Real-time Machine Learning in the Cross-Section of Stock Returns *

Bin Li
Wuhan University

Alberto G. Rossi
Georgetown University

Xuemin (Sterling) Yan
Lehigh University

Lingling Zheng
Renmin University of China

October 20, 2023

Abstract

Recent studies document strong performance for machine-learning-based investment strategies. These strategies use anomaly variables discovered ex-post as predictors of stock returns and cannot be implemented in real time. We construct real-time machine learning strategies based on a "universe" of fundamental signals. While positive and significant, the out-of-sample performance of these strategies is significantly weaker than those documented by prior studies. We find qualitatively similar results when examining a "universe" of past-return-based signals. Our results offer a more tempered view of the economic gains associated with machine learning strategies relative to prior literature.

*Bin Li is at the Economics and Management School, Wuhan University, China. Email: binli.whu@whu.edu.cn. Alberto G. Rossi is at the McDonough School of Business, Georgetown University, U.S.A. Email: agr60@georgetown.edu. Xuemin (Sterling) Yan is at the College of Business, Lehigh University, U.S.A. Email: xuy219@lehigh.edu. Lingling Zheng is at the School of Business, Renmin University of China, China. Email: zhenglingling@rmbs.ruc.edu.cn. We thank Marcin Kacperczyk, David Solomon, Allan Timmermann, and seminar and conference participants at the Beihang University, Chinese Finance Annual Meeting, CNCC Intelligent Finance Workshop, Duke Kunshan University, Hunan University, Lehigh University, Nanjing University, Southwest Jiaotong University, Sun Yat-Sen University, Taiwan Finance Association Asset Pricing Conference, Tianjin University, Tongji University, and Xian Jiaotong University for helpful comments and discussions. Bin Li acknowledges financial support from the National Natural Science Foundation of China (No. 71971164, 72371191). Lingling Zheng acknowledges financial support from the National Natural Science Foundation of China (No. 72122021). All errors are our own.

1 Introduction

Machine learning methods have received considerable attention in the recent asset pricing literature, particularly in the area of return prediction (see, e.g., Chen, Pelger, and Zhu (2022), Freyberger, Neuhierl, and Weber (2020), Gu, Kelly, and Xiu (2020) and Leippold, Wang, and Zhou (2022)). The general conclusions of the existing studies are remarkably similar—machine learning models are superior to traditional models in predicting the cross-section of stock returns. Given the inherent focus of machine learning methods on *out-of-sample* prediction, these studies leave readers with the impression that machine learning methods routinely lead to large improvements in investment performance. Indeed, a common theme among many existing studies is constructing long-short investment strategies based on machine learning forecasts and demonstrating that these strategies are highly profitable.

While prior studies have clearly established the potential for large economic gains to investors using machine learning forecasts, an important issue that has yet to be fully addressed in the literature is the real-time implementability of machine learning strategies. Specifically, existing studies use published anomaly variables as predictors of stock returns and implicitly assume that they are known to investors at the beginning of the training period, even though most anomalies are discovered years or decades later.[1] While this approach is very natural if the objective is to measure risk premium or estimate the stochastic discount factor, in which case we can take an econometrician's perspective and analyze data ex-post, such an approach raises the issue of whether the resulting machine learning strategies could have been implemented by real-time investors. Take the asset growth anomaly of Cooper, Gulen, and Schill (2008) as an example. It is unlikely that investors would have been able to single out asset growth as a stock return predictor before that research was published in the mid-2000s. Assuming otherwise would lead to a look-ahead bias. Moreover, published anomalies tend to exhibit strong in-sample performance partly because of the publication bias (Harvey, Liu, and Zhu, 2016). Having hindsight of this strong performance can lead to an overly optimistic assessment of past trading decisions; hence, the economic gains from using machine learning forecasts documented by prior studies are potentially overstated for real-time investors.

In this paper, we examine machine learning strategies based on a "universe" of over 18,000 fundamental signals. Because these signals are constructed from financial statement variables using permu-

[1] There are exceptions. For example, Kozak, Nagel, and Santosh (2020) use shrinkage and selection method to construct a stochastic discount factor (SDF) from a comprehensive set of 70 financial ratios compiled by WRDS. Avramov, Kaplanski, and Subrahmanyam (2022) use machine learning methods to construct a fundamental index from 105 signals that are based on deviations of accounting variables from their recent averages.

tational arguments (Yan and Zheng, 2017), our strategies are implementable in real-time. Moreover, examining a universe of fundamental signals, rather than selecting a subset of them based on whether they have been published in academic journals, allows us to side-step the issue of data mining and look-ahead bias. As a consequence, our strategies should more accurately reflect the economic benefit of using machine learning forecasts for real-time investors.

We focus on fundamental signals for several reasons. First, fundamental analysis dates back at least to Graham and Dodd (1934), so it is natural to expect investors to consider fundamental signals as a potential class of return predictors (i.e., no look-ahead bias). Second, financial economists have long emphasized the importance of economic intuition behind predictors of expected returns (Fama and French, 1996; Cochrane, 2011), especially in a machine learning environment (Arnott, Harvey, and Markowitz, 2019). Fundamental signals are inherently related to firm cash flows and valuations and, therefore, have stronger economic foundations than most other classes of predictors. Third, one can construct a "universe" of fundamental signals using permutational arguments (Yan and Zheng, 2017). This is important because real-time investors have no way of knowing which signals turn out to be significant ex-post, so they have to learn from the universe of available signals.

The primary machine learning method we use is Boosted Regression Trees (BRT). We focus on BRT for several reasons. First, previous studies have shown that BRT exhibit strong predictive performance in finance applications. Gu, Kelly, and Xiu (2020), for example, show that BRT and neural networks are the two best-performing machine learning methods in predicting stock returns. Second, BRT are ideally suited for handling large, high-dimensional data sets because of their computational efficiency. This is important for us because our predictor set, which contains more than 18,000 signals, is much larger than those examined by previous studies. Third, BRT are robust to missing values, outliers, and the addition of irrelevant input variables. Finally, BRT are not "black boxes" like many other machine learning methods, and they are instead known for their interpretability.

We follow Gu, Kelly, and Xiu (2020) and partition our sample period 1963-2019 into a training period, a cross-validation period, and an out-of-sample test period. We form long-short portfolios based on machine learning predicted returns, i.e., buying stocks with high predicted returns and shorting stocks with low predicted returns. We find that the equal-weighted long-short portfolio generates an average return of 0.95% per month (t-statistic=6.63) and an annualized Sharpe ratio of 1.02 during the out-of-sample period 1987-2019. The performance of the value-weighted long-short portfolio is much weaker, earning an average return of 0.40% per month (t-statistic=2.34) and exhibiting a Sharpe ratio of 0.30.

The long-short returns and Sharpe ratios for our machine learning strategies, although statistically significant, are considerably lower than those documented by prior studies. Gu, Kelly, and Xiu (2020), for example, show that the long-short portfolios formed based on neural network forecasts earn an average return of 3.27% per month and an annualized Sharpe ratio of 2.45 in equal-weighted portfolios and an average return of 2.12% per month, and a Sharpe ratio of 1.35 in value-weighted portfolios. Similarly, Chen, Pelger, and Zhu (2022) and Freyberger, Neuhierl, and Weber (2020) report that the hedge portfolios constructed based on their models deliver an out-of-sample Sharpe ratio of 2.6 and 2.75, respectively. Thus, compared to the previous literature, our results indicate that the economic gains to real-time investors from using machine learning forecasts are much more modest.

Institutional investors are more likely to have the resources and sophistication to use machine learning methods. Previous studies (e.g., Gompers and Metrick (2001)) have shown that institutional investors prefer large, liquid stocks because they are more investable. To evaluate whether our machine learning strategies are profitable among large stocks, we repeat our analysis for subsamples of stocks sorted by firm size. We find that the out-of-sample performance of our machine learning strategies is statistically significant among small stocks but only marginally significant and, in some cases, insignificant among large stocks. The weak evidence of out-of-sample predictability among large stocks suggests that the economic benefit of using machine learning forecasts may be even more limited for institutional investors.

The relatively weak performance of our machine-learning strategies is not specific to BRT. We find even weaker evidence of out-of-sample predictability using neural network (NN) forecasts. Specifically, while the long-short returns are generally significant in equal-weighted portfolios, they are insignificant in value-weighted portfolios. Consistent with Gu, Kelly, and Xiu (2020), we find that shallow learning performs better than deep learning in neural networks.

Our analyses so far have focused on fundamental signals. The main reason for this focus is that we can construct a "universe" of fundamental signals (Yan and Zheng, 2017). Past return-based signals are another class of predictors for which we can construct an "exhaustive" list of signals. In particular, we follow Martin and Nagel (2022) and use the past 120 months (excluding the most recent month) of stock returns. As in our analysis of fundamental signals, we continue to use BRT as the primary machine-learning method. We find that the machine-learning strategy based on past-return signals earns an average return of 1.38% per month (t-statistic=4.93) and exhibits an annualized Sharpe ratio of 1.04 in equal-weighted portfolios. The performance of value-weighted portfolios is weaker; the average long-short return is 0.78% per month (t-statistic=2.41), while the Sharpe ratio is 0.46. The results based on neural network forecasts are qualitatively similar, with shallow networks performing similarly to BRT and deep networks performing significantly worse than BRT. Risk-adjusted returns indicate that the performance is significantly reduced when we control for the momentum factor. Overall, our analyses based on past-return signals paint a similar picture to that based on fundamental signals. That is, the performance of our machine learning strategies based on past-return signals is positive and significant but economically and statistically weaker than those reported by the prior literature.

One might be concerned that the relatively weak performance of our machine learning strategies is perhaps because our ML implementation is not as powerful as those employed by previous studies. To mitigate this concern, we replicate our machine learning analyses—both Boosted Regression Trees (BRT) and Neural Networks (NN)—on samples of published anomalies. We use both the Chen and Zimmermann (2022, CZ) covariates as well as the set of anomalies included in Green, Hand, and Zhang (2017, GHZ) and used by Gu, Kelly, and Xiu (2020). For the GHZ sample, we find that both BRT and shallow NNs deliver an out-of-sample long-short return in excess of 3.5% per month for equally weighted portfolios and a Sharpe Ratio between 2.35 and 2.95. These numbers are in line with the results in Gu, Kelly, and Xiu (2020) and substantially stronger than what we document for our "unmined" universe of predictors. The results that use the CZ sample of anomalies are even more impressive. For example, BRTs generate an equal-weighted long-short return of 5.18% per month and a Sharpe Ratio of 3.68. Taken together, these results indicate that our ML implementation is capable of generating rather strong performance when we use published predictors. As stated earlier, machine learning strategies based on subsequently discovered anomalies cannot be implemented in real time, and their performance is likely inflated due to a look-ahead bias. Nevertheless, the fact that we are able to replicate the strong performance of previous studies when we use published predictors indicates that our ML implementation is not the reason why the performance of our real-time implementable ML strategies is relatively weak.

Another potential explanation for the performance difference between our machine learning strategies and those employed by existing studies is the omission of short-term reversal from our predictor set. We perform two tests to evaluate this possibility. First, we remove short-term reversal from samples of published factors and construct machine learning strategies based on the remaining published factors. We again consider the samples of published anomalies from Green, Hand, and Zhang (2017) and Chen and Zimmermann (2022). We find that excluding short-term reversal reduces the machine learning strategy performance. However, the strong performance of the machine learning portfolios remains intact even after excluding short-term reversal from the samples of published factors. For example, excluding short-term reversal from the GHZ anomalies reduces the equal-weighted long-short return for the BRT strategy from 3.57% (t-stat=8.91) to 3.04% (t-stat=9.11). The corresponding numbers for the CZ anomalies are 5.18% (t-stat=9.91) and 4.87% (t-stat=10.05). The magnitude of the performance reduction is modest, and the long-short returns of these machine learning strategies, even after excluding short-term reversals, remain considerably higher than those based on our universe of fundamental signals.

In the second test, we add short-term reversal to our set of past-return signals and repeat our analysis. The results reveal a modest increase in the machine learning performance after short-reversal is added to the predictor set. For example, the equal-weighted long-short return for the BRT strategy is 1.38% (t-stat=4.93) without short-term reversal and is 1.81% (t-stat=6.40) with short-term reversal. The value-weighted long-short returns increase from 0.78% (t-stat=2.41) without short-term reversal to 0.98% (t-stat=3.14) with short-term reversal. Overall, we find that the long-short returns are modestly increased after including short-term reversal; however, they remain significantly lower than those for the machine learning strategies based on the GHZ and CZ samples of published factors. These results suggest that short-term reversal alone cannot explain the performance difference between our ML strategies and those based on published factors.

We focus on the gross, i.e., the before-trading cost performance of our machine learning strategies in this paper to facilitate comparison with prior literature (e.g., Gu, Kelly, and Xiu (2020); Freyberger, Neuhierl, and Weber (2020); Chen, Pelger, and Zhu (2022)). There is, however, growing attention to trading costs in both the anomaly literature and the machine learning literature (e.g., Novy-Marx and Velikov (2016); Chen and Velikov (2022); Jensen, Kelly, Malamud, and Pedersen (2022)). Next, we examine the after-trading-cost performance of our machine-learning strategies. We use Chen and Velikov (2022)'s low-frequency effective spreads as our trading cost measure and follow their approach to calculate the net returns of long-short strategies. We find that the net returns to our BRT strategies based on fundamental signals are positive: 0.73% per month for equal-weighted portfolios and 0.25% for value-weighted portfolios. In contrast, the net returns to strategies based on past-return signals are consistently negative. For example, the net return for BRT strategies is -0.97% per month for equal-weighted portfolios and -0.29% for value-weighted portfolios. Adding short-term reversal to the predictor set improves the gross returns, but the net returns remain negative.

We perform a number of robustness tests and additional analyses. We repeat our analysis using a rolling-window approach instead of a recursive one. If the relations between fundamental signals and future stock returns are unstable over time, then the rolling-window approach should perform better. Contrary to this argument, we find that our machine-learning strategies perform slightly worse under the rolling-window approach than under the recursive-window approach.

One might argue that the modest performance of our machine-learning strategies is due to our large universe of fundamental signals. In particular, if most of these signals are uninformative about future stock returns, machine-learning strategies based on our universe could be sub-optimal. As noted earlier, BRTs are robust to the addition of irrelevant predictor variables, so this is not a big concern for us. Nevertheless, to examine whether the performance of our machine learning strategies is hampered by the large size of our predictor set, we construct various subsets of our universe based on the prominence of the underlying accounting variables. Overall, we find no evidence that machine learning strategies based on smaller universes of fundamental signals perform significantly better.

We also compare the in-sample and out-of-sample performance of our machine-learning strategies. This analysis is motivated by Martin and Nagel (2022), who demonstrate that, in the age of Big Data, when investors face a high-dimensional prediction problem, there should be a substantial wedge between in-sample and out-of-sample predictability. Our results are consistent with this prediction. We find that, in contrast to the modest out-of-sample predictability, our fundamental signals exhibit strong in-sample predictability.

Our paper builds on and contributes to the recent literature employing machine learning methods in empirical asset pricing. Gu, Kelly, and Xiu (2020) use machine learning methods to measure risk premium and show that machine learning models, particularly trees and neural networks, significantly outperform linear regression models in predicting stock returns. Chen, Pelger, and Zhu (2022) estimate the SDF using deep neural networks and show that their model outperforms all other benchmark models. Freyberger, Neuhierl, and Weber (2020) use the adaptive group LASSO for model selection and show that their model exhibits superior out-of-sample performance. Kozak, Nagel, and Santosh (2020) use shrinkage and selection methods to construct an SDF that summarizes the joint explanatory power of a large cross-section of return predictors.[2] These studies have established the

[2]For additional studies that use machine learning methods in asset pricing, please also see, e.g., Rapach, Strauss, and Zhou (2013), Chinco, Clark-Joseph, and Ye (2019), Feng, Polson, and Xu (2020), Bryzgalova, Pelger, and Zhu (2020),

potential for large economic gains to investors using machine learning strategies. We complement the existing studies by taking the perspective of real-time investors. Specifically, we construct real-time-implementable machine learning strategies and show that they are significantly less profitable than those considered by prior literature.

Our paper is closely related to Avramov, Kaplanski, and Subrahmanyam (2022), who use machine learning methods to construct a fundamental index from 105 signals based on deviations of accounting variables from their recent averages. They show that this fundamental deviation index significantly predicts future stock returns. Our paper is also related to several earlier studies (Ou and Penman, 1989; Holthausen and Larcker, 1992; Haugen and Baker, 1996) that use machine leaning-like methods to predict future stock returns. Ou and Penman (1989) use a comprehensive set of accounting ratios to predict future unexpected earnings and then form trading strategies based on the predicted sign of future unexpected earnings. Holthausen and Larcker (1992) use the accounting ratios from Ou and Penman to directly predict future stock returns. Haugen and Baker (1996) examine the predictive ability of a comprehensive set of cross-sectional return predictors in the U.S. and globally.

Our paper is also related to Avramov, Cheng, and Metzker (2023), who show that much of the profitability of machine learning-based investment strategies is derived from difficult-to-arbitrage stocks or during periods when limits-to-arbitrage are elevated. Our finding that the out-of-sample predictability is significantly weaker among large stocks is consistent with Avramov, Cheng, and Metzker (2023). Limits-to-arbitrage, however, is not our primary focus. Our main argument is that machine learning strategies that use subsequently discovered anomaly variables as predictors, including those considered by Avramov, Cheng, and Metzker (2023), may not be implementable in real-time.

Finally, our paper is related to Arnott, Harvey, and Markowitz (2019) and Israel, Kelly, and Moskowitz (2020), who caution that machine learning methods may not work as well in finance as in some other disciplines. In particular, machine learning methods face three significant challenges in finance applications: the lack of data (on the time series dimension), the low signal-to-noise ratio, and the adaptive nature of financial markets. The modest performance of our real-time machine learning strategies could be a manifestation of these challenges faced by market professionals and investors.

The rest of our paper proceeds as follows. Section 2 describes our data, sample, and methods. Section 3 presents our main empirical results. Section 4 presents the results for additional analyses

Bianchi, Büchner, and Tamoni (2021), Dong, Li, Rapach, and Zhou (2022), Leippold, Wang, and Zhou (2022), Kelly and Xiu (2023), Geertsema and Lu (2023), Kaniel, Lin, Pelger, and Van Nieuwerburgh (2023) and Bali, Beckmeyer, Mörke, and Weigert (2023).

and robustness tests. Section 5 concludes.

2 Data, Sample, and Methods

This section describes the stock sample and associated fundamental signals we employ in our main analysis. We then describe the cross-sectional prediction problem underlying the portfolio strategies we generate and the main empirical method we use—Boosted Regression Trees (BRT). Third, we describe how we implement our machine learning strategy. Finally, we provide a discussion of the distinction between look-ahead bias and data-mining bias.

2.1 Stock Sample and Associated Fundamental Signals

We obtain monthly stock returns, share price, SIC code, and shares outstanding from the Center for Research in Security Prices (CRSP) and annual accounting data from Compustat. Our sample consists of the NYSE, AMEX, and NASDAQ common stocks (with a CRSP share code of 10 or 11) with the necessary data to construct fundamental signals and compute subsequent stock returns. We exclude financial stocks, i.e., those with a one-digit SIC code of 6. We also remove stocks with a share price lower than \$1. To mitigate backfilling biases, we require that a firm be listed on Compustat for two years before it is included in our sample (Fama and French, 1993). We obtain Fama and French (1996, 2015) factors and the momentum factor from Kenneth French's website and Hou, Xue, and Zhang (2015) q-factors from Lu Zhang's website.[3] Our sample spans from July 1963 to June 2019, and our sample consists of $15,035$ stocks.

We construct the universe of fundamental signals for our sample of stocks following Yan and Zheng (2017).[4] We start with 240 accounting variables (listed in Appendix B) and compute, for each variable, a total of 76 signals (listed in Appendix C). These signals are obtained by taking the original accounting variables and transforming them by computing changes, ratios, and other potentially economically meaningful transformations. The final number of fundamental signals we include in our analysis is $18,113$, which is slightly smaller than $18,240$ (240×76) because not all combinations of the accounting variables result in meaningful signals, and some of the combinations are redundant.

[3]Kenneth French's data library is located at `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`. The q-factors can be downloaded from `http://global-q.org/index.html`.

[4]To minimize our discretion, we use a pre-existing universe of fundamental signals instead of constructing one specifically for this study. Chordia, Goyal, and Saretto (2020) extend Yan and Zheng (2017) and construct a universe of over 2 million fundamental signals. We choose not to use this universe because real-time investors are unlikely to have the computing power to evaluate these many predictive variables in a machine learning context.

For brevity, we refer the readers to Yan and Zheng (2017) for complete details regarding selecting accounting variables and constructing fundamental signals.

2.2 Methodology

2.2.1 Prediction Equation. We predict the cross-section of stock returns using the following specification:

$$R_{i,t+1} = f\left(\mathbf{x}_{i,t}|\theta\right) + \epsilon_{i,t+1} \tag{1}$$

where $R_{i,t+1}$ denotes annual excess return for stock i from July of year t to June of year $t+1$, $\mathbf{x}_{i,t}$ denotes a vector of variables used to predict the cross section of returns, and θ denotes the parameters for the prediction function f. Stocks are indexed as $i = 1, \ldots, N$ and years are indexed by $t = 1, \ldots, T$.

The vector of predictive variables includes the $18,113$ fundamental signals described earlier. To make sure the accounting information is publicly available to investors, we follow Fama and French (1992) and pair accounting variables in year $t-1$ with stock returns from July of year t to June of year $t+1$. We follow Gu, Kelly, and Xiu (2020) and transform all fundamental signals as follows. We first rank all non-missing fundamental signals each year and then scale their ranks to the interval $[-1, +1]$. By construction, the cross-sectional median of the transformed fundamental signals is zero.

We predict annual excess returns for two reasons. First, our fundamental signals are constructed from annual financial statements and are updated annually. Second, the number of signals considered in our study is substantially larger than those in prior studies. Predicting annual returns is computationally more efficient than predicting monthly returns.[5]

2.2.2 Machine Learning Methods vs Linear Regressions. Traditionally, it was common in the literature to assume linearity of the f function and estimate Equation (1) using linear regression (LR) methods. More recently, the finance literature has instead started adopting more advanced Machine Learning (ML) methods.

One may expect that ML methods should have an advantage compared to linear regression methods because they feature 1) variable selection, 2) model combination, and 3) regularization/shrinkage, which allow them to handle large sets of conditioning information and stabilize their predictions by making them less sensitive to outliers.

[5]We conduct most of our empirical analyses on a high-performance cluster of 14 computing nodes, each of which is equipped with 128GB of RAM.

ML methods also allow to capture nonlinearities in the relations between the target variable and the regressors. When viewed through the lenses of the bias-variance trade-off, including nonlinearities allows for a smaller bias at the cost of a higher variance which positively relates to the instability of the predictions. In fact, a growing field in computer science, referred to as "adversarial machine learning," shows that even very small perturbations of the predictor variables can result in large changes in ML predictions.[6]

Similar effects could arise naturally in finance, where the data-generating process relating regressands and regressors constantly evolves. As profitable strategies are arbitraged away by smart money in a Shumpeterian creative destruction cycle, ML methods could potentially overfit certain temporary patterns that exist only in certain periods. This is particularly true for ML models with thousands (millions or even billions) of parameters that have been trained to capture deep, non-linear interactions because such a process makes them less adaptable to changes in the underlying dynamics of the data. These issues are further complicated by the fact that financial datasets are relatively small compared to those used in other fields, and financial research often faces weak signal-to-noise ratios (Kelly and Xiu, 2023). In these contexts, simpler models, like linear regression, could be more robust to changes in the data-generating process and deliver a more robust performance out-of-sample.

An important question is whether we should expect the advantages and disadvantages of ML models compared to LR models to vary depending on whether the researchers use "unmined" versus "known" predictors in their analysis. The theoretical literature does not provide a definitive answer to this question. Intuitively, on the one hand, we can expect ML methods to have a greater advantage compared to LR methods in the "unmined" predictor setting than in the "known" predictor setting because they feature regularization and variable selection. On the other hand, ML may have a smaller advantage relative to LR among "unmined" predictors because nonlinearities and variable interactions may be less important in higher-dimensional settings, and ML methods may be less robust to time variations in the relation between regressand and regressors. We leave an in-depth analysis of these theoretical and empirical issues to further research, and, in this work, we limit ourselves to highlighting that, while the majority of the literature shows tremendous promise for ML methods in finance, some of the results documented in the literature should be interpreted cautiously.

[6]see `https://towardsdatascience.com/breaking-neural-networks-with-adversarial-attacks-f4290a9a45aa` for an introduction to the topic and additional details.

2.2.3 Boosted Regression Trees.

Our baseline specification includes 18,113 fundamental signals. We choose the "off-the-shelf" machine learning tool called Boosted Regress Trees (BRT), in particular, the LightGBM implementation (Ke, Meng, Finley, Wang, Chen, Ma, Ye, and Liu, 2017) for our baseline analysis.

We choose BRT as our primary machine learning method for several reasons. First, BRT routinely rank among the very best machine learning algorithms in both finance and non-finance applications.[7] Second, BRT can handle large data sets with high dimensionality without overfitting because they simultaneously perform subsampling, model combination, and shrinkage. Third, BRT are robust to missing values and outliers (Hastie, Tibshirani, and Friedman, 2009). In particular, BRT are invariant under all monotone transformations of the individual input variables, making the forecasts generated robust to extreme values. Fourth, BRT are robust to the addition of irrelevant input variables (Friedman, 2001; Hastie, Tibshirani, and Friedman, 2009), because the underlying Classification and Regression Trees (CART) algorithm is designed to perform variable selection. Finally, because BRT are rooted in the CART framework, they possess good interpretability. For example, BRT return the rank and relative importance of all the potential regressors available, known as relative influence measures.[8] This feature distinguishes BRT from harder-to-interpret methods such as neural networks.

