (2017) "Computer Age Statistical Inference". Cambridge University Press.

Fairfield, P.M., J.S. Whisenant, and T.L. Yohn (2003) "Accrued Earnings and Growth: Implications for Future Profitability and Market Mispricing". *Accounting Review* 78 (1): 353–371.

Fama, E.F., and K.R. French (1988) "Dividend Yields and Expected Stock Returns". *Journal of Financial Economics* 22 (1): 3–25.

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

Fama, E.F., and K.R. French (2006) "Profitability, Investment and Average Returns". *Journal of Financial Economics* 82 (3): 491–518.

Fama, E.F., and K.R. French (2015) "A Five-Factor Asset Pricing Model". *Journal of Financial Economics* 116 (1): 1–22.

Fama, E.F., and K.R. French (2016) "Dissecting Anomalies with a Five-Factor Model". *Review of Financial Studies* 29 (1): 69–103.

Feng, G., N.G. Polson, and J. Xu (2019) "Deep Learning in Characteristics-Sorted Factor Models". Working Paper.

Freyberger, J., A. Neuhierl, and M. Weber (2020) "Dissecting Characteristics Nonparametrically". *Review of Financial Studies*, forthcoming.

Friedman, J. (2002) "Stochastic Gradient Boosting". *Computational Statistics & Data Analysis* 38 (4): 367–378.

Friedman, J., and B. Popescu (2008) "Predictive Learning via Rule Ensembles". *Annals of Applied Statistics* 2 (3): 916–954.

Greenwell, B.M., B.C. Boehmke, and A.J. McCarthy (2018) "A Simple and Effective Model-Based Variable Importance Measure". Working Paper.

Gu, S., B. Kelly, and D. Xiu (2020a) "Empirical Asset Pricing via Machine Learning". *Review of Financial Studies*, forthcoming.

Gu, S., B. Kelly, and D. Xiu (2020b) "Autoencoder Asset Pricing Models". *Journal of Econometrics*, forthcoming.

Hastie, T., R. Tibshirani, and J. Friedman (2017) "The Elements of Statistical Learning, Data Mining, Inference, and Prediction (Second Edition)". Springer.

Haugen, R.A., and N.L. Baker (1996) "Commonality in the Determinants of Expected Stock Returns". *Journal of Financial Economics* 41 (3): 401–439.

Hou, K., G.A. Karolyi, and B.-C. Kho (2011) "What Factors Drive Global Stock Returns?". *Review of Financial Studies* 24 (8): 2527–2574.

Hou, K., C. Xue, and L. Zhang (2020) "Replicating Anomalies". *Review of Financial Studies*, forthcoming.

Huerta, R., F. Corbacho, and C. Elkan (2013) "Nonlinear support vector machines can systematically identify stocks with high and low future returns". *Algorithmic Finance* 2: 45-58.

Hutchinson, J.M., A.W. Lo, and T. Poggio (1994) "A Nonparametric Approach to Pricing and Hedging Derivative Securities via Learning Networks". *Journal of Finance* 49 (3): 851–889.

Jaffe, J., D.B. Keim, and R. Westerfield (1989) "Earnings Yields, Market Values, and Stock Returns". *Journal of Finance* 44 (1): 135–148.

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

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

Lehmann, B.N. (1990) "Fads, Martingales, and Market Efficiency". *Quarterly Journal of Economics* 105 (1): 1–28.

Levin, A. U. (1995) "Stock Selection via Nonlinear Multi-Factor Models". *NIPS'95: Proceedings of the 8th International Conference on Neural Information Processing Systems:* 966–972.

Litzenberger, R.H., and K. Ramaswamy (1979) "The Effect of Personal Taxes and Dividends on Capital Asset Prices: Theory and Empirical Evidence". *Journal of Financial Economics* 7 (2): 163–195.

Lopez De Prado, M. (2018) "Advances in Financial Machine Learning". Wiley Publishing.

Loughran, T., and J.R. Ritter (1995) "The New Issues Puzzle". *Journal of Finance* 50 (1): 23–51.

Luo, Y., J. Jussa, and S. Wang (2017) "The Big and the Small Sides of Big Data". QES Handbook of Active Investing, Part 1, Wolfe Research.

Messmer, M. (2017) "Deep Learning and the Cross-Section of Expected Returns". Working Paper.

Molnar, C. (2019) "Interpretable Machine Learning – A Guide for Making Black Box Models Explainable". https://christophm.github.io/interpretable-ml-book/.

Moritz, B., and T. Zimmermann (2016) "Tree-Based Conditional Portfolio Sorts: The Relation between Past and Future Stock Returns". Working Paper.

Novy-Marx, R. (2013) "The Other Side of Value: The Gross Profitability Premium". *Journal of Financial Economics* 108 (1): 1–28.

Ou, J.A., and S.H. Penman (1989) "Financial Statement Analysis and the Prediction of Stock Returns". *Journal of Accounting and Economics* 11 (4): 295–329.

Pontiff, J., and A. Woodgate (2008) "Share Issuance and Cross-Sectional Returns". *Journal of Finance* 63 (2): 921–945.

Rasekhschaffe, K.C., and R.C. Jones (2019) "Machine Learning for Stock Selection". *Financial Analysts Journal* 75 (3): 70-88.

Richardson, S.A., R.G. Sloan, M.T. Soliman, and I. Tuna (2005) "Accrual Reliability, Earnings Persistence and Stock Prices". *Journal of Accounting and Economics* 39 (3): 437–485.

Ritter, J.R. (1991) "The Long-Run Performance of Initial Public Offerings". *Journal of Finance* 46 (1): 3–27.

Rosenberg, B., K. Reid, and R. Lanstein (1985) "Persuasive Evidence of Market Inefficiency". *Journal of Portfolio Management* 11 (3): 9–16.

Sloan, R. (1996) "Do Stock Prices Fully Reflect Information in Accruals and Cash Flows about Future Earnings?". *Accounting Review* 71 (3): 289–315.

Soliman, M.T. (2008) "The Use of DuPont Analysis by Market Participants". *Accounting Review* 83 (3): 823–853.

