Multi-Factor Timing with Deep Learning*

Paul Cotturo†, Fred Liu‡§, Robert Proner§

January 17, 2024

Abstract

We develop deep neural networks with economically motivated restrictions that are designed to overcome the main challenges of factor timing. Our critical innovations include integrating multi-task learning to capture the common structure across factors, with long short-term memory neural networks to extract financial and macroeconomic states. This dynamic multi-task neural network significantly outperforms all benchmarks in terms of predictive accuracy and economic gains. We pinpoint tail risk, along with variations on leverage, profitability, and momentum as key predictors, and highlight the importance of capturing their nonlinear interactions. Improved factor timing through neural networks with economic restrictions facilitates more reliable investigation into the economic mechanisms driving factor risk premia, and underscores the value of deep learning for factor investing.

JEL Classification: G10, G11, G12, G17, C14, C22, C45, C58

Keywords: Factor Timing, Deep Learning, Economic Structure, Machine Learning, Multi-task Neural Networks, Big Data.

*This paper was supported by SSHRC Insight Development Grant 430-2022-00399, and the Digital Research Alliance of Canada Resources for Research Groups 2022 competition (ID 4286).

†Department of Statistics and Actuarial Science, University of Waterloo, 200 University Ave W, Waterloo, Ontario, N2L3G1, Canada.

‡Department of Economics and Finance, University of Guelph, 50 Stone Rd E, Guelph, Ontario, N1G2W1, Canada. Email: fred.liu@uoguelph.ca. Corresponding author.

§Department of Economics, University of Western Ontario, 1151 Richmond St, London, Ontario, N6A3K7, Canada.

1 Introduction

Factors in the cross-section of stock returns are persistent sources of risk premia (Fama and French (1993)), forming the basis for factor investing. Factors such as value and momentum deliver high returns over the long run, but can underperform in the short run, creating the opportunity for investors to engage in factor timing. Many investors use multi-factor portfolio strategies that tilt towards (buy) factors that are likely to outperform, and tilt away from (sell) factors that are likely to underperform, which raises several interesting questions. First, can we predict the probability of a factor outperforming or underperforming? Second, what is the economic significance of these predictions? Third, which variables are influential for factor timing, and do their nonlinear interactions matter? This paper is dedicated to answering these questions, consequently enabling a more reliable investigation into the economic mechanisms that drive factor risk premia.

There are four major challenges of factor timing that the literature so far has struggled to overcome in a unified framework. First, factors are a function of many financial and macroeconomic variables (Bender et al. (2018), Dong et al. (2022)). Second, the functional form of factors is unknown and likely complex, depending on nonlinear interactions among the variables (Didisheim et al. (2023), Gu et al. (2020), Kelly et al. (2023)). Third, factors are time-varying and depend on short- and long-term macroeconomic and financial conditions (Ilmanen et al. (2021), Hodges et al. (2017), Polk et al. (2020)). Fourth, factors share a common structure that should be summarized using a few latent features (Haddad et al. (2020), Kagkadis et al. (2023)). While the literature has largely used linear models that can address some of these challenges, none of these models, to our knowledge, can overcome all of them. As a result, the fundamental question concerning the feasibility of factor timing remains contentious (Asness (2016), Asness et al. (2017), Dichtl et al. (2019)).

In this paper, we introduce a deep learning approach for factor timing that addresses the aforementioned challenges by incorporating economically motivated restrictions. Since factor timing is fundamentally a prediction problem, employing neural networks is a natural

choice, as they excel at handling a large number of variables and complex functional forms. However, off-the-shelf neural networks are typically designed for prediction tasks in static environments with high signal-to-noise ratios and a large number of observations. However, factors exhibit time-varying risk premia, low signal-to-noise ratios, and limited historical observations, leading off-the-shelf neural networks to overfit. Hence, economic restrictions play a crucial role in regularization and are essential for improving the out-of-sample predictive accuracy of neural networks.

We demonstrate how to improve deep learning models for factor timing by incorporating economic structure. Our crucial innovation is to develop a multi-task neural network (MT) architecture to jointly predict all factors within a single functional form. MT learns common latent features across factors and uses factor-specific layers to capture each factor's nonlinear exposures to these latent features. This economic restriction offers two main benefits that improve the model's out-of-sample generalizability. First, MTs incorporate the economic restriction that all factors stem from a low-dimensional set of common latent features, which allows the model to capture commonalities among factors and improves performance. This shared representation has been shown to enhance data efficiency (Caruana, 1997), which is critical for applying deep learning to the limited historical observations of factors. In contrast, single-factor models are unable to leverage the low-dimensional common structure across factors, which makes them susceptible to overfitting on factor-specific noise. Second, multi-task learning (MTL) acts as a regularization technique that compels MT to generalize across multiple outputs (Ruder, 2017), thereby preventing overfitting to the noise of a single factor. This safeguard against overfitting is critical for factors, given their low signal-to-noise ratios.

Furthermore, we incorporate a high-dimensional set of variables and capture the time-variation of the factors as a function of macroeconomic and financial conditions using two separate recurrent Long Short-Term Memory neural networks (LSTMs), which perform dimension reduction and extract time series dynamics. The first LSTM takes in a large number

of macroeconomic time series as inputs, and summarizes their dynamics into a small number of macroeconomic state processes; while the second LSTM takes in a large number of financial time series as inputs, and summarizes their dynamics into a small number of financial state processes. Chen et al. (2023) demonstrate that LSTMs can capture short and long-term dependencies, which are necessary for detecting cycles, hence our LSTMs are designed to capture business and financial cycles. The economic restriction of separately modeling the macroeconomic and financial cycle dynamics reduces the number of parameters, and further enhances the neural network's out-of-sample performance. Augmenting the MT architecture with LSTMs gives rise to the dynamic multi-task neural network (DMT) model to address all four major challenges of factor timing.

We study six well-known factors over 57 years from January 1965 to December 2021. These factors include the excess market return (MKT), size (SMB), and value (HML) from Fama and French (1996), profitability (RMW) and investment (CMA) from Fama and French (2015), and momentum (MOM). Our predictors consist of 272 variables, including 123 macroeconomic variables from McCracken and Ng (2016), and 149 financial variables from Chen and Zimmermann (2022) and Welch and Goyal (2008). In each month, we condition on these predictors to forecast the probability of each factor earning a positive risk premium, i.e., the factor goes up. We compare MT and DMT, which predict the probabilities of all factors in a single functional form, against off-the-shelf models that estimate a separate functional form for each factor. These off-the-shelf classification models include logistic regression (LR), penalized logistic regression (EN), random forest (RF), extremely randomized trees (XRF), gradient boosted trees (GBT), support vector machine (SVM), and feed-forward neural network (NN).