Regression Trees

A regression tree is built through a process known as binary recursive partitioning, which is an iterative process that splits the data into partitions or branches. Suppose we have P potential predictor ("state") variables and a single dependent variable over T observations, i.e., (x_t, y_{t+1}) for $t = 1, 2, \ldots, T$, with $x_t = (x_{t1}, x_{t2}, ..., x_{tp})$. Fitting a regression tree requires deciding (i) which predictor variables to use to split the sample space and (ii) which split points to use. The regression trees we use employ recursive binary partitions, so the fit of a regression tree can be written as an additive model:

$$f(x) = \sum_{j=1}^{J} c_j I\{x \in S_j\},$$

where $S_j, j = 1, \ldots, J$ are the regions we split the space spanned by the predictor variables into, $I\{\}$ is an indicator variable, and c_j is the constant used to model the dependent variable in each region. If the L^2 norm criterion function is adopted, the optimal constant is $\hat{c}_j = mean\,(y_{t+1}|x_t \in S_j)$.

[7] See a list of Machine Learning Challenge Winning Solutions on the LightGBM's website at https://github.com/microsoft/LightGBM/tree/master/examples.

[8] To conserve space, we provide a description of the relative influence measures in Appendix D. We also implement the relative influence measure on our data and report the results in Appendix D.

The globally optimal splitting point is difficult to determine, particularly in cases where the number of state variables is large. Hence, we use a sequential greedy algorithm. Using the full set of data, the algorithm considers a splitting variable p and a split point s so as to construct half-planes,

$$S_1(p,s) = \{X|X_p \leq s\} \quad and \quad S_2(p,s) = \{X|X_p > s\},$$

that minimize the sum of squared residuals:

$$\min_{p,s}\left[\min_{c_1}\sum_{x_t\in S_1(p,s)}(y_{t+1}-c_1)^2 + \min_{c_2}\sum_{x_t\in S_2(p,s)}(y_{t+1}-c_2)^2\right].$$

For a given choice of p and s, the fitted values, $\hat{c}_1$ and $\hat{c}_2$, are

$$\begin{aligned}\widehat{c}_1 &= \frac{1}{\sum_{t=1}^{T} I\{x_t \in S_1(p,s)\}}\sum_{t=1}^{T} y_{t+1} I\{x_t \in S_1(p,s)\},\\ \widehat{c}_2 &= \frac{1}{\sum_{t=1}^{T} I\{x_t \in S_2(p,s)\}}\sum_{t=1}^{T} y_{t+1} I\{x_t \in S_2(p,s)\}.\end{aligned}$$

The best splitting pair (p,s) in the first iteration can be determined by searching through each of the predictor variables, $p = 1,\ldots,P$. Given the best partition from the first step, the data is then partitioned into two additional states, and the splitting process is repeated for each of the subsequent partitions. Predictor variables that are never used to split the sample space do not influence the fit of the model, so the choice of splitting variable effectively performs variable selection.

Regression trees are ideally suited for handling high-dimensional data sets, incorporating multi-way interactions among predictors, and capturing non-linear relations between predictors and the predicted variable. However, the approach is sequential, and successive splits are performed on fewer and fewer observations, increasing the risk of fitting idiosyncratic data patterns. Furthermore, there is no guarantee that the sequential splitting algorithm leads to the globally optimal solution. To deal with these problems, we next consider a regularization method known as boosting.

Boosting

Boosting is based on the idea that combining a series of simple prediction models can lead to more accurate forecasts than those available from any individual model. Boosting algorithms iteratively re-weight data used in the initial fit by adding new trees in a way that increases the weight on observations modeled poorly by the existing collection of trees. From above, recall that a regression tree can be written as:

$$\mathcal{T}\left(x; \{S_j, c_j\}_{j=1}^{J}\right) = \sum_{j=1}^{J} c_j I\{x \in S_j\}.$$

A boosted regression tree is simply the sum of regression trees:

$$f_B(x) = \sum_{b=1}^{B} \mathcal{T}_b\left(x; \{S_{b,j}, c_{b,j}\}_{j=1}^{J}\right),$$

where $\mathcal{T}_b\left(x; \{S_{b,j}, c_{b,j}\}_{j=1}^{J}\right)$ is the regression tree used in the b-th boosting iteration and B is the number of boosting iterations. Given the model fitted up to the $(b-1)$-th boosting iteration, $f_{b-1}(x)$, the subsequent boosting iteration seeks to find parameters $\{S_{j,b}, c_{j,b}\}_{j=1}^{J}$ for the next tree to solve a problem of the form

$$\left\{\hat{S}_{j,b}, \hat{c}_{j,b}\right\}_{j=1}^{J} = \min_{\{S_{j,b}, c_{j,b}\}_{j=1}^{J}} \sum_{t=0}^{T-1} \left[y_{t+1} - \left(f_{b-1}(x_t) + \mathcal{T}_b\left(x_t; \{S_{j,b}, c_{j,b}\}_{j=1}^{J}\right)\right)\right]^2.$$

For a given set of state definitions ("splits"), $S_{j,b}$, $j = 1, \ldots, J$, the optimal constants, $c_{j,b}$, in each state are derived iteratively from the solution to the problem

$$\begin{aligned}\hat{c}_{j,b} &= \min_{c_{j,b}} \sum_{x_t \in S_{j,b}} \left[y_{t+1} - (f_{b-1}(x_t) + c_{j,b})\right]^2 \\ &= \min_{c_{j,b}} \sum_{x_t \in S_{j,b}} \left[e_{t+1,b-1} - c_{j,b}\right]^2,\end{aligned}$$

where $e_{t+1,b-1} = y_{t+1} - f_{b-1}(x_t)$ is the empirical error after $b-1$ boosting iterations. The solution to this problem is the regression tree that most reduces the average of the squared residuals $\sum_{t=1}^{T} e_{t+1,b-1}^2$, and $\hat{c}_{j,b}$ is the mean of the residuals in the j-th state.

Forecasts are simple to generate from this approach. The boosted regression tree is first estimated using data from $t = 1, \ldots, t^*$. Then, the forecast of y_{t^*+1} is based on the model estimates and the value of the predictor variable at time t^*, x_{t^*}. Boosting makes it more attractive to employ small trees (characterized by few terminal nodes) at each boosting iteration, reducing the risk that the regression trees will overfit. Moreover, by summing over a sequence of trees, boosting performs a type of model averaging that increases the stability and accuracy of the forecasts.

2.3 Implementation

We implement our BRT model by following Gu, Kelly, and Xiu (2020). We divide our sample period (1963-2019) into 12 years of training sample (1963-1974), 12 years of validation sample (1975-1986), and the remaining 33 years (1987-2019) for out-of-sample testing. We begin the out-of-sample period in 1987 in order to align with Gu, Kelly, and Xiu (2020).

We refit our model every year because our fundamental signals are updated annually. Each time we refit the model, we increase the training sample by one year while maintaining the length of the validation period at 12 years. This recursive window approach allows for the incorporation of all available information in generating forecasts. Every year, we generate return forecasts for all the stocks in our sample. We then construct decile portfolios based on the predicted returns. We hold these portfolios for 12 months and rebalance them every year. Our long-short strategy goes long in the decile portfolio with the highest BRT expected returns and short in the decile portfolio with the lowest BRT predicted returns.

To generate return forecasts, we need to estimate the model's parameters using the training data and specify two key hyper-parameters, i.e., the number of boosting iterations and the BRT shrinkage parameter (also known as the learning rate). To choose these two hyper-parameters, we adopt the commonly used grid search with validation procedure (Hastie, Tibshirani, and Friedman, 2009; Gu, Kelly, and Xiu, 2020).[9] We leave all other tuning parameters at their LightGBM default values.

Specifically, we first use the training sample to estimate the model under each set of hyper-parameter values. We then use the hyper-parameters that show the best performance during the validation period to re-estimate the final model. For example, suppose we want to forecast the cross-section of stock returns for 1987. We fit models under different hyper-parameter values during the training period 1963-1974 and then use the validation period 1975-1986 to gauge the performance of these trained models. We choose the hyper-parameters that deliver the best performance during the validation period and then use these hyper-parameters to re-estimate the final model for the combined training and validation period 1963-1986. When we move forward and forecast the cross-section of stock returns for 1988, our validation period rolls forward by one year and stays at 12 years, i.e., 1976-1987, while our training period increases by one year and goes from 1963 to 1975 (13 years).[10]

Our fundamental signals contain missing values. Although BRT can handle missing values, we

[9] Our grid for the number of boosting iterations is $\{100, 250, 500, 750, 1000\}$, while our grid for the learning rate is $\{0.01, 0.05, 0.10\}$.

[10] We show in Section 4.2 that our main results are robust to alternative training and validation periods.

pre-process the missing values to make BRT forecasts comparable to other machine learning methods that cannot handle missing values. Specifically, we follow the approach of Gu, Kelly, and Xiu (2020) and replace missing values with the cross-sectional median.[11] Recall that we have normalized all non-missing fundamental signals to the $[-1,+1]$ interval by using their cross-sectional ranks. By construction, the cross-sectional median of the transformed signals is zero. We, therefore, assign all missing values as zero.[12]

Performance Evaluation

Each year we sort all sample stocks into deciles based on BRT predicted returns, construct equal- and value-weighted portfolios, and focus on the long-short strategy that buys stocks in the top decile and shorts stocks in the bottom decile. We estimate CAPM 1-factor, Fama-French 3-factor, Carhart 4-factor, Fama-French 5-factor, Fama-French 5-factor + Momentum factor, and Q factor models by running the following time-series regressions:

$$r_t = \alpha + \beta\ MKT_t + \epsilon_t$$

$$r_t = \alpha + \beta\ MKT_t + s\ SMB_t + h\ HML_t + \epsilon_t$$

$$r_t = \alpha + \beta\ MKT_t + s\ SMB_t + h\ HML_t + u\ UMD_t + \epsilon_t$$

$$r_t = \alpha + \beta\ MKT_t + s\ SMB_t + h\ HML_t + r\ RMW_t + c\ CMA_t + \epsilon_t$$

$$r_t = \alpha + \beta\ MKT_t + s\ SMB_t + h\ HML_t + r\ RMW_t + c\ CMA_t + u\ UMD_t + \epsilon_t$$

$$r_t = \alpha + \beta\ MKT_t + s\ SMB_t + r\ ROE_t + i\ IA_t + \epsilon_t$$

where r_t is the long-short portfolio return based on BRT-generated forecasts for month t, and MKT, SMB, HML, UMD, RMW, CMA, ROE, and IA are market, size, value, momentum, profitability, investment (FF5), return on equity, and investment (Q) factors (Carhart, 1997; Fama and French, 2015; Hou, Xue, and Zhang, 2015). We focus on the alpha estimates and their t-statistics estimated using Newey and West (1987) standard errors.

[11] Chen and McCoy (2022) provide a rigorous justification for the use of mean/median imputation in machine learning studies. Specifically, Chen and McCoy (2022) compare different imputation methods in machine learning applications, and they find that simply imputing with cross-sectional averages does a surprisingly good job of capturing expected returns. Specifically, they find that mean/median imputation and sophisticated imputation methods lead to similar results. They argue that cross-sectional returns predictors are largely independent, and the independence implies observed predictors are uninformative about missing predictors, making ad-hoc methods valid.

[12] The performance of the BRT portfolios is similar without pre-processing the missing values.

2.4 Look-ahead bias versus data-mining bias

The main argument of our paper is that, from an investor's perspective, machine learning strategies based on subsequently discovered anomaly variables cannot be implementable in real time. Real-time investors could not have known what anomalies would be published decades later. Assuming otherwise leads to a significant look-ahead bias and potentially inflates the economic gains for real-time investors. We note that this is a distinct point from McLean and Pontiff (2016) in that they analyze what they call statistical biases while we are focusing on the effect of a look-ahead bias in existing ML research.

McLean and Pontiff (2016) use the term "statistical biases" to describe a broad array of biases inherent to the research and publication process, including data mining (i.e., many researchers search through many potential predictive signals in an effort to find significant relations) and the publication bias (it is easier to publish a significant result than a non-result). For ease of exposition, we will refer to these biases collectively as "data-mining bias" in our discussion below.

McLean and Pontiff (2016) show that anomaly returns are 26% lower out-of-sample and 58% lower post-publication. The 26% is their estimate of data-mining effects. The difference between the 58% and the 26%, i.e., the 32%, captures the effect of informed arbitrage according to McLean and Pontiff (2016). It shows that once a certain anomaly is published and disseminated, this information becomes available to market participants who act on it and arbitrage the anomaly away. The implication is that the observed predictability may no longer hold in subsequent periods or may be much weaker. This is not because of any bias in the initial study itself but rather because the publication of the study changed the underlying market dynamics.

In contrast to the data-mining bias, the look-ahead bias is a bias that arises when researchers use information that was not available at the time the strategy would have been implemented in the real world. In other words, it's a problem of using future information to make past decisions. This often occurs when researchers use variables that have been discovered or updated in the literature but which would not have been known or available to traders at the time the trading decisions were supposedly being made. As a result, the profitability of a trading strategy might be significantly overstated since it was actually based on information that could not have been used in real-time trading.

The data-mining bias and the look-ahead bias are distinct from each other in that the former typically arises in the context of evaluating in-sample predictability by econometricians who study the economy ex-post, whereas the latter typically arises in the context of evaluating the out-of-sample performance of a trading strategy by real-time investors.

Both data-mining effects and look-ahead biases can lead to overstatements of predictability or profitability (with the former inflating the in-sample predictability and the latter inflating the out-of-sample or real-time performance of the trading strategies), but they arise from different sources. The statistical bias arises from sampling variation and the selective nature of the research and publication process. The look-ahead bias arises from the inappropriate use of future information in a historical simulation of a trading strategy. Recognizing and understanding these biases is important for both researchers and practitioners in financial markets.

3 Main Results

In this section, we report the main results of our paper. We start by reporting in Section 3.1 the baseline results that compute the out-of-sample realized returns for BRT portfolios. We then report in Section 3.2 the abnormal performance of the BRT portfolios that control for various risk factors. Section 3.3 examines the performance of BRT long-short portfolios across large and small stocks. Section 3.4 uses an alternative machine learning method, i.e., neural networks. Section 3.5 examines the machine learning performance based on a universe of past-return signals. Section 3.6 examines whether our ML implementation can generate high long-short returns and Sharpe ratio using published predictors. Section 3.7 quantifies the role of short-term reversal in explaining the performance of our and previous machine learning strategies. Finally, Section 3.8 examines the after-trading-cost performance of our machine learning strategies.

3.1 Baseline Results

Table 1 shows the results of our baseline analysis. As stated earlier, we sort stocks into deciles each year based on one-year-ahead BRT predicted returns. We then construct a long-short portfolio that buys stocks with the highest BRT predicted returns and sells stocks with the lowest BRT predicted returns. We track the performance of these portfolios for 12 months. Following Gu, Kelly, and Xiu (2020), we report in Table 1 the BRT predicted returns (i.e., the sorting variable), the average realized returns, the standard deviation of realized returns, and the Sharpe ratios of BRT-sorted portfolios.

The left panel of Table 1 focuses on equally weighted portfolios. The first column shows the BRT predicted return, which is by construction monotonically increasing from decile 1 (-0.04% per month) to decile 10 (1.69% per month). The second column reports the out-of-sample average realized return for each portfolio: our primary variable of interest. We find that the performance of BRT portfolios increases nearly monotonically from decile 1 (-0.01%) to decile 10 (0.94%). The long-short portfolio earns an average return of 0.95% per month (or 11.4% per year), with a highly significant t-statistic of 6.63.

The standard deviation of the realized returns is U-shaped across the BRT decile portfolios, i.e., the portfolios with the lowest and the highest BRT predicted returns have higher volatilities than the other portfolios. Not surprisingly, we find that the long-short portfolio has a much lower volatility than the long-only portfolios. Finally, the last column of the left panel reports the annualized Sharpe ratio, which ranges from -0.01 to 0.48 across the ten BRT decile portfolios. The Sharpe ratio of the long-short portfolio is much higher at 1.02, which is primarily driven by the lower volatility of the long-short portfolio.

Equally weighted portfolios tend to overweight small-cap stocks that can be harder and more expensive to trade (e.g., Fama and French (2008) and Novy-Marx and Velikov (2016)). To mitigate this issue, we examine in the right panel of Table 1 the value-weighted portfolio returns. The BRT predicted return is again by construction monotonically increasing from decile 1 (0.00%) to decile 10 (1.61%). More importantly, the realized average portfolio return also increases from decile 1 (0.40%) to decile 10 (0.80%), although the relation is far from monotonic. The spread between decile 10 and decile 1 is 0.40% per month, or 4.8% per year.[13] Even though this spread is statistically significant at the 5% level, its magnitude is less than half of the spread for equally weighted portfolios. The Sharpe ratio exhibits a similar pattern, higher for decile 10 (0.47) than for decile 1 (0.22). The Sharpe ratio for the long-short portfolio is underwhelming at 0.30. As a comparison, the Sharpe ratio for the market portfolio over the same period is 0.45. Therefore, the out-of-sample performance of our real-time machine-learning strategies is economically modest.

Overall, we show in Table 1 that long-short portfolios formed based on BRT forecasts earn statistically significant returns, especially in equal-weighted portfolios. The magnitude of the long-short performance, however, is much lower than those documented in the prior literature. For example, the BRT models in Gu, Kelly, and Xiu (2020) achieve an equally weighted monthly long-short portfolio return of 2.14% per month and a Sharpe ratio of 1.73. The corresponding numbers for value-weighted

[13]These returns are before trading costs. We report the before-trading cost performance of our machine learning strategies for ease of comparison with prior literature (e.g., Chen, Pelger, and Zhu (2022); Freyberger, Neuhierl, and Weber (2020); Gu, Kelly, and Xiu (2020)). In Section 3.8, we examine the after-trading-cost performance of our machine learning strategies.

portfolios are 0.99% per month and a Sharpe ratio of 0.81.[14] The long-short portfolios formed based on neural network forecasts perform even better in Gu, Kelly, and Xiu (2020), earning an average return of 3.27% per month and an annualized Sharpe ratio of 2.45 in equal-weighted portfolios and an average return of 2.12% per month and a Sharpe ratio of 1.35 in value-weighted portfolios. Similarly, Chen, Pelger, and Zhu (2022) report an out-of-sample Sharpe ratio of 2.60 and Freyberger, Neuhierl, and Weber (2020) report that their model delivers an out-of-sample Sharpe ratio of 2.75. To sum up, while our results indicate that machine learning methods show promise in predicting stock returns, they are less extreme than those presented in the prior literature. The main difference between our paper and prior studies is that we employ a universe of fundamental signals that could have been employed *ex-ante* by a real-time investor. The conditioning information set we adopt is, therefore, free from data-mining concerns and look-ahead biases. Once we control for a more realistic information set, our results indicate that the economic gains to real-time investors from using machine learning methods are substantially smaller than previously documented.

3.2 Controlling for Common Risk Factors

The results in Table 1 do not control for risk exposures. It could be that the long-short portfolios based on BRT forecasts have positive and significant returns because they are exposed to well-known sources of risk, such as value or profitability. Table 2 shows the risk-adjusted performance of our BRT portfolios once we control for risk exposures using the six models described in Section 2.3. Irrespective of whether we use the CAPM model (columns 1-2), the Fama-French 3-factor model (columns 3-4), the Carhart 4-factor model (columns 5-6), the Fama-French 5-factor model (columns 7-8), the Fama-French 5-factor model augmented with momentum (columns 9-10) or the q-factor model (columns 11-12), we find that portfolios with higher BRT predicted returns have higher average realized risk-adjusted returns. Taking the Carhart 4-factor model as an example, we find that the alpha of decile 1 is negative and significant at -0.71% per month (t-statistic=-4.65), while the alpha of decile 10 is 0.37% per month (t-statistic=2.64). The resulting long-short portfolio has a monthly alpha of 1.08% and is statistically significant with a t-statistic of 6.43.

The results for value-weighted risk-adjusted returns are weaker than the equal-weighted results—in line with the findings in Table 1. Across the various risk-adjustment models, the monthly abnormal

[14]We note that we implement our BRT model using LightGBM, while Gu, Kelly, and Xiu (2020) implement using scikit-learn. When we implement our model using scikit-learn in conjunction with our fundamental signals, we obtain even less significant results than what we currently report in the paper.

performance ranges from a minimum of 0.46% (5.52% annualized) for the CAPM to a maximum of 0.80% (9.60% annualized) for the Fama-French 5-factor model with momentum. In all cases, the alphas of the long-short portfolios are statistically different from zero.

Consistent with the findings reported in Section 3.1, our results suggest that machine learning tools indeed can help predict stock returns. Still, the degree of predictability is significantly lower than what has been reported in the literature once we use as covariates the universe of signals that investors could have constructed in real-time and not the ones that have shown to be successful *ex-post* in predicting the cross-section of stock returns.

3.3 Focusing on Stocks with Different Market Capitalizations

The strategies we examine in this paper are more likely to be implemented by institutional investors rather than individual investors because they are the ones with the resources and sophistication to use machine learning methods. Previous studies (e.g., Gompers and Metrick (2001)) have shown that institutional investors prefer large-capitalization stocks because they are more liquid and more investable. To evaluate whether the profitability of our machine learning strategies varies across stocks with different capitalizations, each year we divide our sample stocks into two groups based on the median market capitalization: those above the median are large stocks and those below the median are small stocks. We then repeat our baseline analysis for each of these two groups of stocks and report the results in Table 3.

The top panel of Table 3 reports the results for equal-weighted portfolios. We find that the raw and risk-adjusted long-short returns are positive and significant for both large and smalls stocks. More importantly, the long-short performance is significantly higher for small stocks than for large stocks. Specifically, the long-short return is 0.63% per month (t-statistic=2.93) for the large stocks, and is 1.13% (t-statistic=6.14) for small stocks. The lower predictive performance for large stocks is not surprising. These stocks are likely to incorporate new information more quickly and are hence less likely to be predictable using machine learning algorithms.

The results for value-weighted portfolios are qualitatively similar. The average long-short return for large stocks is only 0.27% (t-statistic=1.23). The long-short returns for large stocks do become marginally significant when we control for risks using the Carhart 4-factor model, the Fama-French 5-factor model, the Fama-French 5-factor augmented with momentum, and the q-factor model. In comparison, the average long-short return for small stocks is economically and statistically significant whether we examine raw or risk-adjusted returns. For example, the average long-short return for small stocks is 1.16% (t-statistic=5.50).

Overall, the results in Table 3 indicate that the long-short performance of BRT portfolios is weaker for large stocks than for small stocks. This finding suggests that machine learning methods are better at predicting the returns of smaller stocks, for which news is incorporated more slowly into asset prices. As institutional investors are reluctant to trade smaller capitalization stocks because they are less liquid and less scalable, our results suggest that the benefits of machine learning strategies may be even more limited for institutional investors.

3.4 Neural Networks

In our baseline analysis, we use BRT, which is one of the most powerful machine learning methods for stock return predictions. Nevertheless, one might be concerned that our main results are model-specific and may not extend to other machine-learning methods. To ensure this is not the case, we extend our analysis to Neural Networks (NNs) mainly because—together with boosted regression trees—NNs are among the top performers when it comes to return prediction (Gu, Kelly, and Xiu, 2020; Bianchi, Büchner, and Tamoni, 2021). We follow Gu, Kelly, and Xiu (2020) and conduct our analysis using NNs with 1 to 5 hidden layers.

Our results, reported in Table 4, reveal several important findings. First, the equal-weighted long-short returns based on NNs are generally significant, while the value-weighted long-short returns are insignificant. Second, among equal-weighted portfolios, we find that shallow NNs perform better than deep NNs. For example, NNs with 1 and 2 hidden layers achieve long-short returns of 0.66% (t-statistics=3.19) and 0.85% per month (t-statistics=4.18), respectively. NNs with 3 or 4 hidden layers exhibit much lower but still significant long-short returns, while NNs with 5 hidden layers generate insignificant long-short portfolio returns. This finding is consistent with Gu, Kelly, and Xiu (2020), who show that shallow learning performs better than deep learning. Third, the performance of long-short portfolios based on neural network forecasts is much weaker than those documented by prior machine learning studies, particularly for value-weighted portfolios. Gu, Kelly, and Xiu (2020), for example, show that the long-short portfolios formed based on neural network forecasts earn an average return of 3.27% per month in equal-weighted portfolios and an average return of 2.12% in value-weighted portfolios. Overall, similar to BRT, our results based on neural networks suggest that the real-time performance of machine learning strategies is more modest than that portrayed by prior studies.

3.5 Past-return Signals

Our analyses so far have focused on fundamental signals. The main reason for this focus is that we can construct a "universe" of fundamental signals (Yan and Zheng, 2017). Past return-based signals represent another class of signals for which we can construct an "exhaustive" list. In this section, we follow Martin and Nagel (2022) and construct a universe of past return-based signals and then repeat our main analyses.[15] Specifically, we include in our universe the monthly returns during the past 120 months, excluding the most recent month.[16] Therefore, we have 119 past return-based signals for this analysis.

Our stock sample for this analysis consists of the NYSE, AMEX, and NASDAQ common stocks (with a CRSP share code of 10 or 11) with valid past return data. We exclude those stocks with a share price lower than \$1 at the end of month $t-1$. For ease of comparison with our analysis of fundamental signals and previous machine learning studies, the sample period of our past-return analysis spans from July 1963 to December 2019. We employ the same training, cross-validation, and out-of-sample testing periods as in our study of fundamental signals.

We continue to use BRT as the primary machine-learning method. We also examine Neural Networks with 1 to 5 hidden layers. As in our analysis of fundamental signals, we form long-short portfolios of stocks based on the machine learning predicted returns. Specifically, we go long in the stocks with the highest predicted returns and short in the stocks with the lowest predicted returns. We track the performance of these portfolios for one month and compute the return spread between the long and short portfolios. For performance evaluation, we report alphas for the long-short portfolio using the CAPM, the Fama-French three-factor model, and the Carhart four-factor model, the Fama-French five-factor alphas, Fama-French five-factor plus momentum factor alphas, and q-factor alphas. We report results for both equal-weighted and value-weighted portfolios.

Table 5 report the results.[17] We find that the BRT strategy based on past-return signals earns an average return of 1.38% per month (t-statistic=4.93) and exhibits an annualized Sharpe ratio of 1.04 in equal-weighted portfolios. The performance of value-weighted portfolios is significantly

[15] Moritz and Zimmermann (2016) and Murray, Xiao, and Xia (2022) also examine machine learning strategies based on past-return signals.