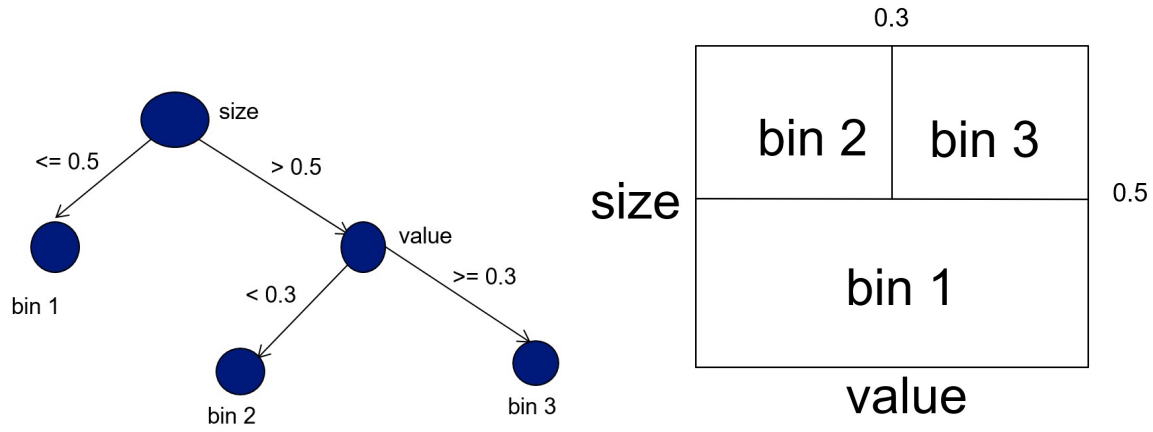
Suhonen, A., M. Lennkh, and F. Perez (2017) "Quantifying Backtest Overfitting in Alternative Beta Strategies". *Journal of Portfolio Management* 43 (2): 90–104.

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

Wright, J., S.C. Yam, and S.P. Yung (2014) "A Test for the Equality of Multiple Sharpe Ratios". *Journal of Risk* 16 (4): 3–21.

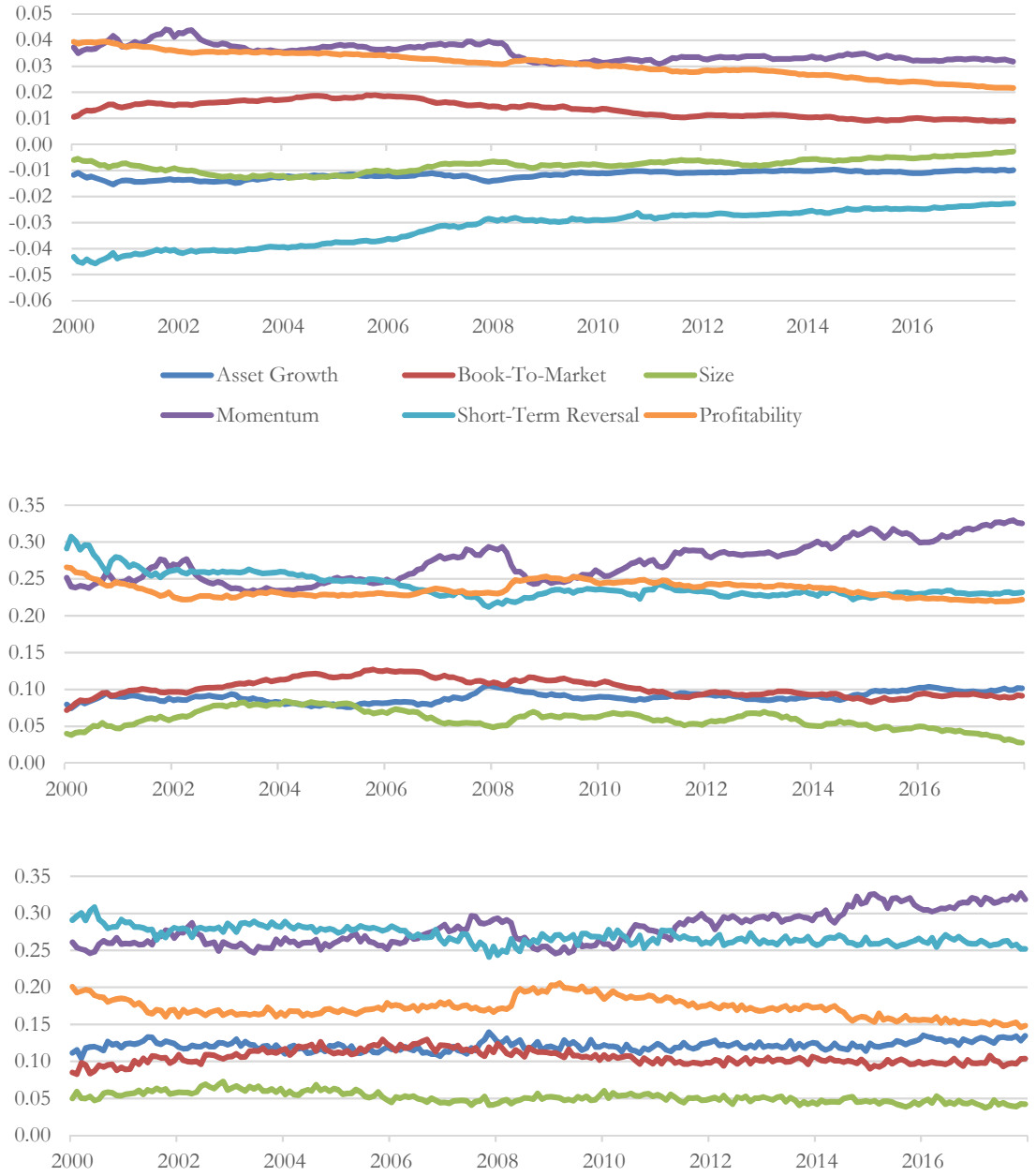
Zhu, Z., and K. Yung (2016) "The Interaction of Short-Term Reversal and Momentum Strategies". *Journal of Portfolio Management* 42 (4): 96–107.