Our results advance the knowledge on factor timing in five main dimensions. First, we assess the predictive power of the different models for factor timing using the out-of-sample predictive accuracy metric, defined as the proportion of correctly classified excess return directions in the out-of-sample period from January 1990 to December 2021. We show that

economic restrictions matter for the average accuracy across factors. Linear models, LR and EN, post accuracies of 53.9% and 54.1%, respectively, underperforming the 55.3% accuracy of the buy-and-hold benchmark (Buy) that always predicts a positive excess return. In contrast, nonlinear models based on decision trees RF, XRF, and GBT deliver higher accuracies of 56.8%, 55.9%, and 54.6%, respectively. SVM and NN exhibit the lowest accuracies of 53.5% and 52.3%, respectively, since they are too flexible and overfit to factor-specific noise. However, with the imposition of an economically motivated restriction of a common structure across factors, MT raises the accuracy to 55.7%, suggesting that the regularization and data efficiency benefits gained from MTL improve forecasting accuracy. Furthermore, by incorporating time series dynamics, DMT delivers the highest average accuracy of 57.2%. Notably, it is the only model to outperform the Buy for every factor, and is the most accurate model for MKT, RMW, CMA, and MOM.

We also conduct pairwise comparisons between different machine learning models using the Diebold and Mariano (1995) test statistic, and use the average log loss across factors to compare probability forecasts between models. DMT significantly outperforms all other models with t-statistics ranging from 2.97 to 13.27, underscoring the importance of incorporating economic structure and time series dynamics into factor timing models.

Second, we study whether factor timing using machine learning models can be exploited in an economically significant trading strategy. We employ a strategy that buys factors if the model predicts a positive return and invests in the risk-free rate otherwise. Our multi-factor strategy is then an equal-weighted portfolio of these strategy excess returns across all factors. We find that the economic significance of models aligns very closely with their average accuracies. Linear models LR and EN, with Sharpe ratios of 0.84 and 0.83 respectively, underperform compared to the multi-factor Buy Sharpe ratio of 0.98, attributable to their less accurate forecasts. In contrast, nonlinear models RF and XRF, with Sharpe ratios of 1.07 and 1.01, outperform both linear models and the multi-factor Buy, whereas GBT and SVM, each with a Sharpe ratio of 0.88, surpass the performance of linear models.

Turning to deep learning models, NN records the lowest Sharpe ratio of 0.78, since its lack of economic structure causes the model to overfit on factor-specific noise. In contrast, MT earns a Sharpe ratio of 1.14 by incorporating a common structure across factors, surpassing the leading off-the-shelf model (RF) Sharpe ratio of 1.07. Additionally, DMT offers the highest Sharpe ratio of 1.26 by further incorporating time series dynamics. DMT also achieves the highest alpha of 1.68% (t-stat of 4.29) with respect to the Buy, surpassing MT and RF's alphas of 0.98% (t-stat of 3.16) and 0.8% (t-stat of 2.38), respectively. Even after subtracting large transaction costs of 14 basis points, DMT earns an impressive Sharpe ratio of 1.17 and alpha of 1.33% (t-stat of 3.42). These results highlight the economic gains from incorporating economic structure and time series dynamics into deep learning models for factor timing.

Third, we dissect the multi-factor strategy by studying the economic significance of machine learning predictions for each individual factor. We find that DMT consistently outperforms, achieving a higher Sharpe ratio than the Buy for every factor. DMT is also the best performing model for MKT and CMA. Particularly for market timing, DMT achieves an alpha of 4.78% (t-stat of 4.17) and Sharpe ratio of 0.89, respectively, significantly outperforming MKT's Sharpe ratio of 0.6. Even after incorporating large transaction costs of 14 basis points, DMT's market timing strategy earns a Sharpe ratio of 0.87 and alpha of 4.5% (t-stat of 3.94). Additionally, we find that nonlinear models generally outperform linear models for the investment and momentum factors, suggesting that incorporating nonlinear interactions is important for understanding the economic mechanisms driving these two factors.

Fourth, we quantify the importance of different predictors for multi-factor timing using Shapley values, which approximate changes in the model probability predictions had we excluded certain predictors in its estimation. Models generally agree on the most important variables across factors. Influential financial variables include tail risk beta; price trends (return seasonality years 11 to 15, momentum based on ff3 residuals, industry return of big

firms); and accounting variables in the leverage (organizational capital, industry concentration, composite debt issuance), value (net payout yield, equity duration, book to market), and profitability (earnings consistency, dividend omission, earnings surprise of big firms) categories. Influential macroeconomic variables fall into the money (total reserves of depository institutions, real estate loans at all commercial banks, consumer motor vehicle loans outstanding), output (ip: nondurable consumer goods, ip: residential utilities), labor (civilians unemployed for 15-26 weeks, average duration of unemployment), and inflation (cpi: apparel, ppi: metals and metal products, ppi: finished goods, personal cons. exp: durable goods) categories.

Finally, we analyze DMT's most influential variables for each factor, revealing heterogeneity across factors. For MKT, Industrial Production (IP) related to nondurable consumer goods emerges as the most influential predictor. Additionally, we find that price trends, as well as variables from every accounting category, play a significant role in market timing. In contrast, for SMB, the most influential variables include tail risk beta and those in the leverage and liquidity categories, reflecting the key vulnerabilities of small firms. Accounting-based factors HML, RMW, and CMA possess a common set of influential variables, particularly those in the leverage, profitability, value, and investment categories. Finally, the influential variables for MOM predominantly include tail risk and price trends. However, we discover that variables in the value, leverage, profitability, and inflation categories also significantly influence the momentum factor. Notably, we find that many asset pricing anomalies provide significant predictive power for factor timing, in agreement with Dong et al. (2022) but contrasting the results of Cakici et al. (2023) and Engelberg et al. (2023). Furthermore, the substantial predictive power of financial variables is enhanced by models that account for nonlinear interactions. For example, we show pairwise nonlinear interactions for each factor, elucidating how the superior performance of DMT is attributable to its ability to capture these complex functional forms.

In this paper, we develop economically motivated deep learning models for factor timing,

contributing to at least two existing strands of literature. Firstly, we build upon the extensive literature that investigates factor predictability through linear models. Market predictability is a central topic in financial economics. Recent comprehensive reviews by Koijen and Van Nieuwerburgh (2011) and Rapach and Zhou (2013) underscore the depth and breadth of research in this domain. Various papers extend the ideas of market predictability to specific factors. For instance, the predictability of the value factor is explored by Baba Yara et al. (2021) and Cohen et al. (2003). Similarly, the predictability of the momentum factor is studied in Cooper et al. (2004) and Daniel and Moskowitz (2016).

Several recent studies delve into multi-factor predictability. Greenwood and Hanson (2012) employ corporate share issuance to predict multiple factors. Moreira and Muir (2017) adopt volatility approaches for multi-factor timing. Haddad et al. (2020) and Kagkadis et al. (2023) harness dimension reduction techniques for their factor predictions. Furthermore, Gupta and Kelly (2019) and Moskowitz et al. (2012) document factor persistence. A common theme among these studies is their reliance on linear models, a limited set of predictors, and a lack of time series dynamics. In contrast, we introduce the DMT model that adeptly captures time series dynamics, nonlinear interactions, a wide range of predictors, and commonalities across factors.