[16] Martin and Nagel (2022) exclude the most recent month to avoid microstructure effects. In Section 3.7, we repeat our analysis by adding the most recent month return to the predictor set.

[17] For brevity, we show decile-by-decile results in Table IA.18 in the Internet Appendix.

weaker. The average long-short return is 0.78% per month (t-statistic=2.41), while the Sharpe ratio is 0.46. The results based on neural network forecasts are qualitatively similar, with shallow networks (NN1 through NN3) performing similarly to BRT and deep networks (NN4 and NN5) performing significantly worse than BRT.

Risk-adjusted returns indicate that the performance is significantly reduced when we control for the momentum factor. For example, the Carhart alpha is 1.09% (t-statistic=6.62) for equal-weighted long-short portfolios and 0.63% (t-statistic=3.05) for value-weighted portfolios. The FF5+MOM alpha is even lower, at 0.78% (t-statistic=5.83) for equal-weighted long-short portfolios and 0.28% (t-statistic=1.55) for value-weighted portfolios. The smaller Carhart alpha and the FF5+MOM alpha are not surprising because much of the predictive ability of past returns is related to the momentum effect of Jegadeesh and Titman (1993).

Overall, our results based on past-return signals are broadly consistent with those based on fundamental signals. Specifically, we find significant long-short returns for our machine learning strategies, suggesting that real-time investors do benefit from using machine learning forecasts. However, the performance of these real-time machine-learning strategies is considerably weaker than those reported in the prior literature.

We would like to point out that although we examine both fundamental signals and past-return signals in our paper, our focus is on the fundamental signals. This is because, although one can construct a "universe" of past return signals (e.g., the past 120 months of returns in our paper), most of these signals are closely related to published, "known" predictors of stock returns. In particular, prior literature has documented (i) the long-run reversal anomaly (e.g., De Bondt and Thaler (1985)), which shows that past 3-5 years of returns negatively predict future returns; (ii) the momentum anomaly (Jegadeesh and Titman, 1993), which shows that past 6 to 12 months of returns positively predict future returns; (iii) the short-term reversal anomaly (e.g., Jegadeesh (1990)), which shows that past one-month return is a negative predictor of future returns, and (iv) seasonality anomaly (Heston and Sadka, 2008), which shows that lag returns at the 12-month frequency (i.e., t-12, t-24, t-36, etc.) positively predict future returns. Collectively, these previous studies have shown that more than half of the returns in the past 120 months are related to published predictors of stock returns.[18] As such, one might argue that the performance of machine-learning strategies based on these past-return signals may also be biased upward due to potential look-ahead concerns.

[18] In comparison, while our universe of fundamental signals also contains published signals (it should; otherwise, it is not a "universe"), published signals represent only a tiny fraction of our universe of fundamental signals.

3.6 ML Implementation

One might be concerned that the relatively weak performance of our machine learning strategies is perhaps due to our ML implementation not being as powerful as those employed in previous studies. To evaluate this possibility, we replicate our ML results—both Boosted Regression Trees (BRT) and Neural Networks (NN)—on samples of published anomalies.

The first sample is the Chen and Zimmermann (2022, CZ) predictors. We download the data from https://www.openassetpricing.com/. We use the March 2022 data release, which includes 207 anomaly predictors.[19] For ease of comparison with GKX, we also use the sample of 94 anomalies listed in Green, Hand, and Zhang (2017, GHZ).[20] We download the SAS code that generates the 94 predictors from Jeremiah Green's Web site at https://sites.google.com/site/jeremiahrgreenacctg/home. The out-of-sample testing period for this analysis is 1987-2019, the same as that for our main analyses based on fundamental signals and past-return signals.[21]

In Table 6, we report the results based on the GHZ sample of anomalies. We find that both BRT and shallow Neural Networks (NN1 through NN3) deliver an out-of-sample long-short return in excess of 3.5% per month for equally weighted portfolios and over 1.5% per month for value-weighted portfolios, in line with the results in Gu, Kelly, and Xiu (2020). We find very similar findings when we focus on risk-adjusted returns, as shown in the remaining columns of Table 6. We also find that BRT and shallow Neural Networks generate a Sharpe Ratio between 2.35 and 2.95 in equal-weighted portfolios. These numbers are consistent with prior literature (e.g., Gu, Kelly, and Xiu (2020)) as our ML implementation could generate Sharpe ratios of around 2.5 using published predictors.

In Table 7, we report the results that use the Chen and Zimmermann (2022) covariates. The results for this set of covariates are even more impressive. For example, BRTs generate an equal-weighted long-short return of 5.18% per month and a VW long-short return of 2.32% per month. Adjusting for risk using standard models reveals very similar findings. Furthermore, BRTs deliver an equally-weighted Sharpe Ratio of 3.68. The results for shallow Neural Networks are somewhat lower than those of BRTs but still very strong.

[19] The definitions of these variables are available at https://www.openassetpricing.com/march-2022-data-release/.

[20] GKX construct their data set based on GHZ's 94 characteristics, and GKX modified some variable definitions, see Gu, Kelly, and Xiu (2020, footnote 30). The GKX shared data set stops at 2016, and we cannot extend the data set. We also performed our analysis using GKX's shared data set and found similar magnitude of performance from 1987 to 2016.

[21] We also consider three alternative out-of-sample testing periods, namely 1987-2016, 1991-2004, and 1991-2014. The performance of machine-learning strategies during these alternative sample periods is qualitatively similar to and quantitatively stronger than that for 1987-2019.

Taken together, these results indicate that, our ML implementation is capable of generating rather strong performance when we use published predictors. Of course, as we have argued, machine learning strategies based on subsequently discovered anomalies cannot be implemented in real time and their performance is likely inflated due to a look-ahead bias. Nevertheless, the fact that we are able to replicate the strong ML performance of previous studies when we use published predictors indicates that our ML implementation is not the reason why the performance of ML strategies based on our universe fundamental signals is relatively weak.

3.7 The Role of Short-term Reversal

As noted earlier, we exclude short-term reversal (past one-month return) from our universe of past-return signals. Prior studies including Gu, Kelly, and Xiu (2020), Freyberger, Neuhierl, and Weber (2020), and Chen, Pelger, and Zhu (2022), however, include short-term reversal in their predictor sets. As such, one might reasonably wonder about the role of short-term reversal in explaining the performance difference between our machine learning strategies and those employed by previous studies. In particular, is the strong performance of previous ML strategies based on published factors due to their inclusion of short-term reversal? Is the relatively weak performance of our ML strategies driven entirely by the omission of short-term reversal in our predictor set?

We perform two analyses to address these questions. In the first analysis, we remove short-term reversal from the samples of published factors and construct machine learning strategies based on the remaining published factors. As before, we consider two samples of published factors: Green, Hand, and Zhang (2017)'s 94 anomaly variables and Chen and Zimmermann (2022)'s 207 anomaly factors. After excluding the short-term reversal factor, the GHZ sample includes 93 predictors, while the CZ sample includes 206 factors. The out-of-sample period is 1987-2019.[22]

We report the performance of the above machine learning strategies in Table 8 (the GHZ sample) and Table 9 (the CZ sample). For ease of comparison, we report, in each table, the performance of the machine learning strategy including the short-term reversal first and then the performance of the machine learning strategy excluding the short-term reversal.[23]

Overall, we find that excluding short-term reversal reduces the ML strategy performance. However,

[22]We also perform this analysis for alternative out-of-sample testing periods. For brevity, we report these results in Table IA.12 and Table IA.13 in the Internet Appendix. The qualitative results for these sub-periods are the same as those for 1987-2019.

[23]For brevity, we report raw long-short returns in Table 8 and Table 9. We report risk-adjusted returns in Table IA.14 and IA.15 in the Internet Appendix.

the strong performance of the machine learning portfolios remains intact even after excluding short-term reversal from the samples of published factors. For example, in Table 8, where we examine the GHZ anomalies, we find that the equal-weighted long-short return for BRT strategy is 3.57% (t-stat=8.91) when short-term reversal is included in the predictor set, and is 3.04% (t-stat=9.11) when the short-term reversal is excluded. Table 9, where we examine the CZ anomalies, we find that the equal-weighted long-short return for the BRT strategy is 5.18% (t-stat=9.91) when short-term reversal is included in the predictor set, and is 4.87% (t-stat=10.05%) when short-term reversal is excluded. The magnitude of the performance reduction is modest at best. Furthermore, the long-short returns of these ML strategies, even after excluding short-term reversals, remain considerably higher than those based on our universe of fundamental signals. Recall that the equal-weighted long-short return of the machine learning strategy based on our universe of fundamental signals is "merely" 0.95% per month. The results based on NN strategies and value-weighted portfolios are qualitatively similar. That is, we find that the performance of NN strategies is quantitatively reduced, but remains qualitatively similar even after excluding short-term reversal, suggesting that short-term reversal alone cannot explain the extraordinary performance of the machine learning strategies based on published factors.

In the second analysis, we add short-term reversal to our set of past-return signals and repeat our machine learning analysis. As a result of this addition, we have 120 past monthly returns in the predictor set. We report the performance of this ML strategy in Table 10. The results reveal a modest increase in the ML performance after short-reversal is added to the predictor set. For example, the EW long-short return for the BRT strategy is 1.38% (t-stat=4.93) without short-term reversal and is 1.81% (t-stat=6.40) with short-term reversal. The VW long-short returns increase from 0.78% (t-stat=2.41) without short-term reversal to 0.98% (t-stat=3.14) with short-term reversal.

Overall, we find that the omission of short-term reversal does have a moderate impact on the performance of our ML strategy. However, the long-short returns, even after including short-term reversal, remain significantly lower than those for the ML strategies based on GHZ and CZ samples of published factors. These results once again suggest that short-term reversal alone cannot explain the performance difference between our machine learning strategies and those based on published factors.

3.8 After-trading-cost Performance

For ease of comparison with prior literature (e.g., Gu, Kelly, and Xiu (2020); Freyberger, Neuhierl, and Weber (2020); Chen, Pelger, and Zhu (2022)), we focus on the gross performance of our machine learning strategies in this paper. There is, however, growing attention to trading costs in the anomaly literature and ML literature (e.g., Novy-Marx and Velikov (2016); Chen and Velikov (2022); Jensen et al. (2022)). In this section, we provide a simple analysis of the net performance (after-trading-cost returns) of our machine learning strategies.

We follow the general approach of Chen and Velikov (2022) to calculate turnover, trading costs, and net returns to long-short trading strategies. We also use their low-frequency (LF) measures of effective spreads as our trading cost measure.[24] These four LF measures are (i) Hasbrouck (2009)'s Gibbs sampler estimate, (ii) Corwin and Schultz (2012)'s high-low measure, (iii) Kyle and Obizhaeva (2016)'s volume-over-volatility measure, and (iv) Abdi and Ranaldo (2017)'s close-high-low measure. Following Chen and Velikov (2022), we use the average of the four low-frequency (LF) measures of effective spreads.

In Table 11, we show that the turnover rate for our BRT strategy based on fundamental signals is fairly low, with a two-sided turnover of 14% per month for both EW portfolios and VW portfolios. These relatively low turnover rates are not surprising because most of the fundamental signals are updated annually and we rebalance our portfolios once a year. We find that trading costs account for significantly less than half of the gross returns to our ML strategy. The net returns to the BRT strategy remain positive, at 0.73% per month for EW portfolios and 0.25% for VW portfolios. The net returns of NN strategies are mixed, with some being positive and others negative. We note that the gross returns reported here are slightly different from those of our baseline analysis. This is because the trading cost data is available only up to 2017, so the sample period for this analysis is slightly shorter than our baseline analysis.

Table 12 reports the corresponding results for our past-return-based machine-learning strategies. In contrast to those for fundamental signals, we find that the turnover rate for past-return-based machine-learning strategies is extremely high, well over 100% in both equal- and value-weighted portfolios. As a consequence, we find that net returns to machine learning strategies are consistently negative. For example, the net return for BRT strategies is -0.97% per month for equal-weighted portfolios and -0.29% for value-weighted portfolios. Adding short-term reversal to the predictor set improves the gross returns but makes the net returns even worse. Specifically, the net return is -1.48% per month for equal-weighted portfolios and -0.40% for value-weighted portfolios after including the short-term

[24] Due to the data availability issue, we do not adopt their high-frequency (HF) measures of effective spreads. We download the LF data from Andrew Chen's website at `https://sites.google.com/site/chenandrewy/`. We note their data is available up to 2017, so our analysis ends in 2017.

reversal. The results for NN strategies are mostly worse than those for BRT strategies.

Chen and Velikov (2022) note that LF spreads are biased upward by 25-50 basis points (compared to HF effective spreads) post decimalization. As such, the net returns to our machine learning strategies reported in Table 11 and Table 12 may be too low. We decided not to make an ad-hoc adjustment related to this bias because despite their upward bias relative to HF spreads, the LF spreads may underestimate the total trading costs because they do not include other important components of trading costs, such as the cost of short selling and price impact. The shorting cost is particularly important for us because our machine learning strategies are long-short strategies.

Overall, we show that the net performance of ML strategies is positive for fundamental signals and negative for past-return signals. We acknowledge that our analysis is exploratory and preliminary. An in-depth trading cost analysis that incorporates HF spreads, shorting cost, and price impact is a promising area of future research in the machine learning literature.[25]

4 Additional Results

In this section, we provide several extensions of our baseline analysis. Section 4.1 employs rolling windows instead of recursive windows in estimating the BRT model. Section 4.2 studies whether our results are robust to alternative training and validation periods. Section 4.3 examines different subsets of our universe of fundamental signals. Section 4.4 we compute the in-sample performance of BRT portfolios and then compare it with the out-of-sample performance. Finally, Section 4.5 investigates whether the performance of BRT portfolios varies with economic and market conditions. For brevity, we report the results of these additional analyses in the Internet Appendix.

4.1 Rolling Windows

We use recursive windows in our baseline specification to align ourselves with the majority of the literature (e.g., Gu, Kelly, and Xiu (2020)). Recursive windows allow for incorporating all available information in generating forecasts, but they can lead to poor forecasts if the data-generating process changes over time. An alternative to recursive windows is rolling windows that generate forecasts based on less information and hence are potentially less precise but are more robust to time variations in the relation between fundamental signals and returns. If the relation between the fundamental signals and

[25] We also examine the after-trading cost performance of machine-learning strategies based on the GHZ and CZ samples of published anomalies. For brevity, we report the results in Table IA.16 and Table IA.17 in the Internet Appendix.

stock returns is time-varying, rolling windows may improve the predictive power of machine learning algorithms. To assess this possibility, we repeat our main analysis using the rolling window approach described below.

We set the initial estimation period to 24 years so that our out-of-sample test period starts from 1987, the same as in the recursive window approach. To select the optimal hyper-parameters, we split the 24 years into training and validation periods following our baseline specification. In particular, our training period is 12 years, and the validation period is 12 years.[26] After obtaining the optimal hyper-parameters, we re-estimate the final model using the 24-year window. Each year we refit the model by moving the 24-year window forward by one year. The estimation period is fixed at 24 years under the rolling window approach. In comparison, under the recursive window approach, the estimation period expands as we roll forward.

Table IA.1 presents the performance of BRT portfolios for the rolling window approach. We find that the equally weighted portfolios achieve a long-short return of 0.83% per month (t-statistic=4.29) and a Sharpe ratio of 0.77. These numbers are lower than their counterparts for the recursive window approach. Specifically, in Table 1 we report that the equal-weighted portfolios exhibit a long-short return of 0.95% (t-statistic=6.63) and a Sharpe ratio of 1.02. The risk-adjusted returns for the rolling window approach are also correspondingly lower than those for the recursive window approach. The results for value-weighted portfolios paint a similar picture. For example, the average long-short return is 0.33% (t-statistic=1.35) under the rolling window approach, compared to the 0.40% (t-statistic=2.34) under the recursive window approach. Overall, we find that the performance of BRT portfolios is somewhat weaker for the rolling window approach than for the recursive window approach.

4.2 Alternative Training and Validation Periods

In our baseline specification, we use an initial training period of 12 years and a validation period of 12 years. In comparison, Gu, Kelly, and Xiu (2020) employ an initial training period of 18 years and a validation period of 12 years. As explained earlier, we choose an initial training period of 12 years because we want to start our out-of-sample test period in 1987, the same as in Gu, Kelly, and Xiu (2020). In this section, we examine whether our results are robust to our choices of the initial training period and validation period. Specifically, we consider nine alternative specifications in which the initial training period varies from 10 to 18 years, while the validation period varies from 10 to 14

[26]We have considered several alternative training and validation periods and find our results to be qualitatively similar.

years. We examine the performance of BRT portfolios under each of these alternative specifications.

Table IA.2 presents the results. The top panel reports the results for equal-weighted portfolios, while the bottom panel reports the results for value-weighted portfolios. For convenience, we reproduce the results for our baseline specification in the first row of each panel. Our baseline specification is denoted as "12+12", meaning 12 years of initial training period and 12 years of validation period. We denote the alternative specifications similarly. For example, "18+12" means 18 years of initial training and 12 years of validation period.

Overall, our results are highly robust across all alternative specifications. For example, the equal-weighted long-short returns range from 0.87% to 1.02% across the alternative specifications, compared to 0.95% for the baseline specification. Similarly, the value-weighted long-short returns range from 0.37% to 0.55% across the alternative specifications, compared to 0.40% for the baseline specification. The level of statistical significance for the long-short returns is also similar between the baseline and alternative specifications. Finally, the results on risk-adjusted returns are also robust to alternative specifications of initial training and validation periods.

4.3 Results Obtained on Subsets of the Fundamental Signals

Our baseline analysis employs a large number of predictor variables. Specifically, we construct a universe of 18,113 fundamental signals based on permutations of 240 accounting variables and 76 financial ratio configurations. One may argue that not all of these signals are actually considered by real-time investors and that the inclusion of these signals weakens the out-of-sample performance of our machine learning strategies. As noted earlier, BRT are known to be robust to the inclusion of irrelevant predictors. Nevertheless, to explore whether our relatively weak out-of-sample performance is driven by the large number of signals in our universe, we repeat our analysis on various subsets of the fundamental signals used in our baseline analysis.

4.3.1 Results using Subsets of the 240 Accounting Variables. In this section, we re-compute our baseline results for subsets of 240 accounting variables ranked based on the percentage of missing values across all stocks for the period 1963-2019. Some accounting variables are missing for all firms before a certain year. For example, all cash flow statement variables are missing before 1988. Including the years for which these variables were missing would artificially inflate their missing value proportions, so when computing the missing rate for an accounting variable, we exclude those years for which the variable is missing for all stocks. Appendix B reports the missing rates for all the 240 accounting variables in our data. Consistent with our expectation, we find that accounting variables with fewer missing values tend to be more important variables. For example, only 0.01% of the "total assets" are missing. For "total sales", the missing rate is also extremely low at 0.05%. In comparison, 71.1% of the "non-recurring discontinued operations" are missing.

The first row in each panel of Table IA.3 shows the BRT performance based on the fundamental signals constructed using the 30 accounting variables with the fewest missing values. The second row expands the set to 60 accounting variables with the lowest missing value rates. Each of the remaining rows increases the number of accounting variables by 30 compared to the previous row. The last row includes all 240 accounting variables, i.e., the entire universe of signals examined in our baseline analysis.

Examining equally weighted portfolio returns reported in the top panel, we find that the long-short performance is the worst when we include only the 30 accounting variables with the lowest missing value rate in our universe. Specifically, the equally weighted long-short return is 0.34% per month, statistically insignificant with a t-statistic of 1.07. As we increase the number of accounting variables in the subset, the portfolio performance generally improves. For example, the long-short portfolio based on 60 accounting variables delivers an average monthly return of 0.71% (t-statistic=2.09), while the one based on 210 accounting variables has an average return of 1.26% (t-statistic=6.67). The best performance is achieved when the number of accounting variables equals 180 or 210. We do observe a decline in performance as we increase the number of regressors from 210 to 240. The bottom panel of Table IA.3 reports the results for value-weighted portfolios. Similar to the equal-weighted results discussed above, we find that the performance of value-weighted portfolios generally increases with the size of the subset and peaks when the number of accounting variables is in the 180-210 range.

Overall, our results indicate that increasing the number of accounting variables included in the analysis enlarges the conditioning information set that can be exploited by real-time investors and generally increases the profitability of the machine learning strategies. This finding is inconsistent with the argument that the relatively weak performance of our machine learning strategies is due to the large size of our universe.

4.3.2 Results using Subsets of Financial Ratio Configurations. In this section, we repeat our baseline analysis on several subsets of the 76 financial ratio configurations. In constructing the universe of fundamental signals, we follow Yan and Zheng (2017) and use 15 base variables (Y) in addition to the 240 accounting variables (X). We refer the readers to Yan and Zheng (2017) for more details. We consider two subsets of Y's based on the importance of such base variables. The first subset includes the three most commonly used base variables—i.e., total assets, total sales, and market cap—which we term "Y3." The second subset (termed "Y5") includes two additional important base variables, i.e., total liability and shareholder's equity.

We also split the 76 financial ratio configurations into five categories based on their functional forms. The first category (P1) includes the ratios of accounting variables to base variables (i.e., the ratios #1 to #15 in Appendix C). The second category (P2) includes the change of ratios in the first category (i.e., the ratios #16 to #30). The third category (P3) contains the percentage change of ratios in the first category, or the ratios #31 to #45). The fourth category (P4) contains changes in accounting variables scaled by lagged base variables (i.e., the ratios #46 to #60). The fifth category (P5) includes the difference between the percentage changes in both accounting variables and base variables (i.e., the ratios #61 to #75).[27]

The top panel of Table IA.4 shows the BRT equal-weighted portfolio performance on the two subsets of the base variables and the five subsets of the financial ratio configurations. The equal-weighted long-short portfolio returns are positive and statistically significant regardless of which subset we examine. For example, when we use the three most important base variables (Y3), BRT achieve a significant long-short return of 0.89% per month, with an associated *t*-statistics of 4.51. Adding two additional base variables (Y5) leads to a long-short portfolio return of 0.90%, also statistically significant. Similar results hold for subsets from P1 to P5, where the average long-short returns range from 0.53% to 0.82%. It is important to note that the long-short returns for all seven subsets of financial ratios are lower than that for the full universe of fundamental signals (0.95%). The bottom panel of Table IA.4 repeats the exercise for value-weighted returns. As in our baseline analysis, value-weighted returns are weaker than equal-weighted returns and often lack statistical significance. We also find that the value-weighted long-short returns for subsets of financial ratios are generally lower than that for the full universe of fundamental signals. The only exception is the subset Y5. Overall, Table IA.4 shows little systematic evidence that the performance of our machine learning strategies would be much better had we considered a significantly smaller universe of fundamental signals.

[27] We note that for each category, we also include the percentage change of the accounting variable itself, i.e., the ratio #76 in Appendix C.

4.4 In-sample Performance versus Out-of-sample Performance

Standard asset pricing models assume rational expectations, i.e., investors know the model or the data-generating process. Martin and Nagel (2022) argue that, in the age of Big Data, this is unrealistic and that investors face a high-dimensional prediction problem instead. A central prediction of Martin and Nagel (2022) is that there should be a substantial wedge between in-sample and out-of-sample predictability. We follow Martin and Nagel (2022) and analyze data from the perspective of real-time investors. Moreover, we test their prediction by comparing the in-sample and out-of-sample performance of a universe of fundamental signals.

In Sections 3.1 and 3.2, we have already examined the out-of-sample performance of our universe of fundamental signals. We find that the long-short portfolios based on BRT forecasts earn positive and significant returns out of sample. For ease of comparison, the sample period for our in-sample analysis is the same as that for the out-of-sample test (i.e., 1987-2019). To conduct the in-sample analysis, we fit our BRT model using the full set of $18,113$ fundamental signals and use the fitted model to predict each year's returns. There is no consensus on how in-sample tests should be conducted in a machine-learning context. For robustness, we perform our test in two ways. In the first, we aim to maintain comparability with the out-of-sample analysis in Section 3.1 and select the optimal hyper-parameters using data from 1963 to 1986. We then use these hyper-parameters to perform an in-sample analysis for 1987-2019. In the second, we align our analysis with Martin and Nagel (2022) and select the optimal hyper-parameters using leave-one-year-out cross-validation over 1987-2019. The procedure uses a particular year t as the validation period and the remaining years as training periods. We choose the combination of hyper-parameters with the highest average validation performance across all years. Finally, we retrain and test the model for the entire 1987-2019 period using the optimally selected hyper-parameters.

We employ two performance metrics for our analyses. The first is the R^2 of the predictive regression. The second is the long-short portfolio return. Figure IA.1 in the Internet Appendix plots the times-series of in-sample and out-of-sample return predictability over the period 1987-2019. In Panel A, we compute cross-sectional predictive R^2s, but instead of averaging across all periods, we compute and plot 12-month moving averages. We depict the results for the two in-sample specifications in orange and red, respectively, and the out-of-sample results in blue. We also plot in dark grey the two-standard-error band around the out-of-sample R^2s. Panel B plots the BRT long-short portfolio returns for the two in-sample specifications and the out-of-sample long-short returns, adopting the same structure as Panel A. Both panels show that the in-sample predictability is consistently higher than the out-of-sample predictability. The gap between in-sample and out-of-sample predictability is often substantial, e.g., in the early 2000s. Overall, across both performance metrics, we find a significant degradation from in-sample performance to out-of-sample performance, consistent with the predictions of the theoretical model developed in Martin and Nagel (2022).

4.5 Testing for Time-varying Predictability

In Table IA.5, we examine whether the profitability of BRT strategies varies with economic and market conditions. Specifically, we split our sample period based on investor sentiment,[28] the VIX index also known as the "fear-gauge", market liquidity (Pástor and Stambaugh, 2003), business cycle indicators as published by NBER, and market state—proxied by the cumulative market returns over the previous 24 months. We also divide our sample period into two halves (1987-2003 and 2003-2019) to examine whether the predictability declines over time.