Figure 1: Regression Tree Example



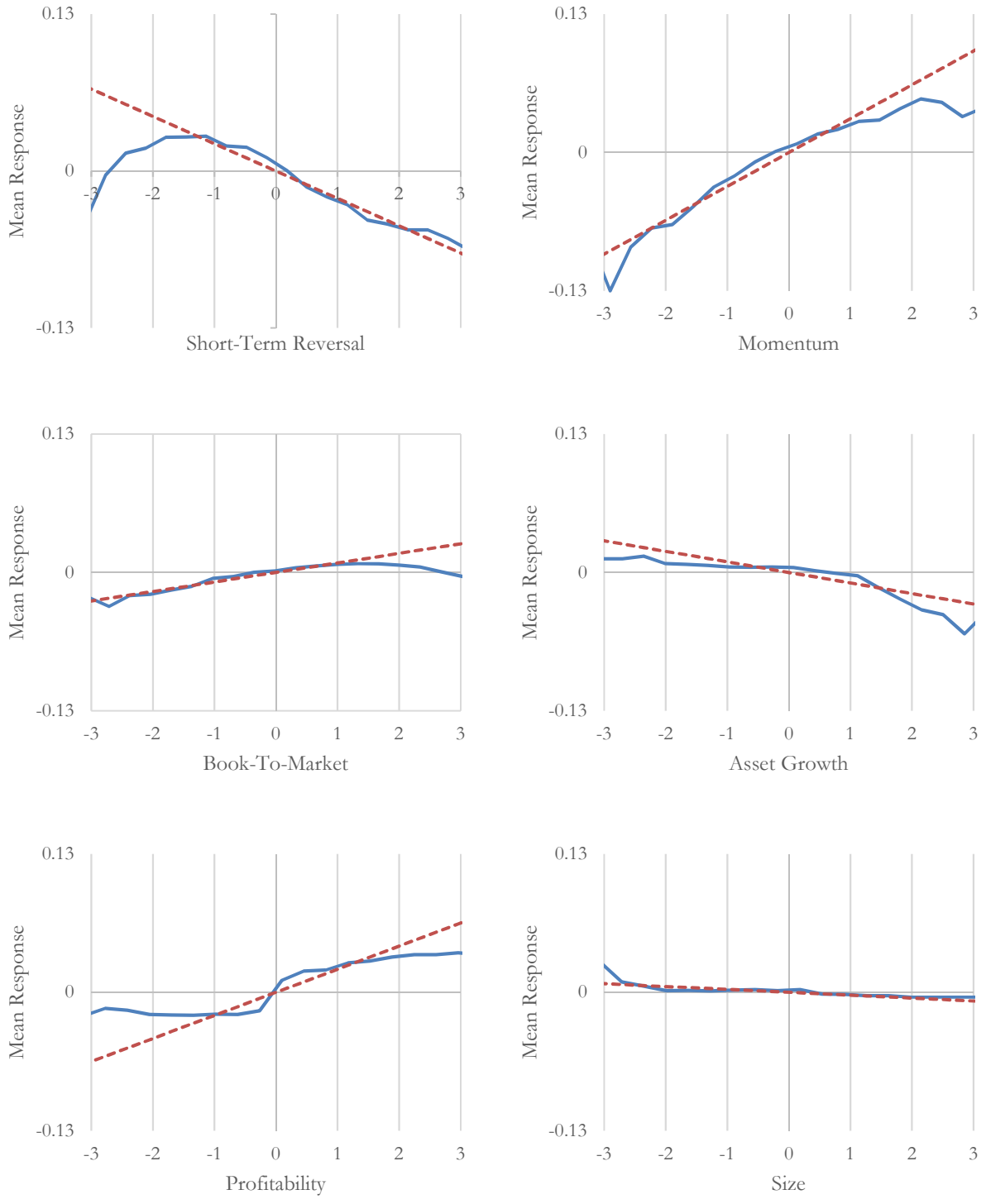
The left chart is a representation of a decision tree. The right chart illustrates how a decision tree effectively splits the space of predictor values into disjoint subsets, which are called leaves. Every leaf represents a value of the outcome variable, where the predictor variables take values consistent with the decision path.

Figure 2: OLS vs GBM Over Time: Betas and Variable Importance



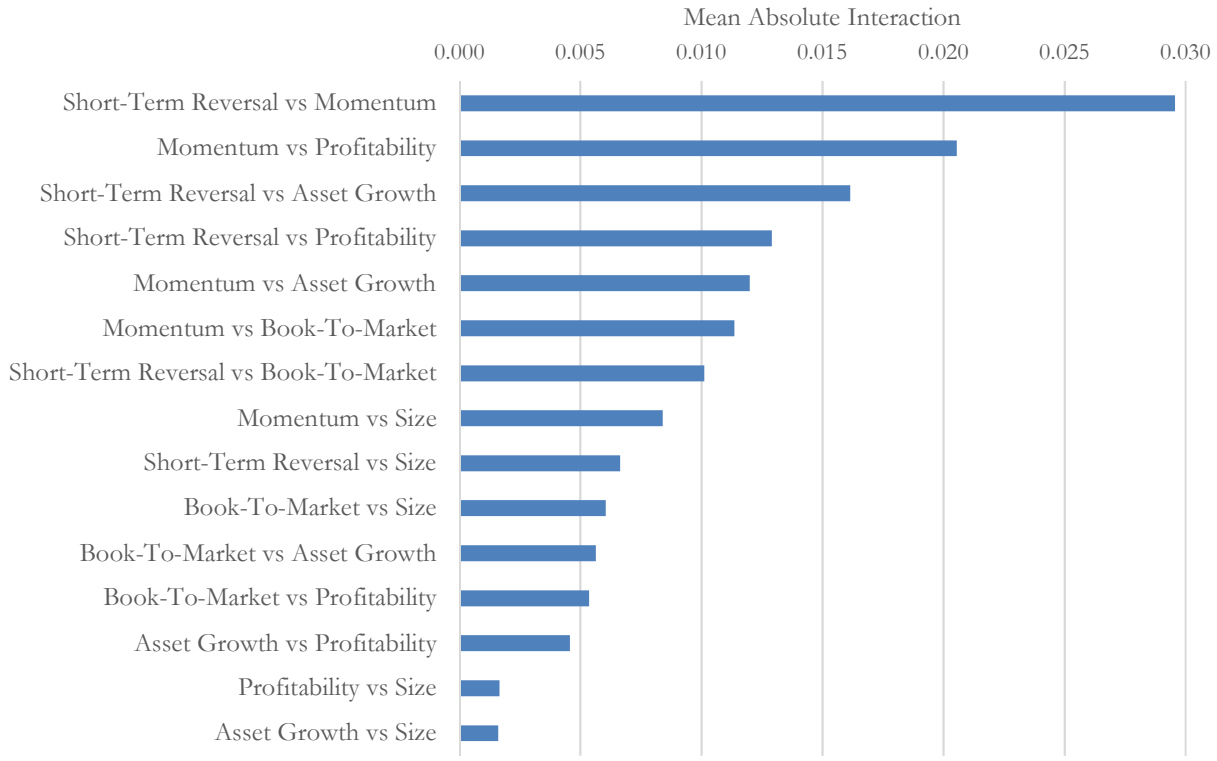
The upper chart shows beta coefficients from expanding window OLS regressions over time. The middle chart plots the percentage of absolute beta, which is the absolute of these beta coefficients over the sum of all six absolute betas. The lower chart depicts the percentage variable contribution from expanding window GBM models. The first model plotted at December 2000 is estimated based on monthly factor scores from January 1991 to November 2000 and monthly stock returns from February 1991 to December 2000. For the second model plotted at January 2001, the date range is increased by one month; for the third model by two months, and so forth.