Second, we contribute to the rapidly expanding literature that uses machine learning techniques to forecast stock returns. In their seminal work, Gu et al. (2020) employ an extensive array of off-the-shelf machine learning techniques to forecast stock portfolio returns. We benchmark our analysis against their top-performing model based on a deep neural network with three hidden layers. Our findings demonstrate that incorporating a common structure across factors and time series dynamics yields superior predictions compared to all off-the-shelf approaches.

Several recent papers introduce machine learning models with economic restrictions to improve forecasts. Bryzgalova et al. (2023), Chen et al. (2023), Feng et al. (2018), Gu et al. (2021), and Kozak et al. (2020) develop machine learning techniques with the economic

restriction of no-arbitrage for predicting individual stock returns. Guijarro-Ordonez et al. (2021) develop a deep learning approach for statistical arbitrage. Liu (2023) develops multi-task neural networks that incorporate an economic restriction based on a common structure across return quantiles, which enhances the accuracy of quantile forecasts, resulting in significant statistical and economic gains. Proner (2023) introduces dynamic multi-task neural networks that jointly forecast inflation and unemployment, incorporating an economic restriction that is motivated by the Phillips Curve. Our paper complements this literature by developing a deep learning approach with economic restrictions, tailoring it specifically for the four main challenges of factor timing.

The rest of the paper is organized as follows. Section 2 lays out the model framework. We detail the data and estimation procedures in Section 3. Section 4 presents the main empirical results. Finally, Section 5 offers concluding remarks.

2 Methodology

We are interested in predicting a factor's probability of outperforming or underperforming. We use a classification approach for factor timing for two main reasons. First, the probability prediction naturally translates into an intuitive trading strategy. Second, factors exhibit heavy tails (Arnott et al. (2019), Daniel and Moskowitz (2016)), which can be detrimental for regression approaches, but has no effect on classification approaches.

In its most general form, we describe a factor's conditional probability of a positive excess return as

$$\pi_{i,t+1} \equiv P_t(r_{i,t+1} > 0) = g_i(x_t), \tag{1}$$

where factors are indexed by $i = 1, ..., N$, months by $t = 1, ..., T$, and the excess return of factor i in month $t + 1$ is denoted as $r_{i,t+1}$. Our objective is to estimate a functional form $g_i(\cdot)$, which links the P-dimensional vector of predictors x_t to the probability of a positive

excess return $\pi_{i,t+1}$. To achieve this objective, we introduce two multi-task neural network architectures that jointly model the $\pi_{i,t+1}$ across all N factors within a single functional form and incorporate several other economically motivated restrictions on g_i, which are detailed in the next sections.

2.1 Multi-task Neural Network

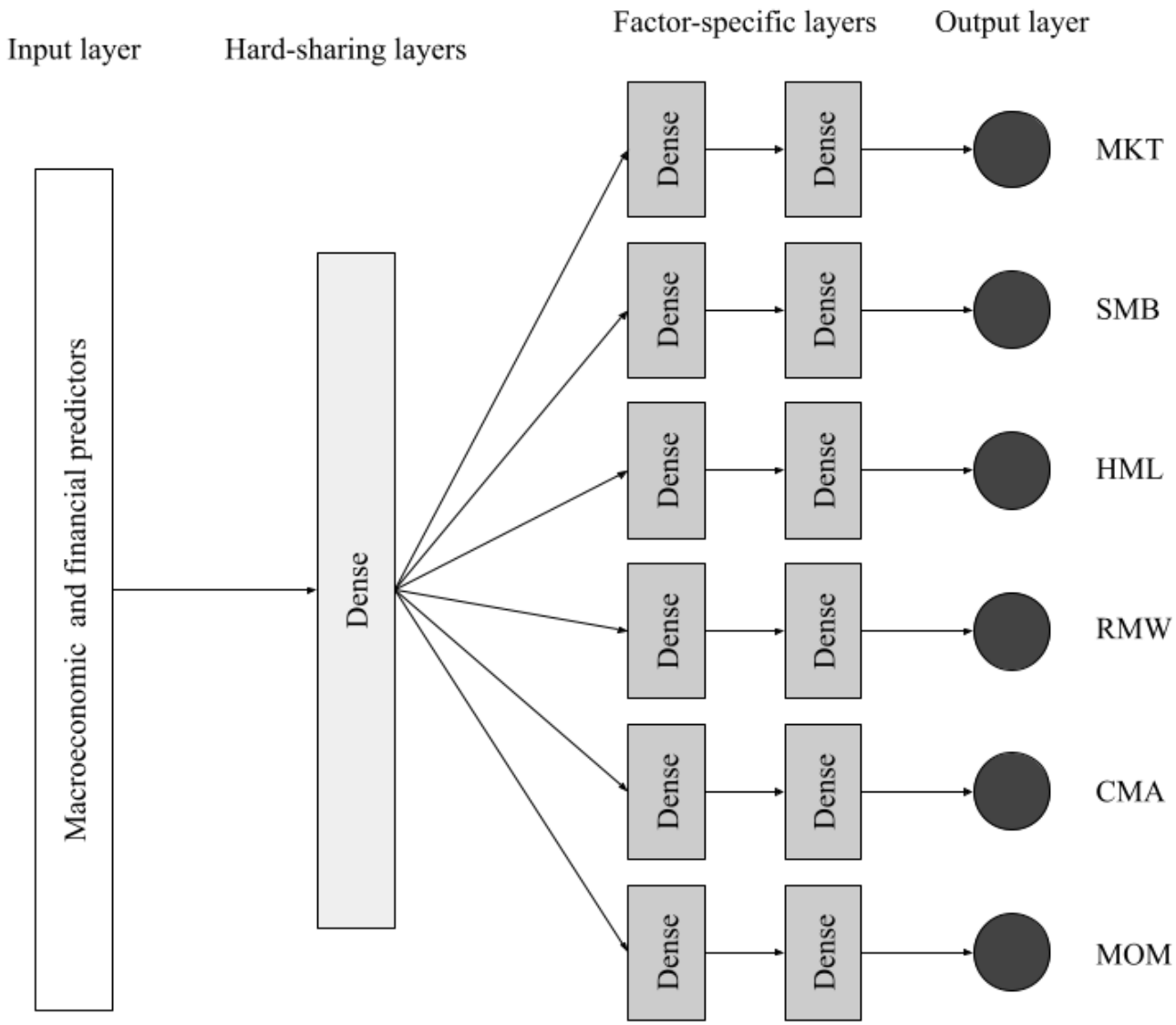
Factors share a low-dimensional common structure (Haddad et al. (2020), Kagkadis et al. (2023)). We incorporate this economic structure with a multi-task neural network architecture. Figure 1 shows an example of a MT. The "hard sharing" hidden layers interact and nonlinearly transform the input predictors into shared latent features, imposing an economically motivated restriction of a low-dimensional common structure across factors. The "factor-specific" layers capture nonlinear exposures to these shared latent features, subsequently aggregating these into a final prediction of $\pi_{i,t+1}$ for each factor.