Panel A shows the long-short portfolio returns for high- and low-sentiment periods. When examining equal-weighted returns, we find significant predictability during both high- and low-sentiment periods. In contrast, value-weighted long-short returns are only marginally significant during low-sentiment periods and insignificant during high-sentiment periods. Whether we look at equal- or value-weighted returns, the difference in long-short returns between high- and low-sentiment periods is statistically insignificant. We find similar results in Panel B, where we divide the sample period into high- and low-VIX periods, and in Panel C, where we divide periods into high- and low-liquidity periods. In each panel, we find significant equal-weighted returns across both subperiods. The value-weighted returns, however, are either insignificant or marginally significant. As in Panel A, we find little significant evidence of differential predictability across subperiods. We also find little difference in predictability between recession and expansion periods in Panel D.

In Panel E, we split the sample period into UP and DOWN market states based on previous 24-month cumulative market returns. We find that the long-short return is higher during UP state than during DOWN state. Specifically, the equal-weighted long-short return is 1.24% during UP state and 0.67% during DOWN state. Similarly, the value-weighted long-short return is 0.79% during UP state and 0% during DOWN state. The differences in long-short returns between the UP and DOWN

[28] We obtain the investor sentiment's data from Wurgler's website at `http://people.stern.nyu.edu/jwurgler/`.

states are economically large and statistically marginally significant. In Panel F, we divide our sample period into two halves and find no statistically significant difference in predictability during the first and second half of our sample period.

Overall, the results in Table IA.5 indicate that the return predictability implied by our real-time machine learning strategies does not change significantly with investor sentiment, market volatility, market liquidity, or business cycle. However, there is some evidence that the profitability of our BRT strategies varies systematically with the market state. Finally, we find no evidence that the return predictability differs significantly across the two halves of our sample period.

5 Conclusions

Recent studies document strong performance for machine learning-based investment strategies. Our analyses paint a more conservative picture of the practical value of machine learning strategies for real-time investors. The machine learning strategies examined by prior studies use subsequently discovered anomaly variables as predictors of stock returns and cannot be implemented in real-time. We construct machine learning strategies based on a "universe" of fundamental signals. The out-of-sample performance of our strategies is positive and significant, but considerably weaker than those documented by previous studies, particularly in value-weighted portfolios. We find similar results when examining a universe of past return-based signals. The relative weak performance of our machine-learning strategies is not due to our ML implementation, as we are able to replicate the strong performance of machine learning strategies based on published anomalies. Nor is it driven by the omission of short-term reversal in our predictor set. Finally, we find that our machine learning strategies based on fundamental signals earn positive returns after trading cost, while those based on past-return signals earn negative net returns. Overall, our results indicate that machine learning strategies enhance investment performance, but the economic gains to real-time investors from using machine learning forecasts are more modest than previously thought.

References

Abdi, F., and A. Ranaldo. 2017. A simple estimation of bid-ask spreads from daily close, high, and low prices. *Review of Financial Studies* 30:4437–80.

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

Avramov, D., S. Cheng, and L. Metzker. 2023. Machine Learning vs. Economic Restrictions: Evidence from Stock Return Predictability. *Management Science* 69:2587–619.

Avramov, D., G. Kaplanski, and A. Subrahmanyam. 2022. Postfundamentals Price Drift in Capital Markets: A Regression Regularization Perspective. *Management Science* 68:7065–791.

Bali, T. G., H. Beckmeyer, M. Mörke, and F. Weigert. 2023. Option Return Predictability with Machine Learning and Big Data. *Review of Financial Studies* 36:3548–602.

Bianchi, D., M. Büchner, and A. Tamoni. 2021. Bond Risk Premiums with Machine Learning. *Review of Financial Studies* 34:1046–89.

Breiman, L., J. Friedman, C. J. Stone, and R. Olshen. 1984. *Classification and Regression Trees*. Chapman and Hall/CRC.

Bryzgalova, S., M. Pelger, and J. Zhu. 2020. Forest through the Trees: Building Cross-Sections of Stock Returns. Working Paper, London Business School.

Carhart, M. M. 1997. On Persistence in Mutual Fund Performance. *Journal of Finance* 52:57–82.

Chen, A. Y., and J. McCoy. 2022. Missing Values and the Dimensionality of Expected Returns. Working Paper arXiv:2207.13071.

Chen, A. Y., and M. Velikov. 2022. Zeroing in on the Expected Returns of Anomalies. *Journal of Financial and Quantitative Analysis* 1–83.

Chen, A. Y., and T. Zimmermann. 2022. Open Source Cross-Sectional Asset Pricing. *Critical Finance Review* 11:207–64.

Chen, L., M. Pelger, and J. Zhu. 2022. Deep Learning in Asset Pricing. *Management Science* forthcoming.

Chinco, A., A. D. Clark-Joseph, and M. Ye. 2019. Sparse Signals in the Cross-Section of Returns. *Journal of Finance* 74:449–92.

Chordia, T., A. Goyal, and A. Saretto. 2020. Anomalies and False Rejections. *Review of Financial Studies* 33:2134–79.

Cochrane, J. H. 2011. Presidential Address: Discount Rates. *Journal of Finance* 66:1047–108.

Cooper, M. J., H. Gulen, and M. J. Schill. 2008. Asset Growth and the Cross-Section of Stock Returns. *Journal of Finance* 63:1609–51.

Corwin, S. A., and P. Schultz. 2012. A simple way to estimate bid-ask spreads from daily high and low prices. *Journal of Finance* 67:719–60.

De Bondt, W. F. M., and R. H. Thaler. 1985. Does the stock market overreact. *Journal of Finance* 40:793–805.

Dong, X., Y. Li, D. Rapach, and G. Zhou. 2022. Anomalies and the Expected Market Return. *Journal of Finance* 77:639–81.

Fama, E. F., and K. R. French. 1992. The Cross-Section of Expected Stock Returns. *Journal of Finance* 47:427–65.

———. 1993. Common Risk Factors In The Returns On Stocks And Bonds. *Journal of Financial Economics* 33:3–56.

———. 1996. Multifactor Explanations of Asset Pricing Anomalies. *Journal of Finance* 51:55–84.

———. 2008. Dissecting anomalies. *Journal of Finance* 63:1653–78.

———. 2015. A Five-Factor Asset Pricing Model. *Journal of Financial Economics* 116:1–22.

Feng, G., N. Polson, and J. Xu. 2020. Deep Learning in Characteristics-Sorted Factor Models. SSRN Scholarly Paper ID 3243683.

Freyberger, J., A. Neuhierl, and M. Weber. 2020. Dissecting Characteristics Nonparametrically. *Review of Financial Studies* 33:2326–77.

Friedman, J. H. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29:1189–232.

Geertsema, P., and H. Lu. 2023. Relative Valuation with Machine Learning. *Journal of Accounting Research* 61:329–76.

Gompers, P. A., and A. Metrick. 2001. Institutional Investors and Equity Prices. *Quarterly Journal of Economics* 116:229–59.

Graham, B., and D. L. F. Dodd. 1934. *Security analysis*. McGraw-Hill, New York.

Green, J., J. R. M. Hand, and X. F. Zhang. 2017. The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns. *Review of Financial Studies* 30:4389–436.

Gu, S., B. T. Kelly, and D. Xiu. 2020. Empirical Asset Pricing via Machine Learning. *Review of Financial Studies* 33:2223–73.

Harvey, C. R., Y. Liu, and H. Zhu. 2016. ... and the Cross-Section of Expected Returns. *Review of Financial Studies* 29:5–68.

Hasbrouck, J. 2009. Trading costs and returns for us equities: Estimating effective costs from daily data. *Journal of Finance* 64:1445–77.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Haugen, R. A. R. A., and N. L. Baker. 1996. Commonality in the determinants of expected stock returns. *Journal of Financial Economics* 41:401–39.

Heston, S. L., and R. Sadka. 2008. Seasonality in the cross-section of stock returns. *Journal of Financial Economics* 87:418–45.

Holthausen, R. W., and D. F. Larcker. 1992. The prediction of stock returns using financial statement information. *Journal of Accounting and Economics* 15:373–411.

Hou, K., C. Xue, and L. Zhang. 2015. Digesting Anomalies: An Investment Approach. *Review of Financial Studies* 28:650–705.

Israel, R., B. T. Kelly, and T. J. Moskowitz. 2020. Can Machines "Learn" Finance? *Journal of Investment Management* 18:23–36.

Jegadeesh, N. 1990. Evidence of Predictable Behavior of Security Returns. *Journal of Finance* 45:881–98.

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

Jensen, T. I., B. T. Kelly, S. Malamud, and L. H. Pedersen. 2022. Machine Learning and the Implementable Efficient Frontier. Working Paper, Swiss Finance Institute Research Paper No. 22-63.

Kaniel, R., Z. Lin, M. Pelger, and S. Van Nieuwerburgh. 2023. Machine-learning the skill of mutual fund managers. *Journal of Financial Economics* 150:94–138.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems 30*, 3146–54. Red Hook, NY, United States: Curran Associates Inc.

Kelly, B. T., and D. Xiu. 2023. Financial Machine Learning. SSRN Scholarly Paper.

Kozak, S., S. Nagel, and S. Santosh. 2020. Shrinking the Cross-Section. *Journal of Financial Economics* 135:271–92.

Kyle, A. S., and A. A. Obizhaeva. 2016. Market microstructure invariance: Empirical hypotheses. *Econometrica* 84:1345–404.

Leippold, M., Q. Wang, and W. Zhou. 2022. Machine learning in the Chinese stock market. *Journal of Financial Economics* 145:64–82.

Martin, I., and S. Nagel. 2022. Market Efficiency in the Age of Big Data. *Journal of Financial Economics* 145:154–77.

McLean, R. D., and J. Pontiff. 2016. Does Academic Research Destroy Stock Return Predictability? *Journal of Finance* 71:5–32.

Moritz, B., and T. Zimmermann. 2016. Tree-based conditional portfolio sorts: The relation between past and future stock returns. Working Paper.

Murray, S., H. Xiao, and Y. Xia. 2022. Charting By Machines. SSRN Scholarly Paper 3853436, Rochester, NY.

Newey, W. K., and K. D. West. 1987. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* 55:703–8.

Novy-Marx, R., and M. Velikov. 2016. A Taxonomy of Anomalies and Their Trading Costs. *Review of Financial Studies* 29:104–47.

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

Pástor, L., and R. F. Stambaugh. 2003. Liquidity Risk and Expected Stock Returns. *Journal of Political Economy* 111:642–85.

Rapach, D. E., J. K. Strauss, and G. Zhou. 2013. International Stock Return Predictability: What Is the Role of the United States? *Journal of Finance* 68:1633–62.

Yan, X. S., and L. Zheng. 2017. Fundamental Analysis and the Cross-Section of Stock Returns: A Data-Mining Approach. *Review of Financial Studies* 30:1382–423.

Table 1: Performance of Portfolios Sorted by BRT Predicted Returns

	Equal-weighted					Value-weighted				
Rank	*Pred*	*Avg*	*t-stat*	*SD*	*SR*	*Pred*	*Avg*	*t-stat*	*SD*	*SR*
1 (Low)	-0.04	-0.01	-0.05	7.51	-0.01	0.00	0.40	1.30	6.18	0.22
2	0.30	0.49	1.58	6.22	0.28	0.30	0.58	2.39	5.53	0.36
3	0.49	0.65	2.12	6.02	0.37	0.50	0.58	2.63	4.73	0.42
4	0.64	0.74	2.65	5.64	0.46	0.64	0.75	3.21	4.60	0.56
5	0.73	0.76	2.72	5.45	0.48	0.73	0.60	2.34	4.61	0.45
6	0.80	0.90	3.23	5.44	0.58	0.80	0.66	2.99	4.51	0.51
7	0.88	0.90	3.18	5.57	0.56	0.88	0.68	2.68	4.93	0.48
8	0.97	0.96	3.17	5.43	0.62	0.97	0.49	1.81	4.82	0.35
9	1.12	0.93	2.84	5.78	0.56	1.11	0.64	2.20	5.15	0.43
10 (High)	1.69	0.94	2.55	6.71	0.48	1.61	0.80	2.51	5.96	0.47
10-1	1.74	0.95	6.63	3.26	1.02	1.61	0.40	2.34	4.68	0.30

This table reports the excess returns of decile portfolios sorted by BRT predicted returns from 1987 to 2019. We predict stock annual excess returns using 18,113 fundamental signals (as described in Section 2.1). We use a recursive window approach and select the optimal hyper-parameters using a cross-validation approach. Our initial estimation period is 1963-1986. The first 12 years is the training period and the second 12 years is the validation period. As we roll forward, the training period expands while the validation period stays at 12 years. The left panel reports equal-weighted portfolio results. In the first column of this panel we report the average predicted monthly returns from the BRT model (*Pred*). The second and third columns report the average realized monthly excess returns (*Avg*) and associated t-statistics (*t-stat*), computed using Newey and West (1987) standard errors with 12 lags. Finally, in the fourth and fifth column we report the portfolios' return standard deviations (*SD*) and Sharpe ratios (*SR*), respectively. The right panel reports the same results using value-weighted portfolio returns. All returns are expressed in percent per month.

Table 2: Risk-adjusted Performance of Portfolios Sorted by BRT Predicted Returns

Rank	CAPM		FF3		Carhart		FF5		FF5+Mom		Q	
Equal Weighted												
	alpha	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
L(ow)	-0.84	-3.72	-0.76	-4.91	-0.71	-4.65	-0.44	-3.43	-0.43	-3.33	-0.30	-1.78
2	-0.21	-1.00	-0.19	-1.50	-0.17	-1.48	-0.09	-0.74	-0.09	-0.72	0.02	0.17
3	-0.07	-0.34	-0.06	-0.64	-0.06	-0.62	-0.03	-0.36	-0.03	-0.34	0.07	0.70
4	0.05	0.29	0.03	0.50	0.03	0.48	0.02	0.23	0.02	0.24	0.09	1.29
5	0.10	0.58	0.07	1.02	0.07	1.06	0.02	0.32	0.03	0.38	0.09	1.16
6	0.24	1.23	0.18	2.37	0.23	2.80	0.11	1.44	0.15	1.91	0.19	2.03
7	0.23	1.26	0.21	3.05	0.20	2.74	0.22	2.79	0.21	2.65	0.26	3.73
8	0.30	1.59	0.28	3.17	0.29	3.34	0.28	3.23	0.29	3.34	0.35	3.84
9	0.22	1.33	0.22	2.22	0.29	3.08	0.35	3.25	0.39	3.74	0.45	4.87
H(igh)	0.18	0.84	0.25	1.60	0.37	2.64	0.58	3.87	0.65	4.43	0.68	4.67
H-L	1.01	6.30	1.01	6.35	1.08	6.43	1.03	5.42	1.08	5.60	0.98	5.11
Value Weighted												
Rank	CAPM		FF3		Carhart		FF5		FF5+Mom		Q	
	alpha	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
L(ow)	-0.36	-2.40	-0.26	-2.30	-0.31	-2.74	-0.13	-1.13	-0.18	-1.58	-0.09	-0.59
2	-0.11	-0.76	-0.03	-0.20	-0.04	-0.28	0.01	0.08	0.00	0.00	0.08	0.54
3	-0.05	-0.53	-0.02	-0.23	-0.03	-0.33	-0.08	-0.82	-0.08	-0.84	-0.02	-0.16
4	0.12	1.61	0.15	2.00	0.10	1.27	0.05	0.68	0.02	0.30	0.08	0.88
5	-0.01	-0.14	-0.01	-0.10	0.01	0.11	-0.12	-1.47	-0.10	-1.37	-0.07	-0.98
6	0.08	0.80	0.04	0.47	0.06	0.70	-0.17	-2.03	-0.13	-1.51	-0.12	-1.69
7	0.04	0.41	0.00	0.03	0.03	0.35	-0.13	-1.37	-0.10	-1.01	-0.11	-1.13
8	-0.13	-1.10	-0.13	-1.02	-0.07	-0.60	-0.17	-1.26	-0.12	-0.99	-0.09	-0.61
9	-0.02	-0.10	0.06	0.43	0.11	1.00	0.14	1.25	0.17	1.60	0.25	1.92
H(igh)	0.10	0.46	0.20	1.23	0.34	2.36	0.54	3.80	0.62	4.24	0.58	4.03
H-L	0.46	2.17	0.46	2.57	0.65	3.66	0.68	3.18	0.80	3.80	0.68	2.94

This table shows the risk-adjusted performance of the BRT portfolios based on the CAPM model, the Fama-French 3-factor model, the Carhart 4-factor model, the Fama-French 5-factor model, the Fama-French 5-factor model augmented with momentum factor, and the q-factor model. The top panel reports results for equal-weighted portfolios. The bottom reports results for value-weighted portfolios. All returns are expressed in percent per month.

Table 3: Performance of Portfolios Sorted by BRT Predicted Returns – Subsets of Stocks Based on Market Capitalization

Equal Weight														
Sub-Samples	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+MOM		Q	
	Avg	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
Large Stocks	0.63	2.93	0.57	0.71	3.17	0.74	3.11	0.77	3.70	0.93	3.39	0.93	3.63	0.90	3.27
Small Stocks	1.13	6.14	1.16	1.18	6.48	1.20	6.21	1.22	5.98	1.23	5.48	1.24	5.35	1.14	5.10

Value Weight															
Sub-Samples	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+MOM		Q	
	Avg	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
Large Stocks	0.27	1.23	0.20	0.35	1.37	0.41	1.77	0.52	2.33	0.60	2.39	0.67	2.58	0.66	2.63
Small Stocks	1.16	5.50	1.04	1.18	5.58	1.21	5.29	1.20	5.08	1.27	5.12	1.25	4.87	1.16	4.41

This table shows the performance of BRT portfolios for large stocks (with above median market capitalization) and small stocks (with below median market capitalization). All results are computed following the procedure described in Table 1. The first three columns report average monthly returns for the long-short portfolios as well as the associated Sharpe Ratios (SR). The remaining columns report risk-adjusted returns—see Table 2 for details. The top panel reports results for equal-weighted returns. The bottom panel reports results for value-weighted returns. All returns are expressed in percent per month.

Table 4: Performance of Portfolios Sorted by NN Predicted Returns

Equal Weight															
Method	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+MOM		Q	
	Avg	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
NN1	0.66	3.19	0.61	0.70	3.02	0.65	3.26	0.83	5.09	0.73	3.27	0.86	4.45	0.74	3.21
NN2	0.85	4.18	0.92	1.02	5.04	0.93	5.33	0.83	4.93	0.61	4.36	0.56	4.01	0.57	3.56
NN3	0.31	1.76	0.40	0.38	1.99	0.41	2.34	0.37	2.23	0.40	2.83	0.37	2.67	0.37	2.49
NN4	0.41	2.07	0.47	0.37	1.62	0.41	1.98	0.44	2.34	0.62	3.67	0.63	3.79	0.61	3.55
NN5	0.09	0.48	0.10	0.06	0.28	0.10	0.53	0.13	0.74	0.12	0.82	0.14	1.00	0.26	1.74

Value Weighted															
Method	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+MOM		Q	
	Avg	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
NN1	0.01	0.03	0.01	0.01	0.03	-0.01	-0.04	0.23	1.18	0.15	0.59	0.32	1.46	0.24	0.99
NN2	0.04	0.15	0.03	0.21	0.79	0.08	0.40	-0.01	-0.04	-0.23	-1.17	-0.27	-1.36	-0.25	-1.13
NN3	-0.01	-0.02	-0.01	0.07	0.28	0.06	0.23	0.16	0.71	-0.10	-0.41	-0.01	-0.06	0.00	-0.01
NN4	0.20	0.60	0.15	0.13	0.30	0.23	0.69	0.22	0.72	0.54	2.02	0.51	1.94	0.54	1.80
NN5	0.20	0.75	0.16	0.11	0.37	0.15	0.56	0.17	0.68	0.12	0.56	0.14	0.67	0.25	1.07

This table shows the performance of long-short portfolios sorted by Neural Network (NN) predicted returns. We consider NN models with hidden layers that range from 1 through 5. The first three columns report average monthly returns for the long-short portfolios as well as the associated Sharpe Ratios. The remaining columns report risk-adjusted returns—see Table 2 for details. The top panel reports results for equal-weighted returns. The bottom panel reports results for value-weighted returns. All returns are expressed in percent per month.

Table 5: Performance of Portfolios Sorted by ML Predicted Returns on Past-Return Signals

Method	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+Mom		Q	
	Avg	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
							Equal Weight								
BRT	1.38	4.93	1.04	1.63	5.86	1.56	6.90	1.09	6.62	1.09	4.82	0.78	5.83	0.78	3.89
NN1	1.17	6.61	1.33	1.32	7.56	1.32	7.81	0.93	9.33	1.14	5.99	0.87	7.96	0.88	4.83
NN2	1.40	7.46	1.45	1.53	7.81	1.50	8.55	1.10	9.69	1.20	5.24	0.93	6.86	0.96	4.31
NN3	1.30	6.34	1.42	1.42	6.90	1.40	7.62	1.01	8.81	1.16	5.74	0.89	7.60	0.93	4.82
NN4	0.74	5.64	0.81	0.75	5.95	0.76	5.75	0.48	2.53	0.62	2.98	0.43	1.99	0.43	1.74
NN5	0.10	0.50	0.11	-0.02	-0.07	0.09	0.54	0.02	0.14	0.33	2.71	0.27	2.00	0.29	2.18
							Value Weight								
BRT	0.78	2.41	0.46	1.17	4.00	1.07	4.23	0.63	3.05	0.56	2.37	0.28	1.55	0.28	1.36
NN1	0.87	4.29	0.72	0.98	4.54	1.02	5.08	0.57	3.34	0.90	4.23	0.58	3.14	0.69	3.08
NN2	1.06	4.34	0.78	1.21	4.73	1.23	4.91	0.77	4.49	1.00	3.36	0.68	3.66	0.83	2.81
NN3	1.00	3.79	0.80	1.09	4.22	1.12	4.35	0.63	3.76	0.98	3.32	0.63	3.70	0.78	2.44
NN4	0.38	1.68	0.31	0.39	1.67	0.43	1.97	0.10	0.37	0.39	1.50	0.16	0.58	0.27	0.93
NN5	0.32	1.30	0.26	0.19	0.76	0.32	1.73	0.17	0.88	0.56	3.20	0.43	2.53	0.49	2.17

This table reports the returns and risk-adjusted performance for the long-short portfolios sorted by ML-predicted returns from 1987 to 2019. We predict stock monthly excess returns using 119 past-return signals (as described in Section 3.5). Our full sample starts from 1963 and the out-of-sample periods begin from 1987, which is consistent with our baseline specifications on fundamental signals. We use a recursive window approach and select the optimal hyper-parameters using a cross-validation approach. Our initial estimation period is 1963-1986. The first 12 years is the training period and the second 12 years is the validation period. As we roll forward, the training period expands while the validation period stays at 12 years. The risk-adjusted performance are calculated based on the CAPM model, the Fama-French 3-factor model, the Carhart 4-factor model, the Fama-French 5-factor model, the Fama-French 5-factor model augmented with momentum factor, and the Q-factor model. The top panel reports results for equal-weighted portfolios. The bottom reports results for value-weighted portfolios. All returns are expressed in percent per month.

Table 6: Performance of Portfolios Sorted by ML Predicted Returns on the GHZ Sample

Method	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+Mom		Q	
	Avg	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
							Equal Weight								
BRT	3.57	8.91	2.35	3.67	9.36	3.69	9.13	3.37	8.61	3.66	7.85	3.43	8.05	3.35	6.71
NN1	3.78	9.74	2.69	3.86	9.87	3.90	10.06	3.60	9.48	3.78	9.34	3.57	9.10	3.67	8.82
NN2	3.87	9.93	2.95	3.94	10.00	3.96	10.08	3.66	9.43	3.78	9.60	3.57	9.36	3.61	9.00
NN3	3.94	10.09	2.84	4.03	10.31	4.05	10.40	3.75	9.57	3.90	9.39	3.69	9.23	3.79	8.98
NN4	2.94	9.59	2.27	3.01	9.78	2.99	9.78	2.78	7.82	2.89	8.98	2.75	7.71	2.71	7.88
NN5	1.70	5.92	1.33	1.62	5.29	1.69	5.44	1.71	5.37	1.77	5.41	1.78	5.37	1.90	5.50
							Value Weight								
BRT	1.51	4.53	0.71	1.66	5.50	1.70	5.14	1.13	4.08	1.56	3.25	1.15	3.37	1.09	2.20
NN1	1.68	4.84	0.85	1.90	5.56	1.96	6.01	1.30	4.41	1.70	4.49	1.24	4.11	1.33	3.54
NN2	1.87	5.18	0.99	2.04	5.56	2.08	6.13	1.42	4.74	1.74	4.73	1.29	4.34	1.43	3.56
NN3	1.71	4.93	0.90	1.94	5.38	1.95	6.08	1.25	5.45	1.48	3.87	1.00	3.98	1.16	3.13
NN4	1.37	6.32	0.82	1.51	6.23	1.48	7.41	0.98	4.19	1.25	4.71	0.90	3.33	0.89	3.20
NN5	0.69	3.78	0.49	0.76	3.67	0.81	4.22	0.57	2.59	0.81	4.02	0.63	2.93	0.72	3.55

This table reports the returns and risk-adjusted performance for the long-short portfolios sorted by ML-predicted returns on the GHZ sample of signals from 1987 to 2019. We predict stock monthly excess returns using the 94 signals collected by Green, Hand, and Zhang (2017). Gu, Kelly, and Xiu (2020) derived their 94 signals from Green, Hand, and Zhang (2017). We use a recursive window approach and select the optimal hyper-parameters using a cross-validation approach. Our initial estimation period is 1963-1986. The first 12 years is the training period and the second 12 years is the validation period. As we roll forward, the training period expands while the validation period stays at 12 years. The risk-adjusted performance are calculated based on the CAPM model, the Fama-French 3-factor model, the Carhart 4-factor model, the Fama-French 5-factor model, the Fama-French 5-factor model augmented with momentum factor, and the q-factor model. The top panel reports results for equal-weighted portfolios. The bottom reports results for value-weighted portfolios. All returns are expressed in percent per month.