Figure 3: Partial Dependence Plots for GBM and OLS Six-Factor Models



Partial dependence plots for the six-factor GBM model (solid blue lines) and partial dependence plots for the six-factor OLS model (red dotted lines). Both models are fitted on monthly factor scores from January 1991 to November 2018 and monthly stock returns from February 1991 to December 2018.

Figure 4: Pairwise Mean Absolute Interactions

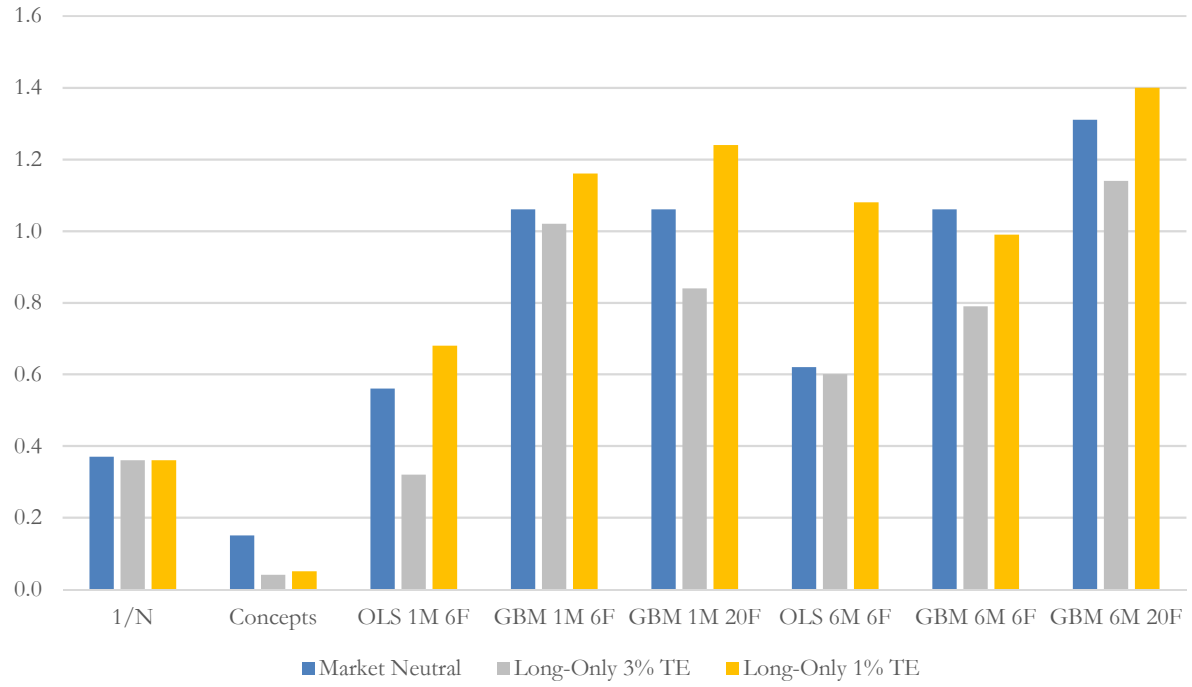


Six-factor GBM model predicting one-month standardized forward returns fitted on monthly factor scores from January 1991 to November 2018 and monthly stock returns from February 1991 to December 2018. We calculate the difference between the two-dimensional partial dependence function and the sum of the two one-dimensional partial dependence functions on a product rule grid consisting of 121 points (in each dimension we use 11 different score values to create the grid). The size of the bar corresponds to the absolute value averaged over the 121 grid points.

Figure 5: Interaction Effect of Momentum and Short-Term Reversal

Six-factor GBM model predicting one-month standardized forward returns fitted on monthly factor scores from January 1991 to November 2018 and monthly stock returns from February 1991 to December 2018. The color of the cell corresponds to the size of the interaction.

Figure 6: Net Information Ratios for Market-Neutral, Long-Only 3% TE and Long-Only 1% TE Strategies



Performance for simulated portfolios rebalanced monthly between December 2000 and November 2018, i.e. strategy returns are from January 2001 to December 2018. Turnover is constrained at 27.5% per month for market-neutral strategy, at 7% for the long-only 3% tracking error strategy, and at 5% for long-only 1% tracking error strategy. Factor "Short-Term Reversal" is lagged by five days. The benchmark for the market-neutral strategy is the three-month USD-LIBOR rate. The benchmark for the two long-only strategies is the MSCI World index excluding financial stocks.

Figure 7: Robustness Check: Sub-Period Performance

Performance for simulated portfolios rebalanced monthly between December 2000 and November 2018, i.e. strategy returns are from January 2001 to December 2018. Turnover is constrained at 27.5% per month for market-neutral strategy, at 7% for the long-only 3% tracking error strategy, and at 5% for long-only 1% tracking error strategy. Factor "Short-Term Reversal" is lagged by five days. The benchmark for the market-neutral strategy is the three-month USD-LIBOR rate. The benchmark for the two long-only strategies is the MSCI World index excluding financial stocks.