MT employs MTL, which is a form of regularization that reduces overfitting to factor-specific noise, especially in low signal-to-noise environments. Simultaneous prediction of multiple factors forces the neural network's hard-sharing parameters to be sufficiently versatile, providing signals for most or all factors. This simultaneous learning of related tasks, known as inductive transfer, improves generalizability, as what is learned for each factor can help other factors be learned better (Ruder, 2017). Consequently, MTL increases the effective sample size, due to the extra information contained in the training signals of related tasks (Caruana, 1997).

2.2 Dynamic Multi-task Neural Network

Factors are time-varying and depend on short- and long-term macroeconomic and financial conditions (Ilmanen et al. (2021), Hodges et al. (2017), Polk et al. (2020)). MT and off-the-shelf models only incorporate predictor values in the preceding period, which is insufficient for the model to learn long-term business and financial cycle dynamics. To address this,

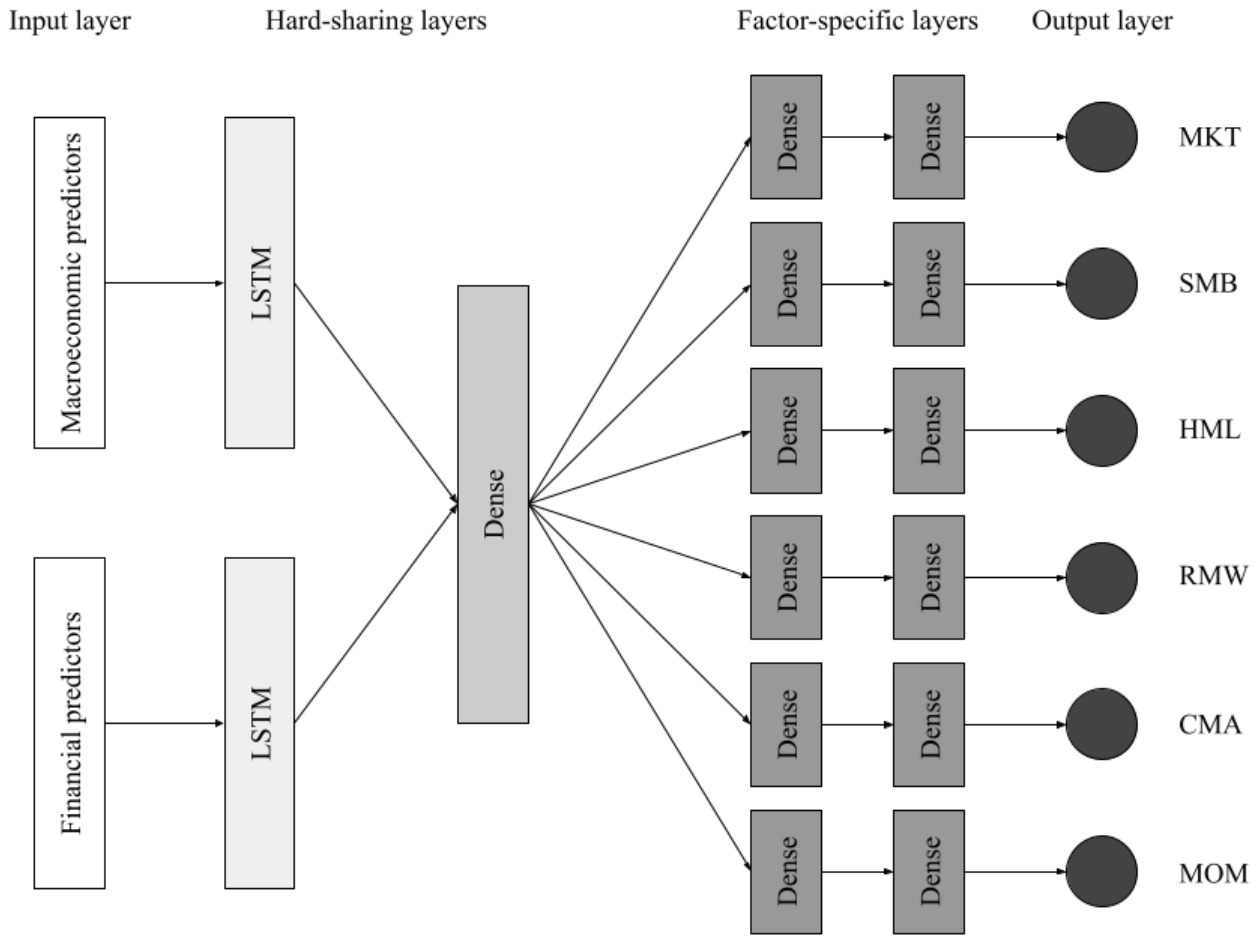
Figure 1: Multi-task Neural Network Example



Macroeconomic and financial predictors are input into a fully connected (dense) layer, where they are nonlinearly interacted. Next, the network splits into branches to learn factor-specific loadings and predict probabilities for each factor.

following Chen et al. (2023), we incorporate long-term nonlinear dependencies using LSTMs (Hochreiter and Schmidhuber, 1997). LSTMs allow for complex nonlinear interactions between the dynamics of the macroeconomic and financial time series, rather than just the values of the stationary time series at time t.

Figure 2: Dynamic Multi-task Neural Network Example



Macroeconomic and financial predictors are input into the network separately and propagate through LSTM layers constructing nonlinear dynamic features. These nonlinear dynamic macroeconomic and financial features are then nonlinearly interacted in a fully connected (dense) layer. Finally, the network splits into branches to learn factor-specific loadings and predict probabilities for each factor.

The DMT architecture, as illustrated in Figure 2, integrates memory by employing two separate LSTM layers for macroeconomic and financial variables, positioned before the first fully connected (dense) layer. The top LSTM nonlinearly interacts the macroeconomic time series inputs to estimate a small number of hidden macroeconomic state variables. Likewise, the bottom LSTM nonlinearly interacts the financial time series inputs to estimate a small number of hidden financial state variables. Drawing on Chen et al. (2023), these LSTM layers extract business and financial cycle dynamics through their ability to identify cyclical

patterns. The need to independently capture business and financial cycle dynamics drives our economic restriction to separate macroeconomic and financial input variables in the LSTM layer. A subsequent hard sharing layer nonlinearly interacts the high-level macroeconomic and financial hidden state variables, producing a low-dimensional set of shared latent features for the multi-task network.

Given the high-dimensionality and strong cross-sectional dependence of financial and macroeconomic input variables, there's overlapping information that can be captured by a low-dimensional model. Functionally, our LSTM layers serve to reduce dimensions, akin to principal component analysis (PCA), while also extracting dynamics akin to a state space model within a broader nonlinear framework. By consolidating the high-dimensional inputs into a small number of hidden state processes, we prevent the model from overfitting on the noise specific to each macroeconomic or financial time series. Internet Appendix IA1.2.5 provides a detailed description of the LSTM architecture.

2.3 Off-the-shelf Models

We consider a variety of off-the-shelf machine learning models from Hastie et al. (2009) as comparative benchmarks. These models estimate a separate functional form g_i for each factor i and can be highly susceptible of overfitting to factor-specific noise. Additionally, they do not leverage the common structure across factors. Off-the-shelf models studied in this paper include linear models (logistic regression and penalized logistic regression) and nonlinear models (random forest, extremely randomized trees, gradient boosted trees, support vector machines, and a feed-forward neural network with three hidden layers). These models are described in detail in Internet Appendix IA1.