Table 7: Performance of Portfolios Sorted by ML Predicted Returns on the CZ Sample

Method	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+Mom		Q	
	Avg	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
Equal Weight															
BRT	5.18	9.91	3.68	5.32	9.94	5.28	10.33	5.00	10.37	5.06	9.69	4.87	10.00	4.91	9.31
NN1	3.72	7.63	2.82	3.83	7.63	3.82	8.00	3.69	7.56	3.60	7.69	3.53	7.49	3.59	7.26
NN2	3.26	6.32	2.23	3.41	6.36	3.40	6.57	3.33	6.25	3.15	6.23	3.12	6.09	3.21	6.04
NN3	3.18	6.22	2.32	3.36	6.44	3.31	6.72	3.24	6.47	3.03	6.59	3.00	6.45	3.14	6.44
NN4	3.00	6.07	2.32	3.14	6.19	3.11	6.35	3.00	6.00	2.87	6.11	2.81	5.96	2.89	5.84
NN5	1.63	5.14	1.23	1.68	5.00	1.77	5.54	1.72	5.29	1.93	6.43	1.88	6.08	1.93	6.18
Value Weight															
BRT	2.32	6.99	1.28	2.58	7.85	2.58	7.41	1.97	7.49	2.31	4.99	1.89	6.11	2.00	4.32
NN1	2.27	5.93	1.38	2.41	6.16	2.40	6.36	1.98	5.72	2.15	5.18	1.86	5.29	1.96	4.61
NN2	1.72	5.08	1.03	1.89	5.13	1.87	5.34	1.47	4.65	1.56	4.21	1.29	4.02	1.37	3.93
NN3	1.80	5.21	1.14	2.03	5.62	2.01	6.08	1.63	5.81	1.70	5.07	1.45	5.02	1.52	4.61
NN4	1.86	4.49	1.17	2.02	4.75	2.01	5.03	1.70	4.46	1.74	4.43	1.54	4.23	1.67	3.82
NN5	0.93	3.62	0.65	0.99	3.27	1.06	3.75	0.79	2.75	1.02	4.16	0.83	3.32	0.87	3.45

This table reports the returns and risk-adjusted performance for the long-short portfolios sorted by ML-predicted returns on the CZ sample of signals from 1987 to 2019. We predict stock monthly excess returns using the March 2022 data release from openassetpricing.com/data, which contains the 207 signals collected by Chen and Zimmermann (2022). We use a recursive window approach and select the optimal hyper-parameters using a cross-validation approach. Our initial estimation period is 1963-1986. The first 12 years is the training period and the second 12 years is the validation period. As we roll forward, the training period expands while the validation period stays at 12 years. The risk-adjusted performance are calculated based on the CAPM model, the Fama-French 3-factor model, the Carhart 4-factor model, the Fama-French 5-factor model, the Fama-French 5-factor model augmented with momentum factor, and the q-factor model. The top panel reports results for equal-weighted portfolios. The bottom reports results for value-weighted portfolios. All returns are expressed in percent per month.

Table 8: Performance of ML Portfolios on the GHZ Sample with and without Short-term Reversal

Method	GHZ94			GHZ93		
	Ret.	*t-stat*	SR	*Ret.*	*t-stat*	SR
			Equal Weight			
BRT	3.57	8.91	2.35	3.04	9.11	1.80
NN1	3.78	9.74	2.69	3.22	9.99	2.46
NN2	3.87	9.93	2.95	3.37	10.68	2.30
NN3	3.94	10.09	2.84	3.27	10.57	2.48
NN4	2.94	9.59	2.27	2.30	8.91	1.84
NN5	1.70	5.92	1.33	0.72	3.65	0.62
			Value Weight			
BRT	1.51	4.53	0.71	1.11	3.67	0.52
NN1	1.68	4.84	0.85	1.42	4.42	0.77
NN2	1.87	5.18	0.99	1.60	5.75	0.85
NN3	1.71	4.93	0.90	1.50	4.34	0.84
NN4	1.37	6.32	0.82	0.97	3.45	0.58
NN5	0.69	3.78	0.49	0.30	1.46	0.22

This table reports the long-short returns for the portfolios sorted by ML predicted returns on the GHZ sample of signals with and without short-term reversal from 1987 to 2019. We predict stock monthly excess returns using signals from Green, Hand, and Zhang (2017). GHZ94 denotes the original 94 signals used by Gu, Kelly, and Xiu (2020), and GHZ93 denotes the 93 signals excluding short-term reversal (or *mom1m*). We use a recursive window approach and select the optimal hyper-parameters using a cross-validation approach. Our initial estimation period is 1963-1986. The first 12 years is the training period and the second 12 years is the validation period. As we roll forward, the training period expands while the validation period stays at 12 years. The top panel reports results for equal-weighted portfolios. The bottom reports results for value-weighted portfolios. All returns are expressed in percent per month.

Table 9: Performance of ML Portfolios on the CZ Sample with and without Short-term Reversal

Method	CZ207 *Ret.*	CZ207 *t-stat*	CZ207 SR	CZ206 *Ret.*	CZ206 *t-stat*	CZ206 SR
			Equal Weight			
BRT	5.18	9.91	3.68	4.87	10.05	3.21
NN1	3.72	7.63	2.82	3.32	7.58	2.55
NN2	3.26	6.32	2.23	3.21	6.39	2.27
NN3	3.18	6.22	2.32	3.20	6.30	2.28
NN4	3.00	6.07	2.32	2.86	5.83	1.99
NN5	1.63	5.14	1.23	1.46	4.86	1.20
			Value Weight			
BRT	2.32	6.99	1.28	2.14	6.26	1.03
NN1	2.27	5.93	1.38	2.02	5.09	1.26
NN2	1.72	5.08	1.03	1.96	5.39	1.09
NN3	1.80	5.21	1.14	1.82	4.77	1.12
NN4	1.86	4.49	1.17	2.07	4.56	1.20
NN5	0.93	3.62	0.65	1.01	4.49	0.77

This table reports the long-short returns for the portfolios sorted by ML-predicted returns on the CZ sample of signals with and without short-term reversal from 1987 to 2019. We use the March 2022 release from `openassetpricing.com/data`, which contains the 207 signals collected by Chen and Zimmermann (2022). CZ207 denotes the 207 signals, and CZ206 denotes the 206 signals after excluding short-term reversal (*Streversal*). We use a recursive window approach and select the optimal hyperparameters using a cross-validation approach. Our initial estimation period is 1963-1986. The first 12 years is the training period and the second 12 years is the validation period. As we roll forward, the training period expands while the validation period stays at 12 years. The top panel reports results for equal-weighted portfolios. The bottom reports results for value-weighted portfolios. All returns are expressed in percent per month.

Table 10: Performance of ML Portfolios on Past-return Signals with and without Short-term Reversal

Method	TECH120			TECH119		
	Ret.	t-stat	SR	Ret.	t-stat	SR
			Equal Weighted			
BRT	1.81	6.40	1.77	1.38	4.93	1.04
NN1	1.61	7.30	1.98	1.17	6.61	1.33
NN2	1.75	8.82	2.28	1.40	7.46	1.45
NN3	1.52	7.99	1.94	1.30	6.34	1.42
NN4	0.98	6.80	1.21	0.74	5.64	0.81
NN5	0.66	3.73	0.84	0.10	0.50	0.11
			Value Weighted			
BRT	0.98	3.14	0.66	0.78	2.41	0.46
NN1	0.72	2.90	0.56	0.87	4.29	0.72
NN2	1.26	4.87	1.02	1.06	4.34	0.78
NN3	1.07	4.10	0.80	1.00	3.79	0.80
NN4	0.92	4.19	0.78	0.38	1.68	0.31
NN5	0.50	2.45	0.43	0.32	1.30	0.26

This table reports the long-short returns of the portfolios sorted by BRT and NN predicted returns on past-return signals with and without short-term reversal from 1987 to 2019. We predict stock monthly excess returns using past-return signals. TECH120 denotes the previous 120 months return signals, and TECH119 denotes the previous 120 months return signals, excluding the most recent month. Our full sample period is 1963-2019 and the out-of-sample period is 1987-2019, consistent with our fundamental signals analysis. We use a recursive window approach and select the optimal hyper-parameters using a cross-validation approach. Our initial estimation period is 1963-1986. The first 12 years is the training period and the second 12 years is the validation period. As we roll forward, the training period expands while the validation period stays at 12 years. The top panel reports results for equal-weighted portfolios. The bottom reports results for value-weighted portfolios. All returns are expressed in percent per month.

Table 11: Performance of ML Portfolios after Trading Cost – Fundamental Signals

Methods	a Gross Return	b Turnover (2-Sided)	c Ave. Spread Paid	d≈ b× c Return Reduction	e=a-d Net Return
			Equal Weight		
BRT	97	14	177	24	73
NN1	67	14	174	25	43
NN2	85	14	187	27	58
NN3	30	14	178	26	4
NN4	38	14	185	24	14
NN5	7	12	177	21	-14
			Value Weight		
BRT	35	14	69	10	25
NN1	2	15	64	10	-8
NN2	-1	15	69	10	-11
NN3	2	15	68	10	-9
NN4	26	14	69	9	17
NN5	18	12	70	8	10

This table presents the gross returns, turnover, trading cost, and net returns for long-short portfolios sorted by BRT and NN predicted returns based on fundamental signals from 1987 to 2017. We predict stock annual excess returns using 18,113 fundamental signals (as described in Section 2.1). We use a recursive window approach and select the optimal hyper-parameters using a cross-validation approach. Our out-of-sample period ends in 2017 because the low-frequency effective spreads data are only available until 2017 on Andrew Chen's website. Column a denotes the gross return before trading cost. Column b is turnover (2-sided), calculated following Chen and Velikov (2022). In particular, we calculate the actual weight at the end of each month and calculate the turnover as the sum of absolute deviation from the next weight and its actual weight. We then take the average of long and short legs and average over the time series. Column d denotes the return reduction or the incurred trading cost, calculated as the sum of absolute deviation multiplied by low-frequency spreads. Column c is the average spread paid, calculated as the return reduction divided by the turnover. Column e reports the net return after trading cost. All results except column b are expressed in basis points per month. Column b (or turnover (2-sided)) is expressed in percent per month.

Table 12: Performance of ML Portfolios after Trading Cost – Past-Return Signals

	Panle A: TECH119					Panel B: TECH120				
Methods	a	b	c	d≈ b× c	e=a-d	a	b	c	d≈ b× c	e=a-d
					Equal Weight					
BRT	141	124	190	238	-97	190	144	232	338	-148
NN1	123	153	201	305	-182	167	156	212	330	-163
NN2	146	149	206	309	-163	182	155	219	340	-158
NN3	136	152	210	318	-182	157	156	216	338	-181
NN4	76	153	205	317	-241	104	160	213	340	-236
NN5	9	106	210	221	-212	69	130	213	277	-207
					Value Weight					
BRT	72	131	77	101	-29	95	155	87	136	-40
NN1	86	171	65	111	-24	68	173	69	119	-50
NN2	106	173	67	115	-9	127	175	71	124	3
NN3	102	174	68	119	-16	106	175	70	123	-17
NN4	36	165	70	115	-78	88	174	71	123	-34
NN5	33	108	77	76	-44	49	136	74	98	-49

This table shows the gross return, turnover, trading cost, and net return for the long-short portfolios sorted by BRT and NN predicted returns based on past-return signals from 1987 to 2017. We predict stock monthly excess returns using past-return signals. TECH120 denotes the previous 120 months return signals, and TECH119 denotes the the previous 120 months return signals, excluding the most recent month. Our out-of-sample period ends in 2017 because the low-frequency effective spreads data are only available until 2017 on Andrew Chen's website. Column a denotes the gross return before trading cost. Column b is turnover (2-sided), calculated following Chen and Velikov (2022). In particular, we calculate the actual weight at the end of each month and calculate the turnover as the sum of absolute deviation from the next weight and its actual weight. We then take the average of long and short legs and average over the time series. Column d denotes the return reduction or the incurred trading cost, calculated as the sum of absolute deviation multiplied by low-frequency spreads. Column c is the average spread paid, calculated as the return reduction divided by the turnover. Column e reports the net return after trading cost. All results except column b are expressed in basis points per month. Column b (or turnover (2-sided)) is expressed in percent per month.

Appendix A: Grids of Hyper-Parameters for Cross Validation

BRT	NN
# of iteration $\in \{100, 250, 500, 750, 1000\}$	L1 penalty $\lambda_1 \in \{10^{-5}, 10^{-3}\}$
learning rate $\in \{0.01, 0.05, 0.1\}$	Learning Rate LR$\in \{0.001, 0.01\}$
	Batch Size = 10000
	Epochs = 100
	Patience = 5
	Adam Para. = Default

This table shows the grids of hyper-parameters used in the cross validation of Boosted Regression Trees (BRT) and Neural Networks (NN). We follow Gu, Kelly, and Xiu (2020) to select the grids of hyper-parameters.

Appendix B: List of Accounting Variables

#	*Variable*	*Description*	*Missing Rate*	*Start Year*
1	ACCHG	Accounting changes - cumulative effect	39.29%	1988
2	ACO	Current assets other total	0.76%	1963
3	ACOX	Current assets other sundry	2.20%	1963
4	ACT	Current assets - total	2.13%	1963
5	AM	Amortization of intangibles	33.03%	1965
6	AO	Assets – other	0.06%	1963
7	AOLOCH	Assets and liabilities other net change	38.36%	1988
8	AOX	Assets – other - sundry	2.22%	1963
9	AP	Accounts payable – trade	4.88%	1963
10	APALCH	Accounts payable & accrued liabilities increase/decrease	53.14%	1988
11	AQC	Acquisitions	12.98%	1972
12	AQI	Acquisitions income contribution	32.50%	1975
13	AQS	Acquisitions sales contribution	32.26%	1975
14	AT	Assets – total	0.01%	1963
15	BAST	Average short-term borrowing	74.28%	1978
16	CAPS	Capital surplus/share premium reserve	2.08%	1963
17	CAPX	Capital expenditure	2.18%	1963
18	CAPXV	Capital expenditure PPE Schedule V	1.39%	1963
19	CEQ	Common/ordinary equity - total	1.54%	1963
20	CEQL	Common equity liquidation value	1.62%	1963
21	CEQT	Common equity tangible	1.64%	1963
22	CH	Cash	12.33%	1963
23	CHE	Cash and short-term investments	0.72%	1963
24	CHECH	Cash and cash equivalents increase/decrease	28.77%	1972
25	CLD2	Capitalized leases - due in 2nd year	46.55%	1985
26	CLD3	Capitalized leases - due in 3rdyear	46.44%	1985
27	CLD4	Capitalized leases - due in 4thyear	46.18%	1985
28	CLD5	Capitalized leases - due in 5thyear	46.15%	1985
29	COGS	Cost of goods sold	0.09%	1963
30	CSTK	Common/ordinary stock (capital)	1.96%	1963
31	CSTKCV	Common stock-carrying value	28.31%	1963
32	CSTKE	Common stock equivalents – dollar savings	0.06%	1963
33	DC	Deferred charges	28.45%	1965
34	DCLO	Debt capitalized lease obligations	10.08%	1965
35	DCOM	Deferred compensation	72.02%	1980
36	DCPSTK	Convertible debt and stock	2.85%	1963
37	DCVSR	Debt senior convertible	9.89%	1970
38	DCVSUB	Debt subordinated convertible	11.96%	1970
39	DCVT	Debt – convertible	5.80%	1963
40	DD	Debt debentures	10.55%	1965
41	DD1	Long-term debt due in one year	5.05%	1963
42	DD2	Debt Due in 2nd Year	23.27%	1974
43	DD3	Debt Due in 3rd Year	23.32%	1974
44	DD4	Debt Due in 4th Year	23.16%	1974
45	DD5	Debt Due in 5th Year	24.04%	1974
46	DFS	Debt finance subsidiary	79.68%	1992
47	DFXA	Depreciation of tangible fixed assets	65.07%	1970
48	DILADJ	Dilution adjustment	62.54%	1994
49	DILAVX	Dilution available excluding extraordinary items	62.54%	1994
50	DLC	Debt in current liabilities - total	0.72%	1963
51	DLCCH	Current debt changes	60.86%	1974
52	DLTIS	Long-term debt issuance	10.50%	1972

#	*Variable*	*Description*	*Missing Rate*	*Start Year*
53	DLTO	Other long-term debt	9.96%	1965
54	DLTP	Long-term debt tied to prime	38.66%	1975
55	DLTR	Long-term debt reduction	9.84%	1972
56	DLTT	Long-term debt - total	0.20%	1963
57	DM	Debt mortgages &other secured	33.76%	1981
58	DN	Debt notes	10.56%	1965
59	DO	Income (loss) from discontinued operations	3.66%	1963
60	DONR	Nonrecurring discontinued operations	71.10%	1994
61	DP	Depreciation and amortization	0.24%	1963
62	DPACT	Depreciation , depletion and amortization	0.44%	1963
63	DPC	Depreciation and amortization (cash flow)	8.59%	1972
64	DPVIEB	Depreciation ending balance (schedule VI)	19.34%	1970
65	DPVIO	Depreciation other changes (schedule VI)	65.12%	1970
66	DPVIR	Depreciation retirements (schedule VI)	65.14%	1970
67	DRC	Deferred revenue current	73.42%	1994
68	DS	Debt-subordinated	9.93%	1965
69	DUDD	Debt unamortized debt discount and other	29.51%	1963
70	DV	Cash dividends (cash flow)	8.55%	1972
71	DVC	Dividends common/ordinary	0.11%	1963
72	DVP	Dividends - preferred/preference	0.06%	1963
73	DVPA	Preferred dividends in arrears	17.95%	1964
74	DVPIBB	Depreciation beginning balance (schedule VI)	60.82%	1970
75	DVT	Dividends – total	0.11%	1963
76	DXD2	Debt (excl capitalized leases) due in 2nd year	49.31%	1985
77	DXD3	Debt (excl capitalized leases) due in 3rd year	49.25%	1985
78	DXD4	Debt (excl capitalized leases) due in 4thyear	48.96%	1985
79	DXD5	Debt (excl capitalized leases) due in 5thyear	49.36%	1985
80	EBIT	Earnings before interest and taxes	1.36%	1963
81	EBITDA	Earnings before interest	0.21%	1963
82	ESOPCT	ESOP obligation (common) - total	40.69%	1980
83	ESOPDLT	ESOP debt - long term	49.09%	1990
84	ESOPT	Preferred ESOP obligation - total	41.01%	1964
85	ESUB	Equity in earnings -unconsolidated subsidiaries	12.33%	1963
86	ESUBC	Equity in net loss earnings	22.05%	1972
87	EXRE	Exchange rate effect	38.46%	1988
88	FATB	Property, plant, and equipment buildings	51.33%	1985
89	FATC	Property, plant and equipment construction in progress	47.36%	1985
90	FATE	Property, plant, equipment and machinery equipment	53.32%	1985
91	FATL	Property, plant, and equipment leases	57.58%	1985
92	FATN	Property, plant, equipment and natural resources	47.37%	1985
93	FATO	Property, plant, and equipment other	52.84%	1985
94	FATP	Property, plant, equipment and land improvements	51.25%	1985
95	FIAO	Financing activities other	38.35%	1988
96	FINCF	Financing activities net cash flow	38.35%	1988
97	FOPO	Funds from operations other	7.83%	1972
98	FOPOX	Funds from operations - Other excl option tax benefit	76.37%	1992
99	FOPT	Funds from operations total	69.42%	1972
100	FSRCO	Sources of funds other	70.81%	1972
101	FSRCT	Sources of funds total	71.27%	1972
102	FUSEO	Uses of funds other	70.81%	1972
103	FUSET	Uses of funds total	71.61%	1972
104	GDWL	Goodwill	47.13%	1989
105	GP	Gross profit (loss)	0.09%	1963
106	IB	Income before extraordinary items	0.05%	1963

#	*Variable*	*Description*	*Missing Rate*	*Start Year*
107	IBADJ	IB adjusted for common stock equivalents	0.05%	1963
108	IBC	Income before extraordinary items (cash flow)	7.82%	1972
109	IBCOM	Income before extraordinary items available for common	0.05%	1963
110	ICAPT	Invested capital – total	1.54%	1963
111	IDIT	Interest and related income - total	42.18%	1965
112	INTAN	Intangible assets – total	10.02%	1963
113	INTC	Interest capitalized	16.78%	1963
114	INTPN	Interest paid net	43.82%	1988
115	INVCH	Inventory decrease (increase)	43.46%	1988
116	INVFG	Inventories finished goods	41.28%	1970
117	INVO	Inventories other	52.52%	1984
118	INVRM	Inventories raw materials	40.27%	1969
119	INVT	Inventories – total	1.43%	1963
120	INVWIP	Inventories work in progress	43.69%	1970
121	ITCB	Investment tax credit (balance sheet)	3.20%	1963
122	ITCI	Investment tax credit (income account)	37.65%	1963
123	IVACO	Investing activities other	38.35%	1988
124	IVAEQ	Investment and advances – equity	9.07%	1963
125	IVAO	Investment and advances other	7.07%	1963
126	IVCH	Increase in investments	13.68%	1972
127	IVNCF	Investing activities net cash flow	38.35%	1988
128	IVST	Short-term investments – total	12.35%	1963
129	IVSTCH	Short-term investments change	48.38%	1988
130	LCO	Current liabilities other total	4.76%	1963
131	LCOX	Current liabilities other sundry	6.10%	1963
132	LCOXDR	Current liabilities-other-excl deferred revenue	72.40%	1994
133	LCT	Current liabilities – total	1.69%	1963
134	LIFR	LIFO reserve	22.04%	1976
135	LO	Liabilities – other – total	0.72%	1963
136	LT	Liabilities – total	0.50%	1963
137	MIB	Minority interest (balance sheet)	6.37%	1963
138	MII	Minority interest (income account)	10.24%	1963
139	MRC1	Rental commitments minimum 1styear	27.85%	1975
140	MRC2	Rental commitments minimum 2ndyear	28.34%	1975
141	MRC3	Rental commitments minimum 3rdyear	28.46%	1975
142	MRC4	Rental commitments minimum 4th year	28.61%	1975
143	MRC5	Rental commitments minimum 5th year	30.38%	1975
144	MRCT	Rental commitments minimum 5 year total	29.51%	1975
145	MSA	Marketable securities adjustment	18.18%	1976
146	NI	Net income (loss)	0.06%	1963
147	NIADJ	Net income adjusted for common stock equiv.	2.24%	1963
148	NIECI	Net income effect capitalized interest	59.92%	1976
149	NOPI	Non-operating income (expense)	0.10%	1963
150	NOPIO	Non-operating income (expense) other	0.10%	1963
151	NP	Notes payable short-term borrowings	0.80%	1963
152	OANCF	Operating activities net cash flow	38.36%	1988
153	OB	Order backlog	64.22%	1971
154	OIADP	Operating income after depreciation	0.07%	1963
155	PI	Pre-tax income	0.06%	1963
156	PIDOM	Pretax income domestic	74.94%	1981
157	PIFO	Pretax income foreign	75.36%	1981
158	PPEGT	Property, plant, and equipment – total (gross)	0.45%	1963
159	PPENB	Property, plant, and equipment buildings (net)	70.38%	1970
160	PPENC	Property plant equipment construction in progress (net)	65.66%	1970

#	*Variable*	*Description*	*Missing Rate*	*Start Year*
161	PPENLI	Property plant equipment land and improvements (net)	70.26%	1970
162	PPENME	Property plant equipment machinery and equipment (net)	69.73%	1970
163	PPENNR	Property plant equipment natural resources (net)	69.31%	1970
164	PPENO	Property plant and equipment other (net)	69.26%	1970
165	PPENT	Property, plant, and equipment – total (net)	0.11%	1963
166	PPEVBB	Property plant equipment beginning balance (schedule V)	57.03%	1963
167	PPEVEB	Property, plant, and equipment ending balance	15.25%	1963
168	PPEVO	Property, plant, and equipment other changes (schedule V)	62.50%	1963
169	PPEVR	Property, plant and equipment retirements (schedule V)	62.50%	1963
170	PRSTKC	Purchase of common and preferred stock	12.98%	1972
171	PSTK	Preferred/preference stock (capital) – total	0.24%	1963
172	PSTKC	Preferred stock convertible	4.96%	1963
173	PSTKL	Preferred stock liquidating value	0.05%	1963
174	PSTKN	Preferred/preference stock – non-redeemable	1.48%	1963
175	PSTKR	Preferred/preference stock - redeemable	20.89%	1964
176	PSTKRV	Preferred stock redemption value	0.06%	1963
177	RDIP	In process R&D expense	65.68%	1994
178	RE	Retained earnings	2.04%	1963
179	REA	Retained earnings restatement	10.33%	1970
180	REAJO	Retained earnings other adjustments	30.06%	1983
181	RECCH	Accounts receivable decrease (increase)	41.58%	1988
182	RECCO	Receivables – current – other	3.21%	1963
183	RECD	Receivables – estimated doubtful	29.03%	1970
184	RECT	Receivables – total	1.45%	1963
185	RECTA	Retained earnings cumulative translation adjustment	30.39%	1983
186	RECTR	Receivables – trade	17.96%	1967
187	REUNA	Retained earnings unadjusted	29.89%	1983
188	SALE	Sales/turnover (net)	0.05%	1963
189	SEQ	Stockholders' equity – total	2.24%	1963
190	SIV	Sale of investments	16.24%	1972
191	SPI	Special items	3.93%	1963
192	SPPE	Sale of property	28.92%	1972
193	SPPIV	Sale of property plant equipment investments gain (loss)	38.36%	1988
194	SSTK	Sale of common and preferred stock	9.55%	1972
195	TLCF	Tax loss carry forward	23.48%	1963
196	TSTK	Treasury stock – total (all capital)	16.37%	1970
197	TSTKC	Treasury stock – common	26.38%	1974
198	TSTKP	Treasury stock – preferred	25.51%	1963
199	TXACH	Income taxes accrued increase/decrease	56.69%	1988
200	TXBCO	Excess tax benefit stock options -cash flow	76.43%	1992
201	TXC	Income tax – current	16.78%	1963
202	TXDB	Deferred taxes (balance sheet)	3.34%	1963
203	TXDBA	Deferred tax asset - long term	73.84%	1993
204	TXDBCA	Deferred tax asset - current	73.11%	1994
205	TXDBCL	Deferred tax liability - current	74.46%	1994
206	TXDC	Deferred taxes (cash flow)	10.38%	1972
207	TXDFED	Deferred taxes-federal	48.37%	1985
208	TXDFO	Deferred taxes-foreign	45.98%	1985
209	TXDI	Income tax – deferred	6.99%	1963
210	TXDITC	Deferred taxes and investment tax credit	3.34%	1963
211	TXDS	Deferred taxes-state	48.91%	1985
212	TXFED	Income tax federal	16.78%	1963
213	TXFO	Income tax foreign	19.02%	1970
214	TXNDB	Net deferred tax asset (liab) - total	69.95%	1994

#	*Variable*	*Description*	*Missing Rate*	*Start Year*
215	TXNDBA	Net deferred tax asset	72.66%	1994
216	TXNDBL	Net deferred tax liability	72.67%	1994
217	TXNDBR	Deferred tax residual	72.05%	1994
218	TXO	Income taxes - other	33.11%	1963
219	TXP	Income tax payable	5.93%	1963
220	TXPD	Income taxes paid	45.36%	1988
221	TXR	Income tax refund	10.40%	1963
222	TXS	Income tax state	17.76%	1963
223	TXT	Income tax total	0.06%	1963
224	TXW	Excise taxes	24.39%	1976
225	WCAP	Working capital (balance sheet)	2.15%	1963
226	WCAPC	Working capital change other increase/decrease	72.51%	1972
227	WCAPCH	Working capital change total	74.62%	1972
228	XACC	Accrued expenses	19.16%	1963
229	XAD	Advertising expense	64.98%	1963
230	XDEPL	Depletion expense (schedule VI)	68.80%	1970
231	XI	Extraordinary items	1.60%	1963
232	XIDO	Extra. items and discontinued operations	0.06%	1963
233	XIDOC	Extra. items and disc. operations (cash flow)	9.44%	1972
234	XINT	Interest and related expenses – total	5.05%	1963
235	XOPR	Operating expenses – total	0.09%	1963
236	XPP	Prepaid expenses	43.96%	1963
237	XPR	Pension and retirement expense	25.03%	1963
238	XRD	Research and development expense	47.01%	1963
239	XRENT	Rental expense	14.34%	1963
240	XSGA	Selling, general and administrative expense	12.13%	1963

This table lists the 240 accounting variables used in this study and their descriptions. Our sample period is 1963-2019. We begin with all accounting variables on the balance sheet, income statement, and cash flow statement included in the annual Compustat database. We exclude all variables with fewer than 20 years of data or fewer than 1,000 firms with non-missing data on average per year. We exclude per-share-based variables such as book value per share and earnings per share. We remove LSE (total liabilities and equity), REVT (total revenue), OIBDP (operating income before depreciation), and XDP (depreciation expense) because they are identical to TA (total assets), SALE (total sale), EBITDA (earnings before interest) and DFXA (depreciation of tangible fixed assets) respectively. Please refer to Yan and Zheng (2017) for more details.