Table 1: Overview of Factors used as Predictors (or Features)

Description	Concept	Definition with References
Short-Term Reversal	Momentum	Long stocks with a weak previous month's performance and short stocks with a high previous month's performance (Jegadeesh, 1990; and Lehmann, 1990)
Momentum	Momentum	Twelve-month price momentum that controls for the short-term reversal effect by excluding the most recent month (t-1) at time t (Jegadeesh, 1990; Jegadeesh and Titman, 1993)
Long-Term Reversal	Momentum	Following DeMiguel et al. (forthcoming), we chose the horizon to be 36 months. To control for the momentum effect, we excluded the most recent year from our three-year horizon of past performance (De Bondt and Thaler, 1985)
Accruals	Quality	Change in working capital divided by book value (Sloan, 1996)
Asset Growth (Investment)	Quality	Year-on-year change in total assets divided by the total assets in t-2 (Fairfield, Whisenant, and Yohn, 2003; Richardson, Sloan, Soliman, and Tuna, 2005; Titman, Wei, and Xie, 2004; Fama and French, 2006; Cooper, Gulen, and Schill, 2008)
Asset Turnover	Quality	Sales divided by average net operating assets (Soliman, 2008)
Cash Productivity	Quality	Market value of equity plus long-term debt minus total assets divided by cash (Chandrashekar and Rao, 2009)
Change in Long Term Debt	Quality	Year-on-year changes divided by the long-term debt in t-2 (Richardson et al., 2005)
Change in Shares Outstanding	Quality	Year-on-year change in shares outstanding divided by outstanding shares in t-2 (Ritter, 1991; Loughran and Ritter, 1995. See also Daniel and Titman (2006) and Pontiff and Woodgate (2008))
Leverage	Quality	Total liabilities divided by the market value of the company (Bhandari, 1988)
Profit Margin	Quality	Operating income divided by sales (Soliman, 2008)
Return on Assets	Quality	Net income divided by the two fiscal period average of total assets (Balakrishnan, Bartov and Faurel, 2010)
Sales to Cash	Quality	Sales divided by cash (Ou and Penman, 1989)
Sales to Inventory	Quality	Sales divided by inventory (Ou and Penman, 1989)
Size	Size	Market capitalization (Banz, 1981; Fama and French, 1992)
Book to Market	Value	Book value of equity divided by market capitalization (Basu, 1977; Rosenberg, Reid and Lanstein, 1985; Jaffe, Keim and Westerfield, 1989; Chan, Hamao, and Lakonishok, 1991; Fama and French, 1992)
Cash Flow Yield	Value	Cash flows are measured as the sum of funds from operations, extraordinary items and funds from other operating activities divided by market capitalization (Sloan, 1996; Da and Warachka, 2009; Hou, Karolyi, and Kho, 2011)
Dividend Yield	Value	Dividends divided by market capitalization (Litzenberger and Ramaswamy, 1979; Blume, 1980; Fama and French, 1988; Campbell and Shiller, 1988)
Earnings Yield	Value	Earnings divided by market capitalization (Basu, 1977)
Profitability	Value	Annual revenues less cost of goods sold, interest and other expenses divided by book value (Haugen and Baker, 1996; Cohen, Gompers and Vuolteenaho, 2002; Novy-Marx, 2013; Fama and French 2006 and 2016)

Table 2: Correlation Matrix of Signals

	Concepts	1/N	GBM 6M 6F	GBM 1M 6F	OLS 6M 6F	OLS 1M 6F	GBM 1M 20F	GBS 6M 20F
Concepts	1.00							
1/N	0.92	1.00						
GBM 6M 6F	0.43	0.39	1.00					
GBM 1M 6F	0.46	0.35	0.73	1.00				
OLS 6M 6F	0.56	0.48	0.66	0.53	1.00			
OLS 1M 6F	0.65	0.47	0.57	0.65	0.84	1.00		
GBM 1M 20F	0.57	0.51	0.55	0.68	0.46	0.56	1.00	
GBM 6M 20F	0.58	0.59	0.68	0.57	0.59	0.52	0.79	1.00

The signals are the out-of-sample predictions of the eight hyperparameter tuned models calculated monthly between December 2000 and November 2018.

Table 3: Top and Bottom-Decile Average Monthly Returns

Model	Signal Lag	D1	D10	D10 - D1	D10 - D1	D10 - D1	D10 - D1
		Avg. Ret.	*Avg. Ret.*	*Avg. Ret.*	*Std. Dev.*	*Ann. IR*	*t-Stat*
1/N	*0M*	-0.84	0.23	1.07	1.86	1.99	8.47
	1M	-0.81	0.14	0.95	1.82	1.81	7.71
	2M	-0.72	0.13	0.85	1.84	1.60	6.82
Concept	*0M*	-0.86	0.20	1.07	1.90	1.94	8.26
	1M	-0.73	0.14	0.88	1.89	1.61	6.84
	2M	-0.62	0.10	0.72	1.89	1.33	5.64
OLS-1M-6F	*0M*	-0.61	0.10	0.71	2.23	1.10	4.67
	1M	-0.49	-0.05	0.44	1.60	0.96	4.07
	2M	-0.35	-0.06	0.29	1.75	0.57	2.43
GBM-1M-6F	*0M*	-0.68	0.26	0.94	1.82	1.79	7.60
	1M	-0.55	0.12	0.67	1.44	1.60	6.81
	2M	-0.50	0.03	0.53	1.44	1.27	5.40
GBM-1M-20F	*0M*	-0.94	0.40	1.34	1.86	2.50	10.63
	1M	-0.80	0.21	1.02	1.72	2.05	8.70
	2M	-0.66	0.12	0.78	1.63	1.65	7.02
OLS-6M-6F	*0M*	-0.44	0.10	0.54	1.94	0.97	4.14
	1M	-0.44	0.08	0.52	1.85	0.98	4.19
	2M	-0.37	0.05	0.42	1.71	0.86	3.67
GBM-6M-6F	*0M*	-0.73	0.16	0.89	2.34	1.32	5.61
	1M	-0.78	0.18	0.96	2.22	1.51	6.41
	2M	-0.72	0.11	0.83	2.14	1.34	5.69
GBM-6M-20F	*0M*	-0.97	0.28	1.25	2.28	1.90	8.09
	1M	-0.97	0.27	1.24	2.20	1.95	8.30
	2M	-0.88	0.22	1.10	2.16	1.77	7.52

Monthly average returns of equal-weighted decile portfolios sorted according to the models' prediction. Returns are relative to the median stock return in each month, i.e. before calculating the equal-weighted decile portfolio returns for a given month, we subtract the median monthly return across all stocks from every stock return. The simulated period is January 2001 to December 2018. D1 refers to the decile portfolios with the lowest predictions. D10 refers to the decile portfolios with the highest predictions. D10-D1 is the decile spread portfolio and t(D10-D1) is the related t-statistic testing for a null hypothesis of zero return.