3 Data and Estimation Procedure

3.1 Data

Our factor sample spans January 1965 to December 2021, totaling 57 years. We use six monthly factors from Kenneth French's website as response variables: MKT, SMB, HML, RMW, CMA, and MOM. The first three factors correspond to the excess market return, size, and value from Fama and French's original three-factor model (Fama and French, 1996). The following two factors indicate profitability and investment from their five-factor model (Fama and French, 2015), while the last factor represents momentum, which goes long on past winners and short on past losers.

Additionally, we include macroeconomic variables from the FRED-MD database, as detailed in McCracken and Ng (2016), which correspond to categories such as labor, output, and inflation. We apply their transformations to generate stationary time series. After removing variables with missing values before 1990, corresponding to the beginning of the out-of-sample period, we have 123 macroeconomic predictors. Following Bianchi et al. (2021) and Chen et al. (2023), we lag the macroeconomic variables by one additional month (i.e., we use the observation at $t-1$) to account for announcement delays. See Table IA3 in the Internet Appendix for a comprehensive list of macroeconomic variables.

We obtain 149 financial variables from two sources. First, motivated by Dong et al. (2022), we include "anomalies" from the asset pricing literature as predictors, specifically referring to long-short portfolios of individual stocks. We include the anomalies detailed in Chen and Zimmermann (2022), and categorize variables loosely based on Gu et al. (2020) and Jensen et al. (2023), encompassing investment, liquidity, price trend, profitability, quality, risk, and value.[1] We remove predictors with any missing values before 1990, corresponding to the beginning of the out-of-sample period, resulting in 137 predictors, and fill the remaining missing values with the expanding training set mean. Second, we include 12 aggregate

[1]The monthly data are available at www.openassetpricing.com. Our data is collected from the August 2023 Release.

predictors from Welch and Goyal (2008) based on stock and bond markets. Stock market variables include the book-to-market ratio, dividend price ratio, dividend yield, earnings price ratio, dividend payout ratio, stock variance, and net equity expansion; and bond market variables include the t-bill rate, long term yield, long term rate of return, term spread, default yield spread.[2] See Table IA2 in the Internet Appendix for a descriptive list of the financial variables.

3.2 Estimation Procedure

To estimate the models, we use the standard validation set approach and divide our full sample (January 1965 to December 2021) into three disjoint time periods. We start by estimating the model parameters on a training sample of 20 years (1965 - 1984). We then perform an extensive hyperparameter optimization, validating the model's fit in the next five years (1985 - 1989). Lastly, we assess the predictive power in the one-year testing sample (1990). We keep the models fixed for one year and replicate this procedure extending the number of years in the training sample by one year in each iteration, for a total of 32 out-of-sample years (1990 - 2021). In the training, validation, and test sets, we standardize each predictor using its mean and variance from the training set. Section IA3 in the Internet Appendix details the hyperparameter optimization setup.

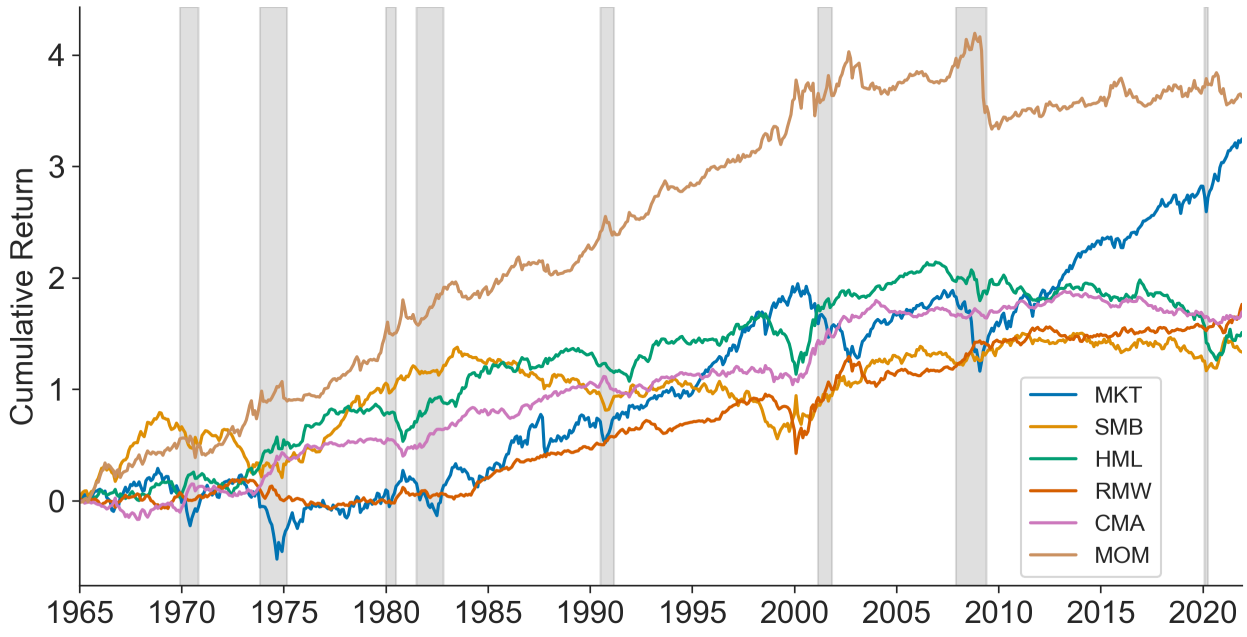
4 Empirical Results

4.1 Time-varying Factor Risk Premia

Figure 3 displays the cumulative returns of the six factors across the entire sample. All factors gain considerable risk premia, with cumulative returns exceeding 100%. SMB fluctuates with periods of high and low returns, resulting in the lowest cumulative return. HML fares well before the financial crisis, but a significant post-crisis drawdown lands it with the second

[2]The monthly data are available at Amit Goyal's website.

Figure 3: Cumulative Returns of Each Factor



The figure illustrates the cumulative log returns for each factor across the entire sample period from January 1965 to December 2021, with NBER recessions shaded in grey.

lowest cumulative return. CMA and RMW exhibit similar cumulative returns with occasional major drawdowns. MKT boasts the second highest cumulative return, due to a high risk premia period after the Global Financial Crisis (GFC). Despite a severe crash during the GFC, MOM still achieves the highest cumulative return. Notably, factors respond differently to adverse financial and macroeconomic conditions, as indicated by NBER recessions.

From January 1990 onward, our out-of-sample period, MOM and MKT demonstrate superior performance, indicating an ex-post tilt toward these factors is beneficial in a multi-factor strategy. Conversely, the underperformance of SMB and HML demonstrates an ex-post tilt away from these factors is advisable. Notably, all factors except MKT yield subpar returns after the GFC, indicating an ex-post tilt away from all but the market factor. The subsequent sections demonstrate that, through the incorporation of economic structure and time series dynamics, DMT acquires the capability to execute these tilts ex-ante, ultimately leading to significant statistical and economic gains.