Appendix C: List of Financial Ratios and Configurations

#	*Description*	#	*Description*	#	*Description*	#	*Description*	#	*Description*
1	X/AT	16	Δ in X/AT	31	%Δ in X/AT	46	ΔX/LAGAT	61	%Δ in X - %Δ in AT
2	X/ACT	17	Δ in X/ACT	32	%Δ in X/ACT	47	ΔX/LAGACT	62	%Δ in X - %Δ in ACT
3	X/INVT	18	Δ in X/INVT	33	%Δ in X/INVT	48	ΔX/LAGINVT	63	%Δ in X - %Δ in INVT
4	X/PPENT	19	Δ in X/PPENT	34	%Δ in X/PPENT	49	ΔX/LAGPPENT	64	%Δ in X - %Δ in PPENT
5	X/LT	20	Δ in X/LT	35	%Δ in X/LT	50	ΔX/LAGLT	65	%Δ in X - %Δ in LT
6	X/LCT	21	Δ in X/LCT	36	%Δ in X/LCT	51	ΔX/LAGLCT	66	%Δ in X - %Δ in LCT
7	X/DLTT	22	Δ in X/DLTT	37	%Δ in X/DLTT	52	ΔX/LAGDLTT	67	%Δ in X - %Δ in DLTT
8	X/CEQ	23	Δ in X/CEQ	38	%Δ in X/CEQ	53	ΔX/LAGCEQ	68	%Δ in X - %Δ in CEQ
9	X/SEQ	24	Δ in X/SEQ	39	%Δ in X/SEQ	54	ΔX/LAGSEQ	69	%Δ in X - %Δ in SEQ
10	X/ICAPT	25	Δ in X/ICAPT	40	%Δ in X/ICAPT	55	ΔX/LAGICAPT	70	%Δ in X - %Δ in ICAPT
11	X/SALE	26	Δ in X/SALE	41	%Δ in X/SALE	56	ΔX/LAGSALE	71	%Δ in X - %Δ in SALE
12	X/COGS	27	Δ in X/COGS	42	%Δ in X/COGS	57	ΔX/LAGCOGS	72	%Δ in X - %Δ in COGS
13	X/XSGA	28	Δ in X/XSGA	43	%Δ in X/XSGA	58	ΔX/LAGXSGA	73	%Δ in X - %Δ in XSGA
14	X/EMP	29	Δ in X/EMP	44	%Δ in X/EMP	59	ΔX/LAGEMP	74	%Δ in X - %Δ in EMP
15	X/MKTCAP	30	Δ in X/MKTCAP	45	%Δ in X/MKTCAP	60	ΔX/LAGMKTCAP	75	%Δ in X - %Δ in MKTCAP
								76	%Δ in X

This table lists the 76 financial ratios and configurations used in this study. Our sample period is 1963-2019. We begin with all accounting variables on the balance sheet, income statement, and cash flow statement included in the annual Compustat database. We exclude all variables with fewer than 20 years of data or fewer than 1,000 firms with non-missing data on average per year. We exclude per-share-based variables such as book value per share and earnings per share. "X" represents the 240 accounting variables listed in Appendix B. "Y" represents the fifteen base variables, including AT (total assets), ACT (total current assets), INVT (inventory), PPENT (property, plant, and equipment), LT (total liabilities), LCT (total current liabilities), DLTT (long-term debt), CEQ (total common equity), SEQ (stockholders' equity), ICAPT (total invested capital), SALE (total sale), COGS (cost of goods sold), XSGA (selling, general, and administrative cost), EMP (number of employees), and MKTCAP (market capitalization). Please refer to Yan and Zheng (2017) for more details.

Appendix D. Relative Influence Measures

One criticism of machine learning algorithms is that they are "Black Boxes" that do not provide a lot of intuition to the researcher and the reader. This criticism hardly applies to BRTs that feature very useful and intuitive visualization tools. The first commonly used measure is referred to as the "relative influence" measure. Consider the reduction in the empirical error every time one of the covariates $x_{l,\cdot}$ is used to split the tree. Summing the reductions in empirical errors (or improvements in fit) across the nodes in the tree gives a measure of the variable's influence (Breiman, Friedman, Stone, and Olshen, 1984):

$$I_l(\mathcal{T}) = \sum_{j=2}^{J} \Delta e(j)^2 \, I(x(j) = l),$$

where $\Delta e(j)^2 = T^{-1} \sum_{t=1}^{T} \left(e_t(j-1)^2 - e_t(j)^2 \right)$ is the reduction in the squared empirical error at the j^{th} node and $x(j)$ is the regressor chosen at this node, so $I(x(j) = l)$ equals 1 if regressor l is chosen, and 0 otherwise. The sum is computed across all observations, $t = 1, \ldots, T$, and over the $J-1$ internal nodes of the tree.

The rationale for this measure is that at each node, one of the regressors gets selected to partition the sample space into two sub-states. The particular regressor at node j achieves the greatest reduction in the empirical risk of the model fitted up to node $j-1$. The importance of each regressor, $x_{l,\cdot}$, is the sum of the reductions in the empirical errors computed over all internal nodes for which it was chosen as the splitting variable. If a regressor never gets chosen to conduct the splits, its influence is zero. Conversely, the more frequently a regressor is used for splitting, and the bigger its effect on reducing the model's empirical risk, the larger its influence.

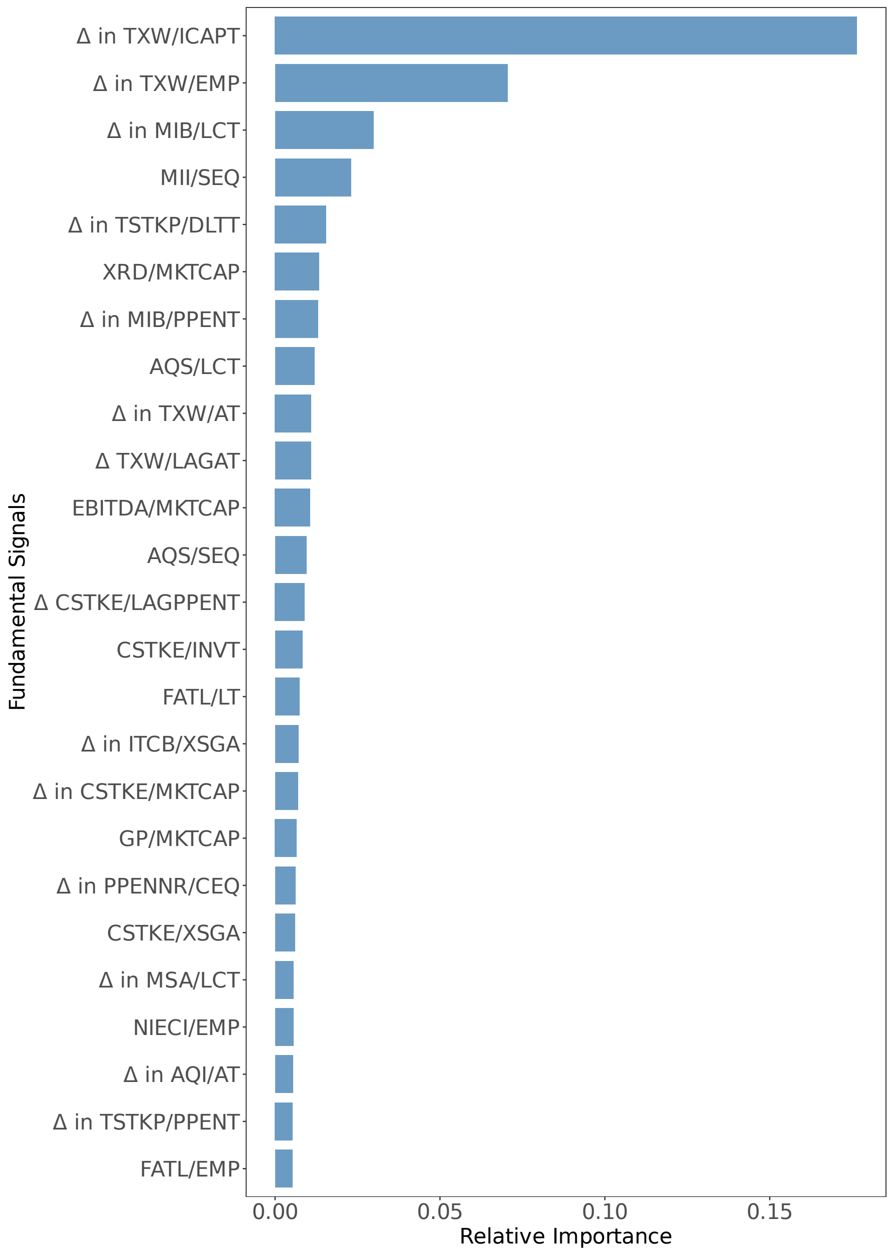
This measure of influence can be generalized by averaging over the number of boosting iterations, B, which generally provides a more reliable measure of influence:

$$\bar{I}_l = \frac{1}{B} \sum_{b=1}^{B} I_l(\mathcal{T}_b).$$

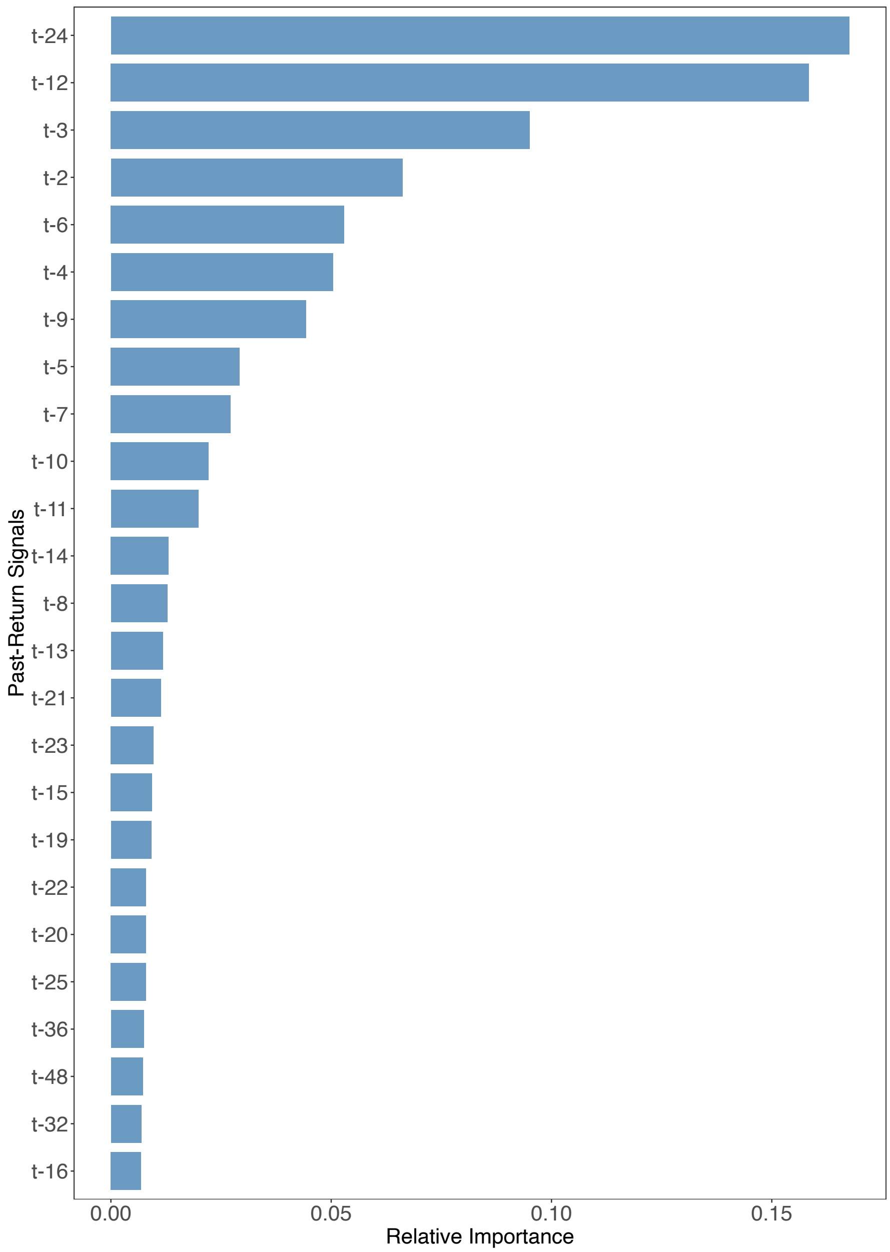
This is best interpreted as a measure of relative influence that can be compared across regressors. We therefore report the following measure of relative influence, $\overline{RI}_l$, which sums to 1:

$$\overline{RI}_l = \bar{I}_l / \sum_{l=1}^{L} \bar{I}_l.$$

The figure below shows the relative influence of the top 25 signals in the baseline BRT model estimated in the paper. We first compute the signals' relative influence in each year of the test period, 1987-2019, and average their values across all test years. Note that the relative importance measure across all signals sums to one every year. We then rank and plot the signals according to their average relative influence. The Y-axis reports the 25 most important signals, while the x-axis presents each signal's average relative influence measure.



The figure below shows the relative influence of the top 25 signals in the baseline BRT model on past return signals. We first compute the signals' relative influence in each month of the test period, 1987-2019, and average their values across all test months. Note that the relative importance measure across all signals sums to one every month. We then rank and plot the signals according to their average relative influence. The Y-axis reports the 25 most important signals in terms of lags, while the x-axis presents each signal's average relative influence measure.

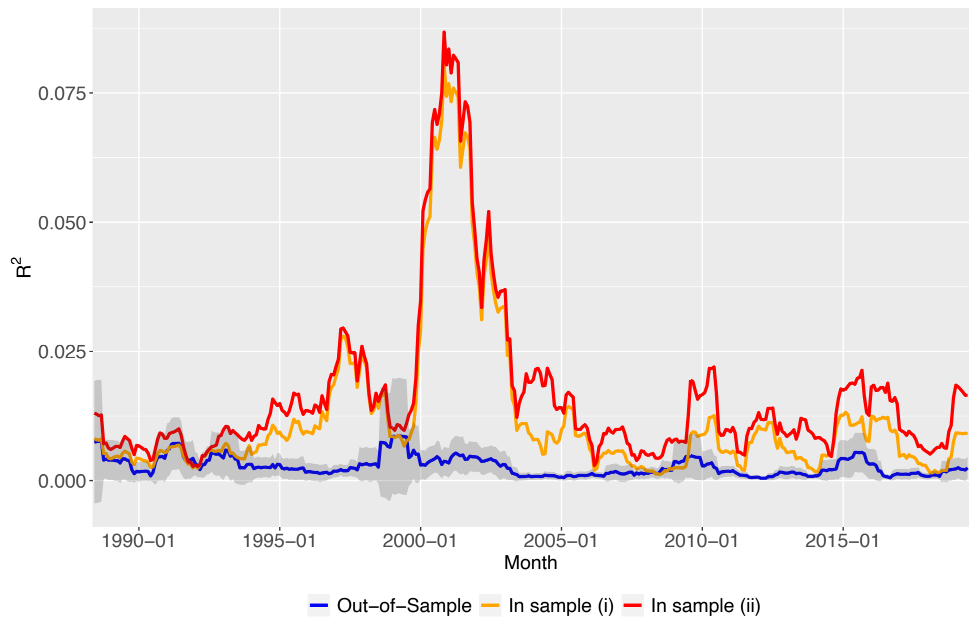


Internet Appendix
"Real-time Machine Learning in the Cross-Section of Stock Returns"

Submitted to *Journal of Financial Economics*
Manuscript # FINEC-D-22-00977

October 19, 2023

(a) In-Sample and Out-of-Sample R^2



(b) In-Sample and Out-of-Sample Long-Short Return

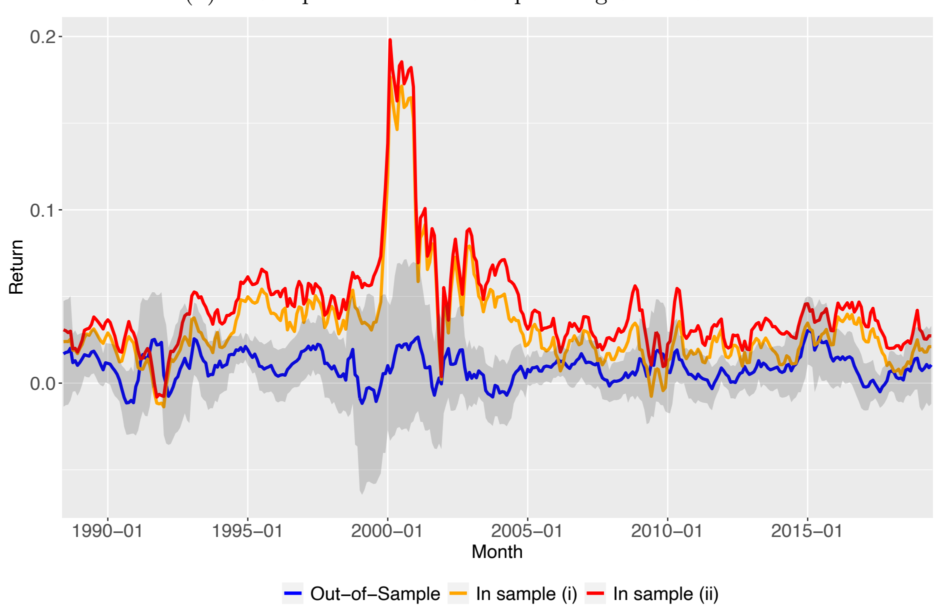


Figure IA.1: In- and Out-of-Sample Return Predictability

This figure shows the in- and out-of-sample BRT return predictability. Panel (a) plots the 12-month moving averages of in- and out-of-sample R^2. In-sample (i) denotes the in-sample results with optimal hyper-parameters chosen over the training period 1962 - 1986. For In-sample (ii), the optimal hyper-parameters are chosen using leave-one-year-out cross-validation over the test period 1987-2019, following Martin and Nagel (2022). Panel (b) shows the in- and out-of-sample returns of a long-short portfolio strategy based on BRT forecasts.

Appendix IA.1: Performance of Portfolios Sorted by BRT Predicted Returns – Rolling-window Approach

	Equal Weight															
	Returns			CAPM		FF3		Carhart		FF5		FF5+Mom		Q		
Rank	*Avg*	*t-stat*	SR	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	
1 (Low)	0.11	0.27	0.05	-0.70	-2.73	-0.63	-3.02	-0.59	-3.32	-0.33	-2.24	-0.33	-2.32	-0.21	-1.05	
2	0.59	1.79	0.31	-0.16	-0.75	-0.15	-1.12	-0.11	-0.86	0.00	-0.04	0.02	0.17	0.12	0.91	
3	0.71	2.18	0.40	0.00	0.00	0.00	-0.02	0.02	0.20	0.08	0.70	0.10	0.82	0.18	1.38	
4	0.77	2.59	0.45	0.07	0.35	0.05	0.55	0.04	0.53	0.07	0.81	0.07	0.80	0.16	1.72	
5	0.71	2.58	0.45	0.05	0.26	0.03	0.33	0.02	0.30	-0.02	-0.30	-0.02	-0.29	0.06	0.69	
6	0.83	3.15	0.55	0.18	1.08	0.14	2.11	0.15	2.23	0.09	1.29	0.10	1.43	0.13	1.67	
7	0.89	3.17	0.58	0.23	1.29	0.20	2.41	0.22	2.60	0.15	1.80	0.17	2.00	0.22	2.77	
8	0.85	2.84	0.55	0.18	1.11	0.16	1.69	0.22	2.61	0.15	1.62	0.20	2.17	0.25	2.23	
9	0.84	2.61	0.49	0.12	0.69	0.14	1.24	0.24	2.16	0.28	2.33	0.34	2.74	0.38	2.84	
10 (High)	0.94	2.47	0.50	0.20	0.83	0.27	1.59	0.33	2.03	0.53	3.08	0.55	3.21	0.61	3.55	
10-1	0.83	4.29	0.77	0.90	3.96	0.90	4.06	0.92	4.42	0.86	4.36	0.87	4.37	0.83	3.79	

	Value Weight														
	Returns			CAPM		FF3		Carhart		FF5		FF5+Mom		Q	
Rank	*Avg*	*t-stat*	SR	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
1 (Low)	0.41	1.15	0.20	-0.37	-1.62	-0.25	-1.06	-0.29	-1.53	-0.01	-0.07	-0.05	-0.34	0.05	0.20
2	0.57	1.79	0.33	-0.12	-0.59	-0.03	-0.16	0.00	-0.03	0.12	0.68	0.13	0.78	0.20	1.02
3	0.61	2.07	0.40	-0.04	-0.24	0.00	-0.01	-0.03	-0.17	0.04	0.38	0.02	0.20	0.05	0.34
4	0.64	2.65	0.42	-0.01	-0.10	0.01	0.10	-0.05	-0.50	0.00	0.03	-0.04	-0.38	0.01	0.10
5	0.58	2.56	0.45	-0.01	-0.18	0.01	0.08	0.05	0.66	-0.02	-0.21	0.02	0.22	0.03	0.31
6	0.78	3.33	0.56	0.15	1.41	0.14	1.53	0.15	1.64	0.01	0.17	0.03	0.33	0.07	0.79
7	0.49	1.81	0.35	-0.16	-1.97	-0.15	-1.74	-0.10	-1.32	-0.25	-2.65	-0.21	-2.44	-0.16	-1.54
8	0.68	2.90	0.52	0.10	1.00	0.10	0.91	0.15	1.67	-0.03	-0.26	0.02	0.20	0.08	0.70
9	0.55	1.80	0.38	-0.09	-0.58	-0.06	-0.41	-0.03	-0.25	-0.09	-0.74	-0.06	-0.60	-0.02	-0.15
10 (High)	0.74	2.18	0.44	0.03	0.18	0.12	0.85	0.22	1.62	0.28	2.17	0.35	2.57	0.39	2.94
10-1	0.33	1.35	0.21	0.40	1.35	0.37	1.30	0.51	2.10	0.29	1.28	0.40	1.77	0.34	1.30

This table shows the performance of BRT portfolios using rolling windows instead of recursive windows. All results are computed following the procedure described in Table 1 in the main paper, except that our estimation period is fixed at 24 years. In particular, we select the optimal hyper-parameters using the most recent 12 years as our validation period and the remaining 12 years as our training period. After obtaining the optimal hyper-parameters, we re-train our model using the rolling window of 24 years. The first three columns report average monthly returns for the decile portfolios and the associated Sharpe Ratios. The remaining columns report risk-adjusted returns—see Table 2 in the main paper for details. The top panel reports results for equal-weighted returns. The bottom panel reports results for value-weighted returns. All returns are expressed in percent per month.