Table 4: Time Average Percentage Variable Importance

Factor	GBM 1M 20F	GBM 1M 20F MOM1L05DL	GBM 6M 20F	GBM 6M 20F MOM1L05DL
Short-Term Reversal	13.8%	10.0%	3.5%	3.8%
Momentum	9.8%	10.3%	8.2%	7.4%
Long-Term Reversal	8.2%	8.3%	5.4%	5.4%
Accruals	4.2%	4.3%	4.1%	4.4%
Asset Growth	3.9%	4.1%	4.3%	4.2%
Asset Turnover	9.2%	9.4%	7.4%	8.4%
Cash Productivity	2.0%	2.1%	2.1%	2.1%
Change in Long Term Debt	3.0%	3.3%	3.1%	3.1%
Change in Shares Outstanding	2.9%	3.1%	3.5%	3.4%
Leverage	3.8%	3.9%	4.2%	4.4%
Profit Margin	3.4%	3.5%	3.5%	3.6%
Return on Assets	3.6%	3.7%	3.5%	3.4%
Sales to Cash	1.8%	1.9%	1.7%	1.8%
Sales to Inventory	6.0%	6.2%	5.8%	6.0%
Size (Market Cap)	4.3%	4.4%	4.8%	4.9%
Book to Market	3.0%	3.2%	5.4%	4.8%
Cash Flow Yield	4.4%	4.6%	6.4%	5.9%
Dividend Yield	3.0%	3.1%	4.3%	4.0%
Earnings Yield	3.8%	4.1%	4.9%	4.7%
Profitability	6.1%	6.5%	13.8%	14.1%

Time average of monthly percentage variable importance for trained and tuned GBM models between December 2000 and November 2018. The term MOM1L05DL indicates that these models have been fitted to the data where the original Short-Term Reversal signal has been replaced by a 5-day lagged version.

Table 5: Market-Neutral Portfolio Simulations

Method	1/N	Concept	OLS	GBM	GBM	OLS	GBM	GBM
Prediction Horizon	-	-	1M	1M	1M	6M	6M	6M
No. of Factors	20F	19F	6F	6F	20F	6F	6F	20F
Panel A: Performance Analysis								
Gross performance								
Return (%)	6.3	5.97	8.22	12.16	12.59	7.13	11.18	11.92
Active Ret. (%)	4.46	4.13	6.38	10.32	10.75	5.29	9.34	10.08
IR	0.81	0.78	1.47	2.32	2.15	1.1	1.89	1.91
Ann. Turnover (%, one-sided)	130	187	224	321	309	119	226	167
Ann. Trans. Costs (bps)	233	319	374	519	502	216	377	289
Net performance								
Return (%)	3.86	2.66	4.27	6.53	7.12	4.85	7.09	8.75
Active Ret. (%)	2.02	0.82	2.43	4.69	5.28	3.01	5.25	6.91
TE (%)	5.53	5.33	4.35	4.44	4.99	4.82	4.95	5.29
IR	0.37	0.15	0.56	1.06	1.06	0.62	1.06	1.31
Drawdown (%)	-24.99	-28.56	-17.74	-9.05	-6.09	-18.04	-12.44	-7.96
Panel B: Paired Test for Equality of Information Ratio p-Values								
1/N	-	8.4%	53.8%	4.4%	1.2%	43.6%	3.3%	0.0%
Concepts	8.4%	-	16.6%	0.6%	0.1%	15.8%	0.6%	0.0%
OLS 1M 6F	53.8%	16.6%	-	2.4%	5.1%	62.7%	1.3%	0.2%
GBM 1M 6F	4.4%	0.6%	2.4%	-	99.2%	8.5%	97.3%	32.2%
GBM 1M 20F	1.2%	0.1%	5.1%	99.2%	-	12.3%	98.8%	17.1%
OLS 6M 6F	43.6%	15.8%	62.7%	8.5%	12.3%	-	3.4%	0.8%
GBM 6M 6F	3.3%	0.6%	1.3%	97.3%	98.8%	3.4%	-	27.0%
GBM 6M 20F	0.0%	0.0%	0.2%	32.2%	17.1%	0.8%	27.0%	-

Panel A gives performance for simulated portfolios rebalanced monthly between December 2000 and November 2018, i.e. strategy returns from January 2001 to December 2018. Turnover is constrained at 27.5% per month. Factor "Short-Term Reversal" is lagged by five days. Benchmark for the market-neutral strategy is the three-month USD-LIBOR rate. Panel B contains p-values for a two-sided test of H0: IR(Signal A)=IR(Signal B). The calculation uses strategy net active returns from January 2001 to December 2018. We use the Chi-Squared test from Wright et. al. (2014) based on a heteroskedasticity and autocorrelation consistent (HAC) covariance matrix estimated using the R *package "Sharpe*R*".*

Table 6: Long-Only 3% Tracking Error Portfolio Simulations

Method	1/N	Concept	OLS	GBM	GBM	OLS	GBM	GBM
Prediction Horizon	-	-	1M	1M	1M	6M	6M	6M
No. of Factors	20F	19F	6F	6F	20F	6F	6F	20F
Panel A: Performance Analysis								
Gross performance								
Return (%)	7.55	6.96	7.8	9.33	8.91	8.22	9.09	9.64
Active Ret. (%)	1.88	1.29	2.13	3.66	3.24	2.55	3.42	3.97
IR	0.67	0.44	0.8	1.61	1.43	0.86	1.28	1.6
Ann. Turnover (%, one-sided)	55	74	81	82	82	48	81	70
Ann. Trans. Costs (bps)	82	111	121	124	124	72	121	106
Net performance								
Return (%)	6.68	5.78	6.51	7.99	7.58	7.46	7.78	8.5
Active Ret. (%)	1.01	0.11	0.84	2.32	1.91	1.79	2.11	2.83
TE (%)	2.81	2.9	2.67	2.28	2.26	2.96	2.67	2.48
IR	0.36	0.04	0.32	1.02	0.84	0.6	0.79	1.14
Drawdown (%)	-11.15	-27.71	-20.84	-3.86	-6.45	-11.48	-8.9	-5.98
Panel B: Paired Test for Equality of Information Ratio p-Values								
1/N	-	8.9%	83.6%	3.4%	14.6%	49.5%	21.1%	1.4%
Concepts	8.9%	-	40.6%	0.2%	0.8%	8.5%	1.9%	0.1%
OLS 1M 6F	83.6%	40.6%	-	0.4%	5.9%	10.2%	5.8%	0.5%
GBM 1M 6F	3.4%	0.2%	0.4%	-	48.1%	9.9%	28.7%	65.3%
GBM 1M 20F	14.6%	0.8%	5.9%	48.1%	-	40.9%	81.8%	19.1%
OLS 6M 6F	49.5%	8.5%	10.2%	9.9%	40.9%	-	47.4%	6.5%
GBM 6M 6F	21.1%	1.9%	5.8%	28.7%	81.8%	47.4%	-	17.9%
GBM 6M 20F	1.4%	0.1%	0.5%	65.3%	19.1%	6.5%	17.9%	-