Table 1: Out-of-sample Prediction Performance

	Buy	LR	EN	RF	XRF	GBT	SVM	NN	MT	DMT
Mean	55.3	53.9	54.1	56.8	55.9	54.6	53.5	52.3	55.7	57.2
MKT	64.3	57.0	60.4	62.5	64.3	57.6	59.4	57.8	63.8	66.1
SMB	51.6	54.9	54.2	55.7	54.7	53.9	53.4	49.7	56.2	52.3
HML	47.9	54.9	53.9	54.7	50.3	52.6	51.0	46.6	51.8	51.6
RMW	58.3	51.3	51.8	55.7	57.0	55.5	54.7	53.4	54.2	59.1
CMA	49.2	50.8	45.6	51.3	48.2	52.3	45.3	49.5	47.9	52.9
MOM	60.4	54.7	58.6	60.9	60.7	56.0	57.3	56.5	60.4	60.9

This table displays the accuracy (in percentage), defined as the percentage of total observations correctly classified for the out-of-sample period January 1990 to December 2021 for all models and the buy-and-hold benchmark (Buy) that always predicts up. The first row consolidates the findings by calculating the average accuracy across all factors, while the following rows individually detail the accuracy for each respective factor.

4.2 Out-of-sample Predictive Power

We evaluate the predictive power of the machine learning models using the out-of-sample predictive accuracy metric, defined as the proportion of correctly classified positive excess returns, i.e., up months, within the test set.[3] Table 1 reports the accuracy (in percentage) of the models for the out-of-sample period from January 1990 to December 2021. We compare nine models, including linear models LR, EN; tree-based models RF, XRF, GBT; kernel-based model SVM; and deep learning models NN, MT, and DMT. As a naive benchmark, we also report the accuracy of the Buy that always predicts up.

The first row of Table 1 reports the average accuracy across all six factors. LR, which utilizes all predictors, offers a low accuracy of 53.9%, falling short of the 55.3% accuracy from the Buy. This is expected, given that LR lacks regularization and is prone to overfitting. The accuracy improves to 54.1% when LR is limited to a sparse parameterization with EN. Accuracy is further boosted to 56.8%, 55.9%, and 54.6% with the incorporation of nonlinear interactions through tree-based models RF, XRF, and GBT. In contrast, kernel-based model,

[3]We report the accuracy metric given that the response is generally balanced. Other metrics, such as the F1-score, produce similar results. However, our findings indicate a stronger correlation between accuracy and the Sharpe ratio compared to other metrics such as the F1-score.

SVM, exhibits a lower accuracy of 53.5%, likely because it's too flexible and overfitting to factor-specific noise.

Turning to the deep learning models, NN, lacking economic structure, posts the lowest accuracy of 52.3%, due to its propensity to overfit on individual factors. However, with the imposition of an economically motivated restriction of a common structure across factors, MT raises the accuracy to 55.7%, suggesting that the regularization and data efficiency benefits from MTL improve forecasting accuracy. Furthermore, by incorporating time series dynamics, DMT delivers the highest accuracy of 57.2%, outperforming the Buy.

Subsequent rows of Table 1 display the accuracy of the nine models for each factor. DMT is the only model that achieves a higher accuracy than the Buy for every factor, underscoring its consistency from overcoming the challenges of factor timing. DMT is also the best performer in four of the six factors. The second row reveals that the market goes up in 64.3% of the months, providing a high benchmark. DMT achieves an accuracy of 66.1% for MKT, significantly outperforming the Buy and best off-the-shelf model (XRF), both of which post accuracies of 64.3%. For RMW in the fifth row, DMT achieves an accuracy of 59.1%, outperforming the Buy accuracy of 58.3% and best off-the-shelf model (XRF) accuracy of 57%. DMT's accuracy of 52.9% for the CMA factor in the sixth row outperforms the Buy accuracy of 49.2% and best off-the-shelf model (GBT) accuracy of 52.3%. DMT also performs well for MOM in the seventh row, beating the Buy accuracy of 60.4% and matching the best off-the-shelf model (RF) accuracy of 60.9%.

MT achieves the highest accuracy of 56.2% for SMB in the third row, outperforming the Buy accuracy of 51.6% and DMT's accuracy of 52.3%. For HML in the fourth row, MT and DMT's accuracies of 51.8% and 51.6%, respectively, outperform the Buy accuracy of 47.9%. However, LR and RF achieve the highest accuracies of 54.9% and 54.7%, respectively, indicating that certain off-the-shelf models can effectively predict the value factor.

To make pairwise comparisons of models, we use a modified Diebold and Mariano (1995) (DM) test that accounts for the potential cross-sectional dependence of factors. Given our

Table 2: Comparison of Probability Predictions Using Diebold-Mariano Tests

	EN	RF	XRF	GBT	NN	MT	DMT
LR	**13.02**	**12.65**	**12.94**	**11.02**	**8.06**	**12.02**	**13.28**
EN		**6.34**	**6.86**	**4.00**	-0.06	**5.47**	**7.59**
RF			**3.56**	**-4.30**	**-5.39**	**-2.80**	**2.97**
XRF				**-5.56**	**-5.99**	**-4.13**	**2.07**
GBT					**-3.47**	1.02	**5.29**
NN						**4.84**	**5.93**
MT							**4.62**

This table presents test statistics from pooled Diebold-Mariano tests. We perform pairwise comparisons of model probability forecasts, using the the average log loss across factors. A positive statistic indicates the column model outperforms the row model, and a bold number denotes significance at the 5% level for each test. We omit SVM, since it does not output probabilities.

focus on classification, we use the cross-sectional average log loss to compare the probability forecasts of our models.[4] The DM test statistic for a comparison of models 1 and 2 is defined as $DM_{12} = \bar{d}_{12}/\sigma_{\bar{d}_{12}}$, where $\bar{d}_{12}$ and $\sigma_{\bar{d}_{12}}$ are, respectively, the mean and Newey-West standard error of $d_{12,t+1}$ over the test sample. $d_{12,t+1}$ is the forecasting error between the two models, calculated as the cross-sectional average of log loss differentials from each model over each period $t+1$,

$$d_{12,t+1} = \frac{1}{N}\sum_{i=1}^{N}(\hat{e}^{(1)}_{i,t+1} - \hat{e}^{(2)}_{i,t+1}), \tag{2}$$

where $\hat{e}^{(j)}_{i,t+1} = y_{i,t+1}\log(\hat{\pi}^{(j)}_{i,t+1}) + (1-y_{i,t+1})\log(1-\hat{\pi}^{(j)}_{i,t+1})$ denotes the log loss of model j for factor i at time $t+1$, and $y_{i,t+1} \equiv I(r_{i,t+1} > 0)$ is an indicator function that takes a value of one if factor i's excess return is positive and zero otherwise.