Appendix IA.2: Performance of Portfolios Sorted by BRT Predicted Returns – Alternative Training and Validation Periods

Equal Weight															
Training	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+MOM		Q	
+Validation	*Avg*	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
12+12	0.95	6.63	1.02	1.01	6.30	1.01	6.35	1.08	6.43	1.03	5.42	1.08	5.60	0.98	5.11
10+10	1.02	6.68	1.11	1.08	6.54	1.09	6.18	1.08	6.57	1.22	5.92	1.20	5.96	1.13	5.53
12+10	1.01	6.37	1.10	1.06	6.26	1.08	6.01	1.09	6.40	1.20	5.81	1.19	5.87	1.12	5.42
14+10	1.01	5.99	1.09	1.05	5.87	1.08	5.67	1.09	5.92	1.18	5.47	1.18	5.50	1.11	5.17
14+12	0.89	6.07	0.95	0.92	5.86	0.92	5.94	0.98	6.03	0.89	5.22	0.94	5.34	0.84	4.93
16+12	0.90	6.24	0.97	0.96	6.20	0.97	6.71	1.02	6.63	0.95	5.74	0.99	5.66	0.90	5.57
18+12	0.87	5.94	0.95	0.93	5.95	0.94	6.45	0.99	6.44	0.92	5.66	0.96	5.65	0.87	5.61
14+14	1.01	6.19	1.12	1.08	6.63	1.13	6.78	1.10	7.44	1.19	5.91	1.17	6.01	1.12	5.93
16+14	1.01	5.88	1.11	1.07	6.32	1.11	6.50	1.08	7.08	1.17	5.73	1.15	5.83	1.09	5.91
18+14	0.93	5.28	1.00	0.99	5.73	1.03	5.86	1.00	6.42	1.09	5.21	1.07	5.31	1.02	5.41

Value Weight															
Training	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+MOM		Q	
+Validation	*Avg*	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
12+12	0.40	2.34	0.30	0.46	2.17	0.46	2.57	0.65	3.66	0.68	3.18	0.80	3.80	0.68	2.94
10+10	0.41	2.50	0.32	0.45	2.21	0.48	2.73	0.57	3.34	0.75	3.61	0.80	3.63	0.74	3.45
12+10	0.41	2.52	0.33	0.45	2.18	0.48	2.63	0.60	3.42	0.78	3.79	0.84	3.93	0.78	3.74
14+10	0.41	2.33	0.32	0.43	2.01	0.47	2.46	0.59	3.21	0.76	3.54	0.83	3.73	0.76	3.48
14+12	0.37	2.08	0.27	0.39	1.81	0.40	2.20	0.59	3.26	0.57	2.79	0.71	3.32	0.57	2.59
16+12	0.39	2.14	0.29	0.44	1.99	0.47	2.89	0.64	3.85	0.66	3.53	0.78	3.84	0.67	3.59
18+12	0.41	2.13	0.30	0.45	1.94	0.48	2.73	0.63	3.60	0.65	3.31	0.75	3.65	0.65	3.36
14+14	0.49	2.42	0.37	0.57	2.32	0.64	3.38	0.70	3.97	0.92	4.64	0.94	4.29	0.93	4.60
16+14	0.55	2.62	0.42	0.61	2.41	0.68	3.39	0.72	3.92	0.94	4.51	0.96	4.30	0.93	4.40
18+14	0.53	2.38	0.39	0.58	2.18	0.65	3.11	0.69	3.57	0.92	4.18	0.94	4.00	0.88	4.04

This table shows the performance of BRT portfolios using alternative initial training and validation periods. We report results for different specifications that vary in the number of years used for initial training and validation. For example, "14+10" refers to the specification that uses an initial training period of 14 years and a validation period of 10 years. All results are computed following the procedure described in Table 1 in the main paper. The first three columns report average monthly returns for the long-short portfolios as well as the associated Sharpe Ratios. The remaining columns report risk-adjusted returns—see Table 2 in the main paper for details. The top panel reports results for equal-weighted returns. The bottom panel reports results for value-weighted returns. All returns are expressed in percent per month.

Appendix IA.3: Performance of Portfolios Sorted by BRT Predicted Returns – Subsets of Accounting Variables

# of X	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+MOM		Q	
Equal Weight															
	Avg	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
30	0.34	1.07	0.25	0.38	1.15	0.28	0.96	0.52	2.17	0.32	0.99	0.49	1.79	0.37	1.01
60	0.71	2.09	0.52	0.79	2.28	0.70	2.09	0.83	2.96	0.52	1.51	0.62	2.13	0.54	1.49
90	0.75	3.26	0.65	0.92	3.77	0.82	4.50	0.73	4.04	0.69	3.51	0.64	3.12	0.56	2.51
120	0.60	3.54	0.64	0.69	3.94	0.70	4.22	0.70	4.17	0.82	4.02	0.81	3.77	0.77	3.67
150	0.97	6.23	0.96	1.12	7.14	1.04	7.28	1.07	7.32	0.92	5.85	0.95	6.10	0.86	4.48
180	1.27	7.01	1.43	1.39	7.87	1.34	7.59	1.32	7.64	1.23	6.13	1.22	6.21	1.21	5.83
210	1.26	6.67	1.30	1.41	7.43	1.33	7.74	1.28	7.78	1.08	6.20	1.06	6.12	1.04	5.52
240	0.95	6.63	1.02	1.01	6.30	1.01	6.35	1.08	6.43	1.03	5.42	1.08	5.60	0.98	5.11
Value Weight															
	Avg	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
30	-0.12	-0.36	-0.07	-0.02	-0.05	-0.11	-0.35	0.06	0.20	-0.06	-0.16	0.06	0.16	0.06	0.16
60	-0.01	-0.04	-0.01	0.05	0.13	-0.07	-0.21	0.03	0.11	-0.22	-0.68	-0.13	-0.43	-0.18	-0.53
90	0.15	0.53	0.11	0.38	1.26	0.30	1.02	0.14	0.56	0.31	0.99	0.19	0.65	0.26	0.76
120	0.28	0.96	0.20	0.43	1.47	0.41	1.38	0.33	1.25	0.60	1.86	0.52	1.65	0.52	1.61
150	0.32	1.44	0.22	0.57	2.60	0.46	2.18	0.53	2.61	0.26	1.00	0.32	1.29	0.31	1.08
180	0.56	1.94	0.34	0.83	2.94	0.70	2.83	0.81	3.76	0.43	1.51	0.53	2.12	0.54	1.77
210	0.57	2.26	0.39	0.84	3.28	0.71	3.41	0.81	4.04	0.34	1.57	0.45	2.17	0.51	2.02
240	0.40	2.34	0.30	0.46	2.17	0.46	2.57	0.65	3.66	0.68	3.18	0.80	3.80	0.68	2.94

This table shows the performance of BRT portfolios when we vary the number of accounting variables (X) used to construct the fundamental signals. In particular, we progressively increase the number of fundamental variables used in the analysis from 30 to 240 on the basis of their missing rates. All results are computed following the procedure described in Table 1 in the main paper. The first three columns report average monthly returns for the long-short portfolios and the associated Sharpe Ratios. The remaining columns report risk-adjusted returns—see Table 2 in the main paper for details. The top panel reports results for equal-weighted returns. The bottom panel reports results for value-weighted returns. All returns are expressed in percent per month.

Appendix IA.4: Performance of Portfolios Sorted by BRT Predicted Returns – Subsets of Financial Ratio Configurations

	Equal Weight															
# of Y	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+MOM		Q	
	Avg	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
Y3	0.89	4.51	0.90	1.01	5.19	0.97	5.01	0.88	4.86	0.79	4.34	0.74	4.08	0.73	3.98
Y5	0.90	4.14	0.83	1.07	5.11	1.02	4.96	0.94	4.81	0.87	4.33	0.82	4.00	0.77	3.65
P1	0.82	3.87	0.82	0.91	4.25	0.88	4.13	0.80	4.43	0.62	3.15	0.59	3.25	0.60	3.18
P2	0.75	4.34	0.82	0.81	4.17	0.73	4.29	0.73	4.73	0.63	4.15	0.64	4.28	0.57	3.55
P3	0.79	4.16	0.87	0.94	5.26	0.83	5.29	0.80	4.92	0.60	4.70	0.60	4.15	0.64	3.70
P4	0.53	3.13	0.56	0.66	3.90	0.61	3.92	0.60	3.86	0.46	2.81	0.46	2.81	0.40	2.21
P5	0.74	3.81	0.76	0.88	4.99	0.78	4.82	0.78	4.26	0.52	3.84	0.54	3.54	0.58	3.38

	Value Weight															
# of X	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+MOM		Q	
	Avg	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
Y3	0.36	1.80	0.27	0.53	2.38	0.46	2.18	0.39	2.21	0.27	1.45	0.23	1.26	0.28	1.55
Y5	0.64	2.49	0.44	0.83	3.28	0.76	3.36	0.67	3.20	0.56	2.66	0.51	2.34	0.53	2.45
P1	0.28	1.09	0.21	0.37	1.35	0.37	1.29	0.25	0.97	0.19	0.79	0.12	0.51	0.24	0.94
P2	0.21	1.17	0.18	0.35	1.94	0.25	1.54	0.33	1.91	0.10	0.52	0.17	0.89	0.15	0.73
P3	0.13	0.59	0.11	0.37	1.97	0.27	1.37	0.17	0.84	0.06	0.32	0.01	0.06	0.00	-0.02
P4	0.23	1.13	0.21	0.37	1.85	0.33	1.64	0.30	1.51	0.18	0.83	0.17	0.78	0.21	0.98
P5	0.29	1.08	0.22	0.55	2.34	0.42	1.95	0.37	1.49	0.01	0.03	0.00	0.02	0.06	0.25

This table shows the BRT performance of different subsets of financial ratio configurations. Row “Y3” builds signals using three base variables, i.e., total assets, total sales, and market capitalization. Row “Y5” builds signals using five base variables, i.e., total assets, total sales, market capitalization, total liability, and shareholder’s equity. Row “P1” denotes the financial ratios configuration #1-15 and #76; “P2” denotes #16-30 and #76; “P3” denotes #31-45 and #76; “P4” denotes #46-60 and #76; “P5” denotes #61-76. Appendix C in the main paper shows the definitions of the 76 financial ratios configurations. All results are computed following the procedure described in Table 1 in the main paper. The first three columns report average monthly returns for the long-short portfolios and the associated Sharpe Ratios. The remaining columns report risk-adjusted returns—see Table 2 in the main paper for details. The top panel reports results for equal-weighted returns. The bottom panel reports results for value-weighted returns. All returns are expressed in percent per month.

Appendix IA.5: Predictability of Stock Returns across Different Economic and Market Conditions

Panel A. Sentiment

Equal Weight			Value Weight		
High	Low	Diff	High	Low	Diff
1.18	0.79	0.39	0.35	0.52	-0.17
(4.87)	(3.45)	(1.17)	(0.90)	(1.90)	(-0.36)

Panel B. VIX

Equal Weight			Value Weight		
High	Low	Diff	High	Low	Diff
0.8	0.96	-0.16	0.15	0.57	-0.42
(2.74)	(5.25)	(-0.47)	(0.34)	(2.28)	(-0.85)

Panel C. Liquidity

Equal Weight			Value Weight		
High	Low	Diff	High	Low	Diff
0.96	0.94	0.02	0.62	0.18	0.44
(4.64)	(3.63)	(0.05)	(2.18)	(0.46)	(-0.93)

Panel D. Business Cycle

Equal Weight			Value Weight		
Recession	Expansion	Diff	Recession	Expansion	Diff
1.27	0.92	0.35	0.87	0.35	0.52
(1.75)	(5.51)	(0.47)	(0.63)	(1.58)	(0.37)

Panel E. Past Market Returns

Equal Weight			Value Weight		
UP	DOWN	Diff	UP	DOWN	Diff
1.24	0.67	0.57	0.79	0.00	0.79
(4.68)	(3.35)	(1.73)	(2.31)	(0.01)	(1.67)

Panel F. Calendar Subperiods

Equal Weight			Value Weight		
1987-2003	2003-2019	Diff	1987-2003	2003-2019	Diff
1.22	0.69	0.53	0.34	0.46	-0.12
(4.32)	(3.94)	(1.58)	(0.82)	(1.87)	(-0.26)

This table reports the average long-short portfolio returns from the BRT model across subperiods sorted by economic and market conditions. For sentiment, VIX, liquidity, and past market returns, we split the sample period into two subperiods based on the median value of the conditioning variable. We then compute the performance of the BRT model across the different subperiods. In Panel D, we split the sample into recession and expansion based on the NBER recession indicator. In Panel F, we split the sample period into 1987-2003 and 2003-2019. In all cases, we report the performance of the long-short BRT portfolio in each subperiod. We also report whether the difference in long-short performance across different subperiods is statistically significant. All returns are expressed in percent per month.

Appendix IA.6: Performance of Portfolios Sorted by ML Predicted Returns on the GHZ Sample (1987-2016)

Method	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+Mom		Q	
	Avg	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
							Equal Weight								
BRT	3.85	9.51	2.48	3.93	9.88	3.95	9.48	3.62	9.01	3.92	8.03	3.69	8.39	3.57	6.73
NN1	4.11	10.93	2.89	4.18	11.01	4.22	11.18	3.90	10.44	4.08	10.22	3.87	10.03	3.95	9.32
NN2	4.16	10.73	3.14	4.22	10.71	4.24	10.77	3.92	9.96	4.04	10.05	3.83	9.85	3.84	9.22
NN3	4.25	11.00	3.02	4.32	11.19	4.34	11.19	4.03	10.21	4.18	9.88	3.97	9.80	4.05	9.35
NN4	3.15	10.16	2.39	3.22	10.39	3.18	10.10	2.97	7.88	3.08	9.09	2.94	7.76	2.87	7.79
NN5	1.87	6.27	1.41	1.78	5.59	1.90	5.86	1.92	5.82	1.98	5.82	1.99	5.81	2.10	5.74
							Value Weight								
BRT	1.63	4.58	0.75	1.78	5.50	1.81	5.07	1.21	4.19	1.66	3.19	1.25	3.48	1.16	2.19
NN1	1.96	5.72	0.95	2.17	6.57	2.23	7.16	1.52	5.19	1.96	5.10	1.47	4.99	1.55	3.96
NN2	2.09	5.60	1.07	2.25	6.05	2.29	6.68	1.60	5.13	1.94	4.97	1.48	4.82	1.59	3.67
NN3	1.83	4.96	0.94	2.07	5.49	2.05	6.20	1.31	5.48	1.56	3.76	1.07	4.11	1.19	2.99
NN4	1.42	6.09	0.82	1.55	6.02	1.49	6.87	0.95	3.61	1.23	4.00	0.87	2.86	0.84	2.63
NN5	0.72	3.56	0.49	0.78	3.46	0.82	3.99	0.56	2.42	0.82	3.73	0.63	2.76	0.70	3.21

This table reports the returns and risk-adjusted performance for the long-short portfolios sorted by ML-predicted returns from 1987 to 2016 using the GHZ sample of signals. We predict stock monthly excess returns using the 94 signals collected by Green, Hand, and Zhang (2017). We use a recursive window approach and select the optimal hyper-parameters using a cross-validation approach. Our initial estimation period is 1963-1986. The first 12 years is the training period and the second 12 years is the validation period. As we roll forward, the training period expands while the validation period stays at 12 years. The risk-adjusted performance is calculated based on the CAPM model, the Fama-French 3-factor model, the Carhart 4-factor model, the Fama-French 5-factor model, the Fama-French 5-factor model augmented with momentum factor, and the q-factor model. The top panel reports results for equal-weighted portfolios, and the bottom reports results for value-weighted portfolios. All returns are expressed in percent per month.

Appendix IA.7: Performance of Portfolios Sorted by ML Predicted Returns on the GHZ Sample (1991-2014)

Method	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+Mom		Q	
	Avg	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
							Equal Weight								
BRT	3.69	8.00	2.22	3.79	8.25	3.78	7.89	3.41	7.54	3.79	6.72	3.51	7.15	3.38	5.51
NN1	4.19	9.46	2.78	4.27	9.32	4.26	9.54	3.92	8.89	4.13	8.82	3.88	8.78	4.02	8.40
NN2	4.27	9.40	3.07	4.33	9.14	4.31	9.16	3.97	8.51	4.08	8.59	3.85	8.64	3.92	8.19
NN3	4.28	9.51	2.89	4.38	9.60	4.34	9.50	4.01	8.66	4.20	8.63	3.96	8.65	4.10	8.35
NN4	3.12	8.71	2.24	3.20	8.86	3.12	8.40	2.89	6.50	3.01	7.44	2.85	6.42	2.78	6.48
NN5	1.69	5.11	1.21	1.59	4.35	1.69	4.51	1.75	4.57	1.83	4.46	1.86	4.51	1.99	4.51
							Value Weight								
BRT	1.57	3.61	0.67	1.76	4.51	1.76	4.09	1.10	3.28	1.59	2.60	1.11	2.80	1.06	1.74
NN1	2.13	5.19	0.96	2.39	6.18	2.41	6.75	1.63	4.50	2.10	4.77	1.53	4.34	1.69	3.97
NN2	2.21	4.98	1.04	2.40	5.31	2.36	6.00	1.59	4.45	1.91	4.26	1.37	4.22	1.61	3.34
NN3	1.89	4.31	0.90	2.22	5.06	2.13	5.89	1.33	4.99	1.58	3.39	1.01	3.80	1.23	2.92
NN4	1.42	5.38	0.76	1.60	5.28	1.46	5.99	0.85	2.86	1.16	3.16	0.73	2.06	0.72	2.05
NN5	0.63	2.67	0.42	0.70	2.56	0.70	2.80	0.42	1.63	0.70	2.53	0.49	1.88	0.59	2.23

This table reports the returns and risk-adjusted performance for the long-short portfolios sorted by ML-predicted returns from 1991 to 2014 using the GHZ sample of signals. We predict stock monthly excess returns using the 94 signals collected by Green, Hand, and Zhang (2017). We use a recursive window approach and select the optimal hyper-parameters using a cross-validation approach. Our initial estimation period is 1963-1990. The first 16 years is the training period and the second 12 years is the validation period. As we roll forward, the training period expands while the validation period stays at 12 years. The risk-adjusted performance is calculated based on the CAPM model, the Fama-French 3-factor model, the Carhart 4-factor model, the Fama-French 5-factor model, the Fama-French 5-factor model augmented with momentum factor, and the *q*-factor model. The top panel reports results for equal-weighted portfolios, and the bottom reports results for value-weighted portfolios. All returns are expressed in percent per month.

Appendix IA.8: Performance of Portfolios Sorted by ML Predicted Returns on the GHZ Sample (1991-2004)

Method	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+Mom		Q	
	Avg	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
							Equal Weight								
BRT	4.83	8.39	2.48	4.98	8.67	4.97	7.63	4.23	8.06	4.91	6.65	4.26	8.40	4.23	4.40
NN1	5.40	12.09	3.12	5.56	13.15	5.56	12.20	4.87	11.37	5.42	11.19	4.82	10.83	5.20	9.49
NN2	5.61	12.95	3.73	5.77	13.88	5.71	12.37	5.02	12.35	5.46	11.51	4.87	12.85	5.14	9.53
NN3	5.52	11.57	3.27	5.70	13.17	5.60	11.36	4.97	9.75	5.48	10.35	4.93	9.74	5.28	8.90
NN4	3.76	9.78	2.41	3.97	10.08	3.73	8.99	3.20	8.58	3.64	7.46	3.18	8.44	3.24	6.38
NN5	2.31	5.75	1.46	2.19	5.06	2.38	5.27	2.51	5.58	2.67	5.25	2.76	5.50	2.92	5.46
							Value Weight								
BRT	2.00	3.14	0.70	2.15	3.57	2.09	2.73	1.08	2.04	1.87	1.96	1.01	1.96	1.09	1.06
NN1	3.02	7.14	1.19	3.26	7.47	3.04	6.49	1.67	4.86	2.70	4.46	1.54	3.93	2.08	3.38
NN2	3.16	6.08	1.28	3.36	5.95	3.16	5.80	1.89	4.02	2.63	4.21	1.56	3.42	2.06	2.89
NN3	2.68	4.39	1.08	3.06	5.06	2.83	4.83	1.47	4.80	2.16	3.21	1.03	3.07	1.51	2.26
NN4	1.77	4.97	0.79	2.06	4.83	1.67	3.88	0.56	1.52	1.34	2.23	0.40	0.85	0.60	0.99
NN5	0.71	2.09	0.40	0.80	1.90	0.78	1.68	0.44	0.90	0.81	1.57	0.51	1.05	0.76	1.42

This table reports the returns and risk-adjusted performance for the long-short portfolios sorted by ML-predicted returns from 1991 to 2004 using the GHZ sample of signals. We predict stock monthly excess returns using the 94 signals collected by Green, Hand, and Zhang (2017). We use a recursive window approach and select the optimal hyper-parameters using a cross-validation approach. Our initial estimation period is 1963-1990. The first 16 years is the training period and the second 12 years is the validation period. As we roll forward, the training period expands while the validation period stays at 12 years. The risk-adjusted performance is calculated based on the CAPM model, the Fama-French 3-factor model, the Carhart 4-factor model, the Fama-French 5-factor model, the Fama-French 5-factor model augmented with momentum factor, and the q-factor model. The top panel reports results for equal-weighted portfolios, and the bottom reports results for value-weighted portfolios. All returns are expressed in percent per month.

Appendix IA.9: Performance of Portfolios Sorted by ML Predicted Returns on the CZ Sample (1987-2016)

Method	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+Mom		Q	
	Avg	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
								Equal Weight							
BRT	5.55	10.49	3.87	5.67	10.48	5.61	10.79	5.32	10.85	5.37	9.98	5.18	10.42	5.21	9.45
NN1	4.08	8.37	3.08	4.17	8.25	4.15	8.71	4.02	8.25	3.92	8.31	3.85	8.19	3.91	7.80
NN2	3.60	6.85	2.42	3.74	6.83	3.72	7.09	3.65	6.74	3.47	6.70	3.44	6.58	3.53	6.42
NN3	3.52	6.79	2.52	3.68	6.95	3.62	7.22	3.55	6.95	3.32	7.05	3.29	6.95	3.43	6.83
NN4	3.31	6.52	2.51	3.44	6.59	3.40	6.74	3.29	6.37	3.13	6.42	3.08	6.31	3.16	6.10
NN5	1.80	5.40	1.30	1.84	5.21	1.97	5.99	1.92	5.73	2.14	6.98	2.09	6.57	2.13	6.59
								Value Weight							
BRT	2.51	7.24	1.33	2.75	8.01	2.74	7.36	2.10	7.49	2.45	4.84	2.02	6.17	2.11	4.18
NN1	2.51	6.45	1.50	2.64	6.63	2.63	6.76	2.18	6.06	2.37	5.38	2.07	5.69	2.17	4.78
NN2	1.90	5.37	1.10	2.07	5.41	2.02	5.54	1.60	4.76	1.71	4.27	1.43	4.16	1.50	3.95
NN3	1.99	5.52	1.22	2.20	5.89	2.17	6.33	1.76	5.94	1.84	5.17	1.58	5.21	1.64	4.63
NN4	2.00	4.48	1.21	2.15	4.73	2.12	4.96	1.79	4.35	1.83	4.26	1.63	4.12	1.75	3.69
NN5	0.99	3.54	0.66	1.04	3.20	1.12	3.73	0.83	2.69	1.09	4.11	0.88	3.24	0.91	3.29

This table reports the returns and risk-adjusted performance for the long-short portfolios sorted by ML-predicted returns from 1987 to 2016 using the CZ sample of signals. We predict stock monthly excess returns using the March 2022 data release from `openassetpricing.com/data`, which contains the 207 signals collected by Chen and Zimmermann (2022). We use a recursive window approach and select the optimal hyper-parameters using a cross-validation approach. Our initial estimation period is 1963-1986. The first 12 years is the training period and the second 12 years is the validation period. As we roll forward, the training period expands while the validation period stays at 12 years. The risk-adjusted performance is calculated based on the CAPM model, the Fama-French 3-factor model, the Carhart 4-factor model, the Fama-French 5-factor model, the Fama-French 5-factor model augmented with momentum factor, and the q-factor model. The top panel reports results for equal-weighted portfolios. The bottom reports results for value-weighted portfolios. All returns are expressed in percent per month.

Appendix IA.10: Performance of Portfolios Sorted by ML Predicted Returns on the CZ Sample (1991-2014)

Method	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+Mom		Q	
	Avg	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
Equal Weight															
BRT	5.79	9.64	3.83	5.96	9.68	5.86	10.13	5.56	10.13	5.67	9.22	5.46	9.64	5.52	9.08
NN1	4.61	9.07	3.38	4.74	8.99	4.69	9.71	4.58	9.25	4.50	9.29	4.43	9.17	4.52	9.06
NN2	4.24	7.98	2.80	4.44	8.28	4.41	8.63	4.35	8.17	4.19	8.25	4.17	8.07	4.25	7.85
NN3	4.10	7.73	2.86	4.31	8.09	4.23	8.53	4.16	8.11	3.90	8.23	3.87	8.04	4.05	8.05
NN4	3.87	7.36	2.83	4.05	7.70	4.01	7.94	3.91	7.42	3.76	7.61	3.70	7.43	3.79	7.11
NN5	1.99	5.29	1.35	2.02	4.92	2.16	5.68	2.08	5.33	2.36	6.93	2.30	6.41	2.34	6.57
Value Weight															
BRT	2.59	6.49	1.30	2.90	7.39	2.87	6.80	2.19	6.71	2.60	4.46	2.11	5.49	2.31	4.07
NN1	2.84	6.59	1.59	3.01	7.08	2.96	7.08	2.48	6.40	2.71	5.60	2.37	6.09	2.53	5.18
NN2	2.24	5.82	1.21	2.50	6.58	2.43	6.58	1.97	5.79	2.18	5.16	1.86	5.19	1.91	4.98
NN3	2.31	6.00	1.33	2.56	6.53	2.51	7.10	2.03	6.57	2.11	5.48	1.78	5.77	1.90	5.03
NN4	2.36	4.73	1.34	2.57	5.30	2.53	5.55	2.17	4.81	2.23	4.78	1.98	4.57	2.15	4.08
NN5	1.20	3.78	0.75	1.27	3.38	1.35	3.87	1.02	2.78	1.32	4.35	1.07	3.34	1.13	3.50

This table reports the returns and risk-adjusted performance for the long-short portfolios sorted by ML-predicted returns from 1991 to 2014 using the CZ sample of signals. We predict stock monthly excess returns using the March 2022 data release from openassetpricing.com/data, which contains the 207 signals collected by Chen and Zimmermann (2022). We use a recursive window approach and select the optimal hyper-parameters using a cross-validation approach. Our initial estimation period is 1963-1990. The first 16 years is the training period and the second 12 years is the validation period. As we roll forward, the training period expands while the validation period stays at 12 years. The risk-adjusted performance is calculated based on the CAPM model, the Fama-French 3-factor model, the Carhart 4-factor model, the Fama-French 5-factor model, the Fama-French 5-factor model augmented with momentum factor, and the q-factor model. The top panel reports results for equal-weighted portfolios. The bottom reports results for value-weighted portfolios. All returns are expressed in percent per month.