Panel A gives performance for simulated portfolios rebalanced monthly between December 2000 and November 2018, i.e. strategy returns from January 2001 to December 2018. Turnover is constrained at 7% per month. Factor "Short-Term Reversal" is lagged by five days. Benchmark is MSCI World excluding financial stocks. Panel B contains p-values for a two-sided test of H0: IR(Signal A)=IR(Signal B). The calculation uses strategy net active returns from January 2001 to December 2018. We use the Chi-Squared test from Wright et. al. (2014) based on a heteroskedasticity and autocorrelation consistent (HAC) covariance matrix estimated using the R package "SharpeR".

Table 7: Long-Only 1% Tracking Error Portfolio Simulations

Method	1/N	Concept	OLS	GBM	GBM	OLS	GBM	GBM
Prediction Horizon	-	-	1M	1M	1M	6M	6M	6M
No. of Factors	20F	19F	6F	6F	20F	6F	6F	20F
Panel A: Performance Analysis								
Gross performance								
Return (%)	6.57	6.43	7.22	7.79	7.9	7.41	7.53	7.8
Active Ret. (%)	0.9	0.76	1.55	2.12	2.23	1.74	1.86	2.13
IR	0.77	0.6	1.42	2.09	2.13	1.41	1.74	1.95
Ann. Turnover (%, one-sided)	30	44	51	58	58	25	50	37
Ann. Trans. Costs (bps)	44	66	76	88	87	38	75	56
Net performance								
Return (%)	6.1	5.73	6.41	6.85	6.97	7.01	6.73	7.2
Active Ret. (%)	0.43	0.06	0.74	1.18	1.3	1.34	1.06	1.53
TE (%)	1.17	1.26	1.1	1.02	1.05	1.24	1.07	1.09
IR	0.36	0.05	0.68	1.16	1.24	1.08	0.99	1.4
Drawdown (%)	-5.55	-10.67	-3.68	-1.2	-1.55	-3.25	-3.8	-2
Panel B: Paired Test for Equality of Information Ratio p-Values								
1/N	-	2.0%	31.1%	0.7%	0.4%	2.2%	3.3%	0.0%
Concepts	2.0%	-	2.8%	0.0%	0.0%	0.1%	0.2%	0.0%
OLS 1M 6F	31.1%	2.8%	-	1.7%	4.7%	0.3%	15.2%	0.9%
GBM 1M 6F	0.7%	0.0%	1.7%	-	77.0%	61.8%	33.6%	31.7%
GBM 1M 20F	0.4%	0.0%	4.7%	77.0%	-	54.7%	27.3%	40.8%
OLS 6M 6F	2.2%	0.1%	0.3%	61.8%	54.7%	-	75.5%	21.4%
GBM 6M 6F	3.3%	0.2%	15.2%	33.6%	27.3%	75.5%	-	5.7%
GBM 6M 20F	0.0%	0.0%	0.9%	31.7%	40.8%	21.4%	5.7%	-

Panel A gives performance for simulated portfolios rebalanced monthly between December 2000 and November 2018, i.e. strategy returns from January 2001 to December 2018. Turnover is constrained at 5% per month. Factor "Short-Term Reversal" is lagged by five days. Benchmark is MSCI World excluding financial stocks. Panel B contains p-values for a two-sided test of H0: IR(Signal A)=IR(Signal B). The calculation uses strategy net active returns from January 2001 to December 2018. We use the Chi-Squared test from Wright et. al. (2014) based on a heteroskedasticity and autocorrelation consistent (HAC) covariance matrix estimated using the R package "SharpeR".

Appendix: Tuning of Hyperparameters

Suppose we use monthly data, we want to tune our model yearly on dates $t_1, t_2, t_3, \ldots, t_n$, i.e. tuning distance $d = 12$, and we have a prediction horizon of six months $h = 6$. And suppose we have already generated a set of hyperparameter vectors $v_1, \ldots, v_k \in V$ which we want to evaluate. We need to estimate all k models for each of the n tuning dates.

Tuning at time t_1, we want to use the tuned model to make out-of-sample predictions for time $t_1 + h$. We use the convention that the y variable is forward-looking. Thus, we cannot use any observations from dates $t \in (t - h, t_1]$, because otherwise we would validate the models on information not available at time t_1.

Given yearly tuning, we want to evaluate predictive abilities of a model every twelve months – therefore the length of our validation set is 12. This implies that, when all k models have been estimated for all n tuning dates, for each model v_i, we have a vector of validation scores (e.g. MAE or R^2) $s_{i,1}, \ldots, s_{i,n}$ calculated on mutually exclusive validation sets $S_1, \ldots, S_n$ which together cover all dates in the interval $(t_1 - d - h, t_n - h]$.

Each validation score $s_{i,j}$ is based on training model v_i on data with $t \leq t_j - h - d - h$. Again, we must ensure that no information from the validation set is contained in the training set – therefore we must again keep a distance of h between training and validation sets. This explains the second h in the previous inequality.

The entire framework is depicted in Figure 8. After estimating all nk models, we use the corresponding vectors of validation scores $s_{\cdot,1}, \ldots, s_{\cdot,n}$ in the following way: At time t we make out-of-sample predictions based on hyperparameter vector v_i, which minimizes[23] average validation scores (MAE in the empirical section of this paper), see Figure 9.

$$avg\{s_{i,j} : t_j \leq t\}$$

[23] In case of R^2, we would naturally maximize the average validation score.