Table 2 presents the DM test statistics across the models. A positive statistic indicates the column model outperforms the row model, and a bold number denotes significance at the 5% level. The first row shows a positive and statistically significant test statistic for all models with DM test statistics ranging from 8.06 to 13.28, indicating that all models significantly

[4]SVMs are excluded from this analysis since they do not provide probabilities. Although Platt scaling can estimate probabilities from SVM predictions, it may yield estimates contradicting the SVM classifications. Hence, we omit SVM from the DM tests.

improve over LR. Additionally, all models significantly outperform EN, except for NN. In contrast, the last column shows DM test statistics ranging from 2.07 to 13.28, showing that DMT significantly outperforms all other models. The second last column indicates that MT significantly outperforms EN and NN, but is significantly outperformed by RF and XRF, illustrating that tree-based models can perform well at forecasting probabilities.

4.3 Multi-factor Timing

Next, we evaluate the economic significance of factor timing strategies derived from each model's forecasts, using a multi-factor portfolio. Motivated by Campbell and Thompson (2008), we apply realistic constraints that restrict short-selling and leverage. For every factor, we take a long position if the model forecasts a positive excess return, otherwise we invest at the risk-free rate, yielding zero excess returns. The strategy return for factor i at time $t+1$ is expressed as

$$r_{i,t+1}^{Strategy} = I(\hat{\pi}_{i,t+1} > 0.5) r_{i,t+1}, \tag{3}$$

where $I(\cdot)$ is an indicator function that takes a value of one if the predicted probability exceeds 50% and zero otherwise. Our multi-factor strategy is then an equal-weighted portfolio of the strategy excess returns across all factors. Motivated by Moreira and Muir (2017), for each model we conduct a time series regression of the strategy on the multi-factor Buy that is an equal-weighted average of the excess returns across all factors. This spanning regression provides the strategy's alpha, beta, and R^2 relative to the multi-factor Buy.

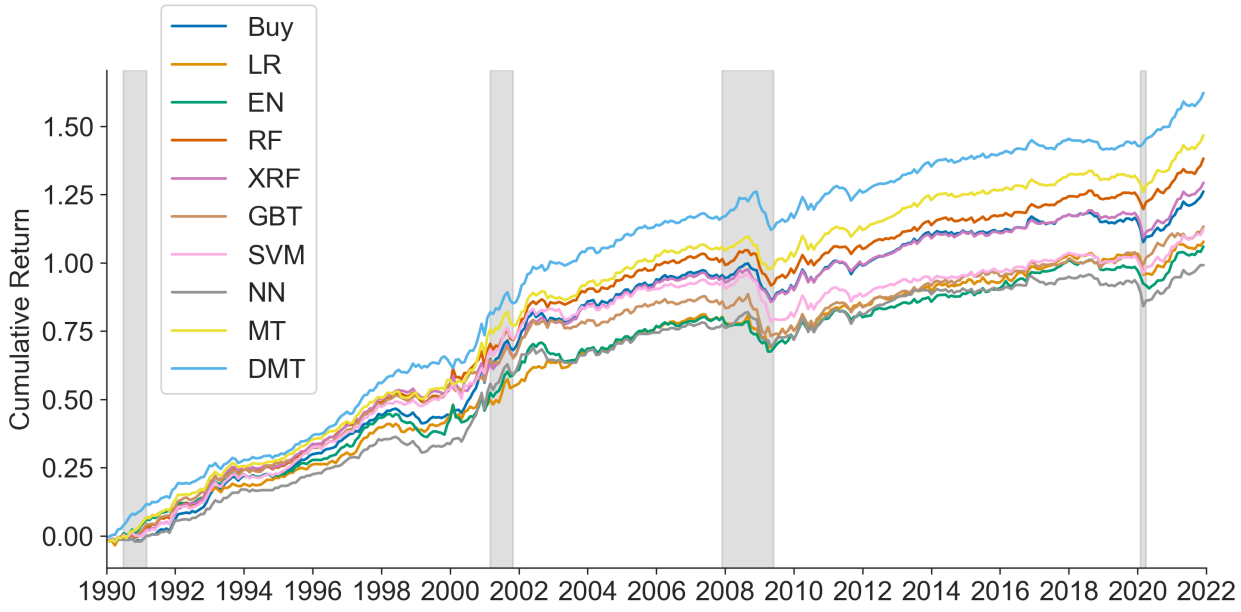
The first row of Table 3 presents the annualized Sharpe ratio for each model, which aligns very closely with the results on the average accuracy reflected in the first row of Table 1. The linear models, LR and EN, yield Sharpe ratios of 0.84 and 0.83, respectively, underperforming the multi-factor Buy Sharpe ratio of 0.98. However, improvements over linear models are seen with the incorporation of nonlinear interactions in RF, XRF, GBT,

Table 3: Multi-factor Timing Strategy Performance

	Buy	LR	EN	RF	XRF	GBT	SVM	NN	MT	DMT
SR	0.98	0.84	0.83	1.07	1.01	0.88	0.88	0.78	1.14	1.26
α		0.56	0.20	0.80	0.43	0.53	0.11	-0.49	0.98	1.68
$t(\alpha)$		1.21	0.47	2.38	1.45	1.20	0.32	-1.77	3.16	4.29
β		0.60	0.72	0.86	0.89	0.66	0.79	0.86	0.81	0.75
R^2		48	62	79	84	57	74	85	80	68

This table presents the performance of the multi-factor timing strategies by model. Reported are the annualized Sharpe ratio (SR), along with the alpha (α), alpha t-statistic (t(α)), beta (β), and percentage R^2 with respect to the multi-factor buy-and-hold benchmark (Buy).

Figure 4: Cumulative Returns of Multi-factor Timing



This figure plots the cumulative log returns of multi-factor timing strategies by model over the out-of-sample period from January 1990 to December 2021. Each model's returns are scaled to have the same volatility as the multi-factor buy-and-hold benchmark (Buy). NBER recessions are shown in grey.

SVM models, resulting in Sharpe ratios of 1.07, 1.01, 0.88, and 0.88 respectively. Notably, RF and XRF are the only off-the-shelf models to achieve a higher Sharpe ratio than the multi-factor Buy.

Turning to deep learning models, NN posts the lowest Sharpe ratio of 0.78, due to its

lack of economic structure causing the model to overfit to factor-specific noise. However, by imposing economic structure, MT achieves a Sharpe ratio of 1.14, which is higher than all off-the-shelf models and the multi-factor Buy. Additionally, DMT, which further incorporates time series dynamics, achieves the highest Sharpe ratio of 1.26, surpassing the Sharpe ratios of the multi-factor Buy by 29% and leading off-the-shelf model (RF) by 18%. This is a larger improvement than the six factor volatility-managed portfolio in Moreira and Muir (2017), which records a 6% Sharpe ratio improvement over the multi-factor Buy. Our Sharpe ratio of 1.26 also greatly exceeds the Sharpe ratios of 0.71 and 0.73 from the PCA models of Haddad et al. (2020) and Kagkadis et al. (2023), respectively.