Appendix IA.11: Performance of Portfolios Sorted by ML Predicted Returns on the CZ Sample (1991-2004)

Method	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+Mom		Q	
	Avg	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
								Equal Weight							
BRT	7.56	12.67	4.59	7.79	13.11	7.52	12.61	7.05	14.68	7.40	12.26	7.00	15.05	7.21	11.02
NN1	5.79	8.90	3.72	5.98	9.14	5.86	10.12	5.71	9.44	5.62	9.73	5.51	9.40	5.69	8.97
NN2	5.35	7.66	3.05	5.61	8.32	5.56	8.87	5.56	8.33	5.37	8.61	5.39	8.29	5.48	7.56
NN3	5.20	7.54	3.21	5.48	8.24	5.19	8.00	5.13	7.62	4.83	7.42	4.80	7.22	5.11	7.51
NN4	5.04	7.40	3.36	5.25	7.83	5.05	7.45	4.99	6.97	4.77	7.08	4.74	6.89	4.89	6.71
NN5	2.36	4.18	1.36	2.36	3.73	2.76	4.97	2.55	4.34	2.92	5.87	2.73	5.05	2.85	5.31
								Value Weight							
BRT	3.45	6.51	1.48	3.82	7.43	3.68	4.98	2.64	5.07	3.48	3.99	2.60	4.82	3.02	3.23
NN1	3.66	6.31	1.77	3.85	6.65	3.68	5.63	2.97	4.99	3.41	4.75	2.81	5.01	3.17	4.12
NN2	2.93	6.17	1.37	3.28	7.03	3.12	5.70	2.45	5.45	2.95	4.87	2.39	5.03	2.56	4.26
NN3	3.11	6.70	1.61	3.40	6.97	3.13	6.20	2.43	6.31	2.77	5.09	2.19	5.37	2.43	4.39
NN4	3.53	6.55	1.83	3.76	6.86	3.61	6.51	3.28	6.22	3.20	5.49	2.95	5.52	3.35	5.08
NN5	1.55	3.21	0.83	1.61	2.71	1.78	3.04	1.22	2.09	1.83	3.72	1.34	2.65	1.36	2.43

This table reports the returns and risk-adjusted performance for the long-short portfolios sorted by ML-predicted returns from 1991 to 2004 using the CZ sample of signals. We predict stock monthly excess returns using the March 2022 data release from openassetpricing.com/data, which contains the 207 signals collected by Chen and Zimmermann (2022). We use a recursive window approach and select the optimal hyper-parameters using a cross-validation approach. Our initial estimation period is 1963-1990. The first 16 years is the training period and the second 12 years is the validation period. As we roll forward, the training period expands while the validation period stays at 12 years. The risk-adjusted performance is calculated based on the CAPM model, the Fama-French 3-factor model, the Carhart 4-factor model, the Fama-French 5-factor model, the Fama-French 5-factor model augmented with momentum factor, and the q-factor model. The top panel reports results for equal-weighted portfolios. The bottom reports results for value-weighted portfolios. All returns are expressed in percent per month.

Appendix IA.12: Performance of ML Portfolios on the GHZ sample with and without Short-term Reversal, different OOS periods

Method	Panel A: 1987-2016						Panel B: 1991-2014						Panel C: 1991-2004					
	GHZ94			GHZ93			GHZ94			GHZ93			GHZ94			GHZ93		
	Ret.	*t-stat*	SR	*Ret.*	*t-stat*	SR	*Ret.*	*t-stat*	SR	*Ret.*	*t-stat*	SR	*Ret.*	*t-stat*	SR	*Ret.*	*t-stat*	SR
							Equal Weight											
BRT	3.85	9.51	2.48	3.19	8.97	1.84	4.83	8.39	2.48	2.97	7.20	1.60	3.69	8.00	2.22	3.73	6.06	1.62
NN1	4.11	10.93	2.89	3.49	11.09	2.62	5.40	12.09	3.12	3.50	9.56	2.47	4.19	9.46	2.78	4.43	11.60	2.64
NN2	4.16	10.73	3.14	3.59	11.25	2.38	5.61	12.95	3.73	3.69	9.85	2.27	4.27	9.40	3.07	4.46	10.25	2.30
NN3	4.25	11.00	3.02	3.43	10.63	2.52	5.52	11.57	3.27	3.55	9.68	2.47	4.28	9.51	2.89	4.36	11.99	2.64
NN4	3.15	10.16	2.39	2.46	9.21	1.90	3.76	9.78	2.41	2.31	7.99	1.71	3.12	8.71	2.24	2.75	8.15	1.72
NN5	1.87	6.27	1.41	0.76	3.71	0.64	2.31	5.75	1.46	0.75	3.21	0.59	1.69	5.11	1.21	1.05	3.15	0.69
							Value Weight											
BRT	1.63	4.58	0.75	1.21	3.73	0.55	2.00	3.14	0.70	1.10	2.80	0.46	1.57	3.61	0.67	1.41	2.55	0.49
NN1	1.96	5.72	0.95	1.59	4.76	0.84	3.02	7.14	1.19	1.46	3.68	0.72	2.13	5.19	0.96	2.30	5.52	0.99
NN2	2.09	5.60	1.07	1.75	5.95	0.89	3.16	6.08	1.28	1.81	5.09	0.86	2.21	4.98	1.04	2.34	4.79	0.92
NN3	1.83	4.96	0.94	1.59	4.27	0.86	2.68	4.39	1.08	1.82	4.16	0.92	1.89	4.31	0.90	2.45	4.27	1.07
NN4	1.42	6.09	0.82	1.11	3.82	0.65	1.77	4.97	0.79	0.97	2.81	0.53	1.42	5.38	0.76	1.26	2.58	0.61
NN5	0.72	3.56	0.49	0.30	1.35	0.21	0.71	2.09	0.40	0.15	0.62	0.10	0.63	2.67	0.42	0.26	0.76	0.14

This table reports the long-short returns for the portfolios sorted by ML-predicted returns on the GHZ sample of signals with and without short-term reversal. This table is similar to Table 8 but features three different out-of-sample (OOS) periods, namely 1987-2016, 1991-2014, and 1991-2004. Panel A compares their OOS performance from 1987 to 2016, Panel B compares the performance from 1991 to 2014, and Panel C compares the performance from 1991 to 2004. We predict stock monthly excess returns using signals in Green, Hand, and Zhang (2017). GHZ94 denotes the 94 signals, and GHZ93 denotes the 93 signals excluding short-term reversal (or mom1m). We use a recursive window approach and select the optimal hyper-parameters using a cross-validation approach. Our initial estimation period is 1963-1986. The first 12 years is the training period and the second 12 years is the validation period. As we roll forward, the training period expands while the validation period stays at 12 years. The top panel reports results for equal-weighted portfolios. The bottom reports results for value-weighted portfolios. All returns are expressed in percent per month.

Appendix IA.13: Performance of ML Portfolios on the CZ sample with and without Short-term Reversal, different OOS periods

Method	Panel A: 1987-2016						Panel B: 1991-2014						Panel C: 1991-2004					
	CZ207			CZ206			CZ207			CZ206			CZ207			CZ206		
	Ret.	*t-stat*	SR	*Ret.*	*t-stat*	SR	*Ret.*	*t-stat*	SR	*Ret.*	*t-stat*	SR	*Ret.*	*t-stat*	SR	*Ret.*	*t-stat*	SR
	Equal Weight																	
BRT	5.55	10.49	3.87	5.21	10.66	3.36	5.79	9.64	3.83	5.48	9.95	3.34	7.56	12.67	4.59	7.09	13.21	3.83
NN1	4.08	8.37	3.08	3.71	9.00	2.88	4.61	9.07	3.38	4.13	9.35	3.11	5.79	8.90	3.72	4.97	8.23	3.25
NN2	3.60	6.85	2.42	3.50	6.81	2.43	4.24	7.98	2.80	4.14	8.01	2.83	5.35	7.66	3.05	5.00	6.69	2.93
NN3	3.52	6.79	2.52	3.52	6.77	2.45	4.10	7.73	2.86	4.11	7.61	2.77	5.20	7.54	3.21	5.04	6.88	2.98
NN4	3.31	6.52	2.51	3.15	6.20	2.13	3.87	7.36	2.83	3.76	7.21	2.42	5.04	7.40	3.36	4.83	6.79	2.64
NN5	1.80	5.40	1.30	1.60	5.00	1.26	1.99	5.29	1.35	1.85	5.17	1.35	2.36	4.18	1.36	2.42	4.69	1.55
	Value Weight																	
BRT	2.51	7.24	1.33	2.36	6.70	1.09	2.59	6.49	1.30	2.50	6.25	1.09	3.45	6.51	1.48	3.21	6.05	1.19
NN1	2.51	6.45	1.50	2.27	5.52	1.39	2.84	6.59	1.59	2.49	5.25	1.43	3.66	6.31	1.77	3.19	4.98	1.61
NN2	1.90	5.37	1.10	2.13	5.56	1.14	2.24	5.82	1.21	2.54	6.12	1.28	2.93	6.17	1.37	3.08	5.29	1.33
NN3	1.99	5.52	1.22	2.01	5.00	1.19	2.31	6.00	1.33	2.42	5.61	1.35	3.11	6.70	1.61	3.25	5.75	1.57
NN4	2.00	4.48	1.21	2.25	4.63	1.25	2.36	4.73	1.34	2.70	5.05	1.41	3.53	6.55	1.83	3.87	5.71	1.74
NN5	0.99	3.54	0.66	1.12	4.67	0.82	1.20	3.78	0.75	1.34	5.09	0.92	1.55	3.21	0.83	1.97	6.22	1.15

This table reports the excess returns for the portfolios sorted by ML-predicted returns on the CZ sample of signals with and without short-term reversal. This table is similar to Table 9 but features three different out-of-sample (OOS) periods, namely 1987-2016, 1991-2014, and 1991-2004. Panel A compares their OOS performance from 1987 to 2016, Panel B compares the performance from 1991 to 2014, and Panel C compares the performance from 1991 to 2004. We predict stock monthly excess returns using signals in Chen and Zimmermann (2022). CZ207 denotes the 207 signals, and CZ206 denotes the 206 signals excluding short-term reversal (or *STreversal*). We use a recursive window approach and select the optimal hyper-parameters using a cross-validation approach. Our initial estimation period is 1963-1986. The first 12 years is the training period and the second 12 years is the validation period. As we roll forward, the training period expands while the validation period stays at 12 years. The top panel reports results for equal-weighted portfolios. The bottom reports results for value-weighted portfolios. All returns are expressed in percent per month.

Appendix IA.14: Performance of Portfolios Sorted by ML Predicted Returns on the GHZ Sample excluding Short-term Reversal

Method	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+Mom		Q	
	Avg	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
							Equal Weight								
BRT	3.04	9.11	1.80	3.24	9.59	3.28	9.36	2.82	7.84	3.18	7.03	2.85	7.27	2.79	5.08
NN1	3.22	9.99	2.46	3.33	10.36	3.36	10.46	3.05	9.19	3.26	9.48	3.05	9.09	3.14	8.55
NN2	3.37	10.68	2.30	3.49	10.87	3.50	10.90	3.03	8.30	3.29	8.69	2.96	7.93	3.05	7.53
NN3	3.27	10.57	2.48	3.39	10.85	3.39	10.88	3.04	8.58	3.19	10.07	2.95	8.99	3.06	8.25
NN4	2.30	8.91	1.84	2.41	9.37	2.42	9.45	2.16	7.26	2.27	9.23	2.09	7.49	2.18	7.98
NN5	0.72	3.65	0.62	0.66	3.23	0.71	3.57	0.57	2.42	0.79	3.81	0.68	2.97	0.70	2.87
							Value Weight								
BRT	1.11	3.67	0.52	1.34	4.85	1.37	4.64	0.78	3.17	1.21	2.87	0.79	2.81	0.78	1.67
NN1	1.42	4.42	0.77	1.63	4.96	1.63	5.36	1.08	3.93	1.44	4.49	1.05	3.87	1.06	3.44
NN2	1.60	5.75	0.85	1.82	6.13	1.82	6.63	1.16	5.56	1.41	4.15	0.97	4.04	1.02	3.19
NN3	1.50	4.34	0.84	1.71	4.72	1.69	5.37	1.03	4.07	1.26	4.05	0.81	3.37	0.96	3.36
NN4	0.97	3.45	0.58	1.15	4.03	1.19	4.63	0.68	3.12	0.98	4.01	0.62	2.72	0.63	2.64
NN5	0.30	1.46	0.22	0.26	1.20	0.32	1.55	0.03	0.13	0.39	1.83	0.17	0.83	0.13	0.49

This table reports the returns and risk-adjusted performance for the long-short portfolios sorted by ML-predicted returns from 1987 to 2019. We predict stock monthly excess returns using 93 signals from the GHZ sample, which we obtain by excluding the short-term reversal (mom1m) from the 94 signals by Green, Hand, and Zhang (2017). We use a recursive window approach and select the optimal hyper-parameters using a cross-validation approach. Our initial estimation period is 1963-1986. The first 12 years is the training period and the second 12 years is the validation period. As we roll forward, the training period expands while the validation period stays at 12 years. The risk-adjusted performance is calculated based on the CAPM model, the Fama-French 3-factor model, the Carhart 4-factor model, the Fama-French 5-factor model, the Fama-French 5-factor model augmented with momentum factor, and the q-factor model. The top panel reports results for equal-weighted portfolios. The bottom reports results for value-weighted portfolios. All returns are expressed in percent per month.

Appendix IA.15: Performance of Portfolios Sorted by ML Predicted Returns on the CZ Sample excluding Short-term Reversal

Method	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+Mom		Q	
	Avg	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
							Equal Weight								
BRT	3.04	9.11	1.80	3.24	9.59	3.28	9.36	2.82	7.84	3.18	7.03	2.85	7.27	2.79	5.08
NN1	3.22	9.99	2.46	3.33	10.36	3.36	10.46	3.05	9.19	3.26	9.48	3.05	9.09	3.14	8.55
NN2	3.37	10.68	2.30	3.49	10.87	3.50	10.90	3.03	8.30	3.29	8.69	2.96	7.93	3.05	7.53
NN3	3.27	10.57	2.48	3.39	10.85	3.39	10.88	3.04	8.58	3.19	10.07	2.95	8.99	3.06	8.25
NN4	2.30	8.91	1.84	2.41	9.37	2.42	9.45	2.16	7.26	2.27	9.23	2.09	7.49	2.18	7.98
NN5	0.72	3.65	0.62	0.66	3.23	0.71	3.57	0.57	2.42	0.79	3.81	0.68	2.97	0.70	2.87
							Value Weight								
BRT	1.11	3.67	0.52	1.34	4.85	1.37	4.64	0.78	3.17	1.21	2.87	0.79	2.81	0.78	1.67
NN1	1.42	4.42	0.77	1.63	4.96	1.63	5.36	1.08	3.93	1.44	4.49	1.05	3.87	1.06	3.44
NN2	1.60	5.75	0.85	1.82	6.13	1.82	6.63	1.16	5.56	1.41	4.15	0.97	4.04	1.02	3.19
NN3	1.50	4.34	0.84	1.71	4.72	1.69	5.37	1.03	4.07	1.26	4.05	0.81	3.37	0.96	3.36
NN4	0.97	3.45	0.58	1.15	4.03	1.19	4.63	0.68	3.12	0.98	4.01	0.62	2.72	0.63	2.64
NN5	0.30	1.46	0.22	0.26	1.20	0.32	1.55	0.03	0.13	0.39	1.83	0.17	0.83	0.13	0.49

This table reports the returns and risk-adjusted performance for the long-short portfolios sorted by ML-predicted returns on the CZ sample from 1987 to 2019. We predict stock monthly excess returns using 206 signals, which exclude the short-term reversal (or *Streversal*) from the 207 signals (Chen and Zimmermann, 2022). We use a recursive window approach and select the optimal hyper-parameters using a cross-validation approach. Our initial estimation period is 1963-1986. The first 12 years is the training period and the second 12 years is the validation period. As we roll forward, the training period expands while the validation period stays at 12 years. The risk-adjusted performance is calculated based on the CAPM model, the Fama-French 3-factor model, the Carhart 4-factor model, the Fama-French 5-factor model, the Fama-French 5-factor model augmented with momentum factor, and the q-factor model. The top panel reports results for equal-weighted portfolios. The bottom reports results for value-weighted portfolios. All returns are expressed in percent per month.

Appendix IA.16: Performance of ML Portfolios after Trading Costs – GHZ Sample (1987-2017)

Methods	Panel A: GHZ94					Panel B: GHZ93				
	a	b	c	d≈ b× c	e=a-d	a	b	c	d≈ b× c	e=a-d
Equal Weight										
BRT	374	137	328	462	-87	312	113	302	343	-31
NN1	401	138	329	459	-58	339	129	317	413	-75
NN2	407	141	318	453	-46	352	132	299	398	-47
NN3	418	141	316	449	-31	340	134	296	397	-57
NN4	308	146	300	439	-131	242	145	284	410	-169
NN5	182	119	282	355	-174	80	109	275	310	-229
Value Weight										
BRT	156	156	97	154	2	116	139	94	131	-15
NN1	187	164	101	164	23	152	161	98	157	-5
NN2	202	169	96	163	38	173	166	91	151	22
NN3	184	168	96	162	22	160	166	90	148	11
NN4	140	167	93	156	-16	108	167	89	148	-39
NN5	70	127	88	119	-49	31	115	87	106	-75

This table reports the gross return, turnover, trading cost, and net return for the long-short portfolios sorted by ML-predicted returns based on the two GHZ sample. GHZ94 in Panel A contains 94 signals from Green, Hand, and Zhang (2017), and GHZ93 in Panel B includes 93 signals excluding short-term reversal. All signals are generated following their baseline specifications, but the sample ends in 2017 because the low-frequency effective spreads data are only available until 2017 on Andrew Chen's website at https://sites.google.com/site/chenandrewy/. Column a denotes the gross return before transaction costs; Column b denotes the turnover (2-sided) calculated following Chen and Velikov (2022). In particular, we calculate the actual weight at the end of the month and calculate the turnover as the sum of absolute deviation from the next weight and the actual weight. We then take the average of long and short legs and average over the time series. Column d denotes the return reduction or the incurred trading cost, calculated as the sum of absolute deviation multiplied by low-frequency spreads. Column c is the average spread paid, calculated as the return reduction divided by the turnover. Column e reports the net return after trading cost. All results except column b are expressed in basis points per month. Column b (or turnover (2-sided)) is expressed in percent per month. Results on other periods (1991-2004, 1991-2014, and 1987-2016) are similar.

Appendix IA.17: Performance of ML Portfolios after Trading Cost – CZ Sample (1987-2017)

	Panel A: CZ207					Panel B: CZ206				
Methods	a	b	c	d≈ b× c	e=a-d	a	b	c	d≈ b× c	e=a-d
					Equal Weight					
BRT	541	140	272	385	156	508	127	251	320	188
NN1	394	145	245	356	38	355	137	237	326	29
NN2	347	146	248	362	-15	339	139	237	330	9
NN3	340	120	250	312	27	340	119	239	286	54
NN4	320	112	248	288	32	305	109	238	255	50
NN5	174	106	228	237	-63	155	98	217	207	-52
					Value Weight					
BRT	245	164	91	150	94	228	160	90	144	84
NN1	243	167	81	135	108	218	163	82	133	84
NN2	182	164	86	141	42	207	162	81	132	75
NN3	194	137	83	123	70	195	141	81	118	76
NN4	196	125	82	113	83	221	125	78	103	117
NN5	98	112	75	93	5	109	102	70	77	33

This table reports the gross return, turnover, trading cost, and net return for the long-short portfolios sorted by ML-predicted returns based on the two CZ samples. CZ207 in Panel A contains 207 signals, and CZ206 in Panel B includes 206 signals excluding the short-term reversal. All signals are generated following their baseline specifications, but the sample ends in 2017 because the low-frequency effective spreads data are only available until 2017 on Andrew Chen's website at https://sites.google.com/site/chenandrewy/. Column a denotes the gross return before transaction costs; Column b denotes the turnover (2-sided) calculated following Chen and Velikov (2022). In particular, we calculate the actual weight at the end of the month and calculate the turnover as the sum of absolute deviation from the next weight and the actual weight. We then take the average of long and short legs and average over the time series. Column d denotes the return reduction or the incurred trading cost, calculated as the sum of absolute deviation multiplied by low-frequency spreads. Column c is the average spread paid, calculated as the return reduction divided by the turnover. Column e reports the net return after trading cost. All results except column b are expressed in basis points per month. Column b (or turnover (2-sided)) is expressed in percent per month. Results on other periods (1991-2004, 1991-2014, and 1987-2016) are similar.

Appendix IA.18: Performance of Portfolios Sorted by BRT Predicted Returns based on Past-Return Signals

Rank	*Pred*	Returns		SR	CAPM		FF3		Carhart		FF5		FF5+Mom		Q	
		Avg	*t-stat*		*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*	*alpha*	*t-stat*
Equal Weighted																
1 (Low)	-0.60	-0.29	-0.61	-0.12	-1.25	-4.31	-1.18	-5.44	-0.76	-4.31	-0.68	-3.03	-0.41	-2.60	-0.36	-1.79
2	0.14	0.16	0.43	0.08	-0.67	-3.14	-0.65	-5.14	-0.38	-3.49	-0.38	-2.99	-0.20	-1.91	-0.13	-0.98
3	0.45	0.44	1.31	0.25	-0.31	-1.62	-0.34	-3.52	-0.12	-1.18	-0.18	-1.56	-0.04	-0.32	0.00	0.00
4	0.60	0.63	2.21	0.41	-0.05	-0.26	-0.08	-1.10	0.06	0.81	-0.01	-0.13	0.09	1.13	0.12	1.16
5	0.70	0.67	2.56	0.47	0.03	0.20	-0.01	-0.24	0.09	1.36	-0.01	-0.20	0.06	0.98	0.07	0.78
6	0.78	0.81	3.33	0.61	0.21	1.30	0.16	2.65	0.22	3.25	0.14	2.29	0.18	2.84	0.20	2.02
7	0.85	0.84	3.43	0.64	0.24	1.52	0.18	2.73	0.20	3.05	0.13	2.00	0.15	2.24	0.16	1.73
8	0.92	0.93	3.74	0.71	0.34	2.07	0.29	4.16	0.29	4.06	0.25	3.44	0.25	3.39	0.26	2.55
9	1.01	0.95	3.67	0.68	0.32	2.03	0.28	3.92	0.26	3.59	0.26	3.54	0.24	3.22	0.26	2.82
10 (High)	1.23	1.09	3.48	0.68	0.38	1.92	0.38	3.25	0.33	2.78	0.41	3.59	0.37	3.20	0.41	3.19
10-1	1.82	1.38	4.93	1.04	1.63	5.86	1.56	6.90	1.09	6.62	1.09	4.82	0.78	5.83	0.78	3.89
Value Weighted																
1 (Low)	-0.46	0.19	0.37	0.07	-0.91	-2.99	-0.78	-3.29	-0.48	-2.19	-0.30	-1.35	-0.12	-0.62	-0.08	-0.39
2	0.16	0.32	0.88	0.16	-0.57	-3.39	-0.51	-3.47	-0.24	-1.77	-0.19	-1.52	-0.02	-0.13	-0.05	-0.41
3	0.46	0.62	2.01	0.37	-0.15	-1.17	-0.15	-1.23	0.02	0.19	0.04	0.37	0.15	1.20	0.11	0.99
4	0.60	0.50	1.73	0.34	-0.20	-2.18	-0.20	-2.12	-0.05	-0.59	-0.14	-1.56	-0.04	-0.39	-0.06	-0.62
5	0.70	0.50	2.12	0.38	-0.13	-1.37	-0.16	-1.88	-0.09	-1.13	-0.20	-2.19	-0.15	-1.78	-0.15	-1.54
6	0.78	0.57	2.85	0.47	0.00	-0.02	-0.04	-0.47	0.01	0.14	-0.14	-1.94	-0.09	-1.32	-0.09	-1.03
7	0.85	0.64	3.36	0.54	0.06	0.80	0.04	0.61	0.00	-0.07	-0.06	-1.24	-0.09	-1.73	-0.09	-1.67
8	0.92	0.73	3.39	0.61	0.15	2.06	0.15	1.92	0.07	0.84	0.06	0.80	0.01	0.15	0.04	0.51
9	1.01	0.74	3.08	0.57	0.12	1.48	0.10	1.27	-0.02	-0.23	0.04	0.46	-0.05	-0.54	-0.06	-0.72
10 (High)	1.18	0.96	3.24	0.63	0.26	1.89	0.30	2.38	0.15	1.36	0.27	2.18	0.16	1.41	0.20	1.63
10-1	1.65	0.78	2.41	0.46	1.17	4.00	1.07	4.23	0.63	3.05	0.56	2.37	0.28	1.55	0.28	1.36

This table reports the excess returns of decile portfolios sorted by BRT predicted returns based on past-return signals from 1987 to 2019. We predict stock monthly excess returns using 119 technical signals (as described in Section 3.5). We use a recursive window approach and select the optimal hyper-parameters using a cross-validation approach. Our initial estimation period is 1963-1986. The first 12 years (or 144 months) is the training period, and the second 12 years (or 144 months) is the validation period. As we roll forward, the training period expands while the validation period stays at 12 years. In the first column, we report the average predicted monthly returns from the BRT model (*Pred*). The second and third columns report the average realized monthly excess returns (*Avg*) and associated t-statistics (*t-stat*), computed using Newey and West (1987) standard errors with 12 lags. Then, we report the portfolios' Sharpe ratios (*SR*). The remaining columns show the risk-adjusted performance is calculated based on the CAPM model, the Fama-French 3-factor model, the Carhart 4-factor model, the Fama-French 5-factor model, the Fama-French 5-factor model augmented with momentum factor, and the q-factor model. The top panel reports results for equal-weighted portfolios, and the bottom reports results for value-weighted portfolios. All returns are expressed in percent per month.