Figure 8: Hyperparameter Tuning – Random Grid Search

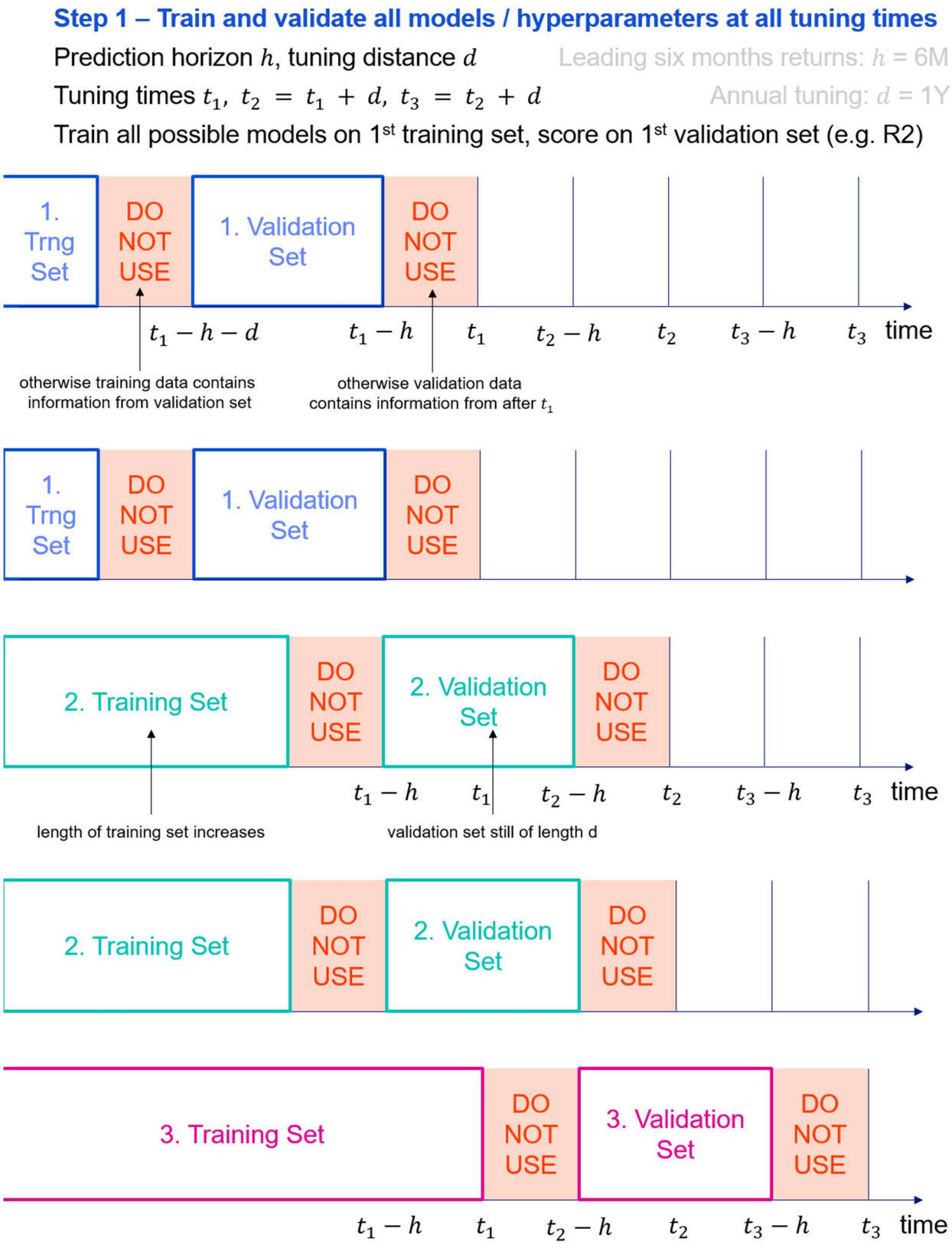
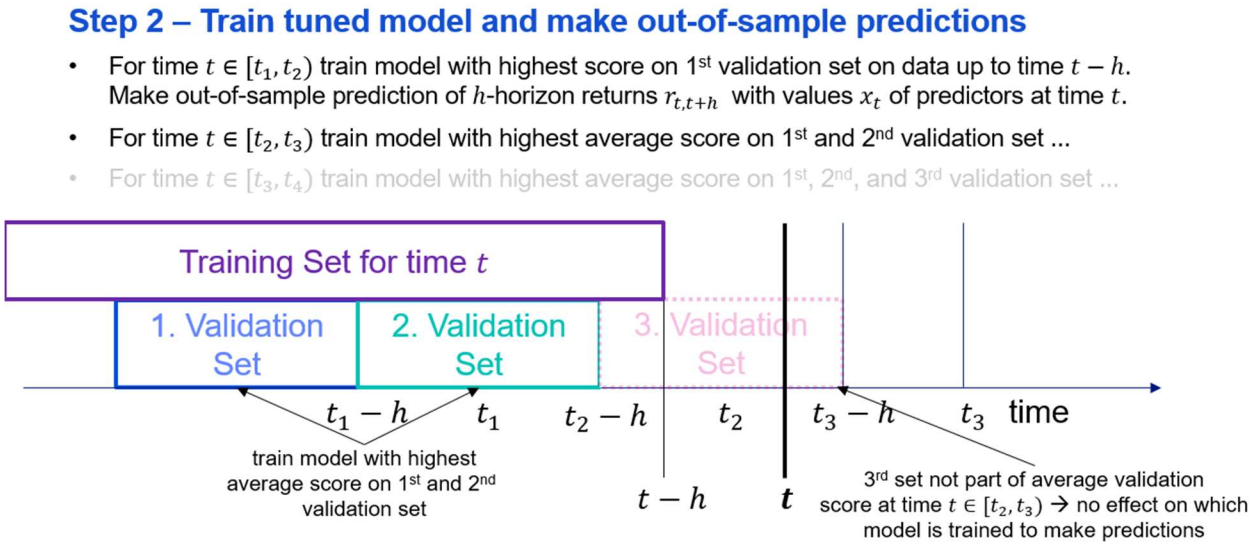


Figure 9: Hyperparameter Tuning – Application of Optimal Model Over Time



Note that since we tune models relatively frequently, this implies that the choice of the best hyperparameter depends only on a short period of time at the beginning. We tune models annually in the empirical part of this article, thus, for the first year of out-of-sample predictions, the choice of the hyperparameter would be based on only one year of data. After one year, it will be based on two years of validation data and so forth (at least in the case of expanding window validation). Therefore, we evaluate out-of-sample predictions of the model only after five years of validation data. In our case, this means that we start evaluating the out-of-sample predictions beginning in December 2000.