The second row of Table 3 displays the annualized alpha earned by each model. Only RF, MT, and DMT achieve alphas that are statistically significant at the 5% level. RF earns an alpha of 0.8% (t-stat of 2.38), which is the best risk-adjusted performance by an off-the-shelf model. NN posts a negative alpha of -0.49% (t-stat of -1.77) due to its propensity to overfit. However, MT earns an alpha of 0.98% (t-stat of 3.16) from imposing economic structure, and DMT boosts the alpha to 1.68% (t-stat of 4.29) from further incorporating time series dynamics. DMT also records a low beta of 0.75 and an R^2 of only 68%, suggesting that the model is capable of avoiding months with factor losses. These results underscore DMT's efficacy in generating superior risk-adjusted returns, underscoring the importance of incorporating economic structure and time series dynamics into deep learning models.

Figure 4 illustrates the cumulative returns of each model and the multi-factor Buy, where each model's returns are scaled to have the same volatility as the multi-factor Buy. RF and XRF earn the highest cumulative returns among the off-the-shelf models, surpassing the multi-factor Buy, but noticeably underperforming DMT and MT. LR, EN, GBT, NN and SVM share similar cumulative returns and underperform the multi-factor Buy.

Examining deep learning models, DMT achieves the highest cumulative return throughout the entire sample, with a minor drawdown during the GFC. Interestingly, DMT performs well in the post-GFC sample and avoids the large drawdown experienced by the multi-factor

Buy and other models during the Covid crisis. MT exhibits the second highest cumulative return, and also performs well in the post-GFC sample. In contrast, NN exhibits the lowest cumulative return due to its propensity to overfit to the noise of individual factors. Taken together, these results demonstrate that incorporating economic structure and time series dynamics is necessary for factor timing with deep learning.

4.4 Single-factor Timing

Next, we break down the multi-factor timing performance of each model by analyzing the performance of strategy (3) for each individual factor. Table 4 reports each model's annualized Sharpe ratio, along with the alpha, alpha t-statistic, beta, and R^2 from a time series regression of the strategy on the respective single-factor Buy. The first panel shows that none of the off-the-shelf models earn a higher Sharpe ratio than MKT's Sharpe ratio of 0.6. However, DMT stands out by significantly outperforming the market with a Sharpe ratio and alpha of 0.89 and 4.78% (t-stat of 4.17), respectively. This demonstrates that incorporating economic structure and time series dynamics is important for market timing. MT earns the same Sharpe ratio as MKT, with a positive but insignificant alpha of 0.11% (t-stat of 0.25). XRF is the best-performing off-the-shelf model, but its strategy is identical to the market as indicated by its beta of 1. All other off-the-shelf models exhibit lower Sharpe ratios than MKT and negative alphas.

The second panel shows that all models, except NN, beat SMB's lackluster Sharpe ratio of 0.16. RF and LR are the best performers with Sharpe ratios of 0.38 and 0.36, respectively, with statistically significant alphas of 2.19% (t-stat of 2.55) and 1.88% (t-stat of 2.03). Additionally, MT and DMT earn Sharpe ratios of 0.35 and 0.22, respectively, suggesting that the benefits of MTL are somewhat limited for the size factor.

The third panel shows that all models, except XRF, SVM, and NN, outperform HML's low Sharpe ratio of 0.11. This suggests that the flexibility of these nonlinear models leads to overfitting on the value factor. Interestingly, LR is the best model with a Sharpe ratio of

Table 4: Single-factor Timing Strategy Performance

Factor		Buy	LR	EN	RF	XRF	GBT	SVM	NN	MT	DMT
	SR	0.60	0.46	0.48	0.57	0.60	0.41	0.48	0.47	0.60	0.89
	α		-0.87	-0.94	-0.31	0.00	-2.00	-1.21	-1.27	0.11	4.78
MKT	$t(\alpha)$		-0.78	-0.95	-1.02	0.00	-2.12	-1.42	-1.47	0.25	4.17
	β		0.79	0.85	0.99	1.00	0.87	0.89	0.89	0.97	0.75
	R^2		79	84	99	100	86	89	89	97	76
	SR	0.16	0.36	0.31	0.38	0.30	0.30	0.30	0.09	0.35	0.22
	α		1.88	1.56	2.19	1.52	1.34	1.50	-0.58	1.78	0.79
SMB	$t(\alpha)$		2.03	1.72	2.55	1.90	1.46	1.69	-0.90	1.91	0.86
	β		0.57	0.63	0.69	0.76	0.39	0.66	0.87	0.55	0.42
	R^2		57	63	69	76	39	66	87	55	42
	SR	0.11	0.26	0.21	0.24	0.05	0.12	0.11	0.02	0.24	0.21
	α		1.34	1.04	1.30	-0.42	0.29	0.13	-0.89	1.37	1.10
HML	$t(\alpha)$		1.40	1.09	1.41	-0.50	0.31	0.15	-1.50	1.71	1.28
	β		0.45	0.52	0.65	0.75	0.58	0.73	0.89	0.77	0.72
	R^2		45	53	65	75	58	72	89	78	72
	SR	0.49	0.45	0.39	0.48	0.52	0.55	0.54	0.34	0.39	0.54
	α		0.71	0.11	0.54	0.79	1.32	1.12	-1.02	-0.67	0.66
RMW	$t(\alpha)$		0.87	0.13	0.72	1.11	1.61	1.41	-2.02	-1.48	1.30
	β		0.47	0.60	0.70	0.74	0.51	0.62	0.90	0.92	0.89
	R^2		47	59	69	74	50	62	89	92	89
	SR	0.32	0.25	-0.02	0.30	0.29	0.28	0.07	0.27	0.30	0.47
	α		0.20	-1.42	0.03	-0.18	0.12	-1.28	-0.20	0.23	1.17
CMA	$t(\alpha)$		0.33	-2.30	0.06	-0.72	0.19	-2.40	-0.48	0.39	2.11
	β		0.40	0.59	0.83	0.96	0.66	0.77	0.87	0.67	0.73
	R^2		39	59	83	96	66	76	87	67	73
	SR	0.34	0.22	0.34	0.38	0.36	0.40	0.34	0.37	0.41	0.37
	α		-0.59	0.43	0.63	0.29	1.63	0.55	0.92	1.16	0.77
MOM	$t(\alpha)$		-0.42	0.40	1.85	1.17	1.17	0.48	0.78	2.11	0.88
	β		0.63	0.83	0.99	0.99	0.64	0.81	0.79	0.96	0.90
	R^2		62	83	99	99	64	81	79	96	90

This table presents the performance of the single-factor timing strategies by model. Reported are the annualized Sharpe ratio (SR), along with the alpha (α), alpha t-statistic (t(α)), beta (β), and percentage R^2 with respect to the single-factor buy-and-hold benchmark (Buy).

0.26, indicating that a simple linear specification outperforms nonlinear models for timing the value factor. MT and DMT also earn high Sharpe ratios of 0.24 and 0.21, respectively.

The fourth panel shows that RMW's Sharpe ratio of 0.49 is surpassed by XRF, GBT,

SVM, and DMT, which achieve Sharpe ratios of 0.52, 0.55, 0.54, and 0.54, respectively, suggesting that nonlinear approaches can work well for timing the profitability factor. Conversely, linear models LR and EN underperform RMW, yielding Sharpe ratios of 0.45 and 0.39, respectively.

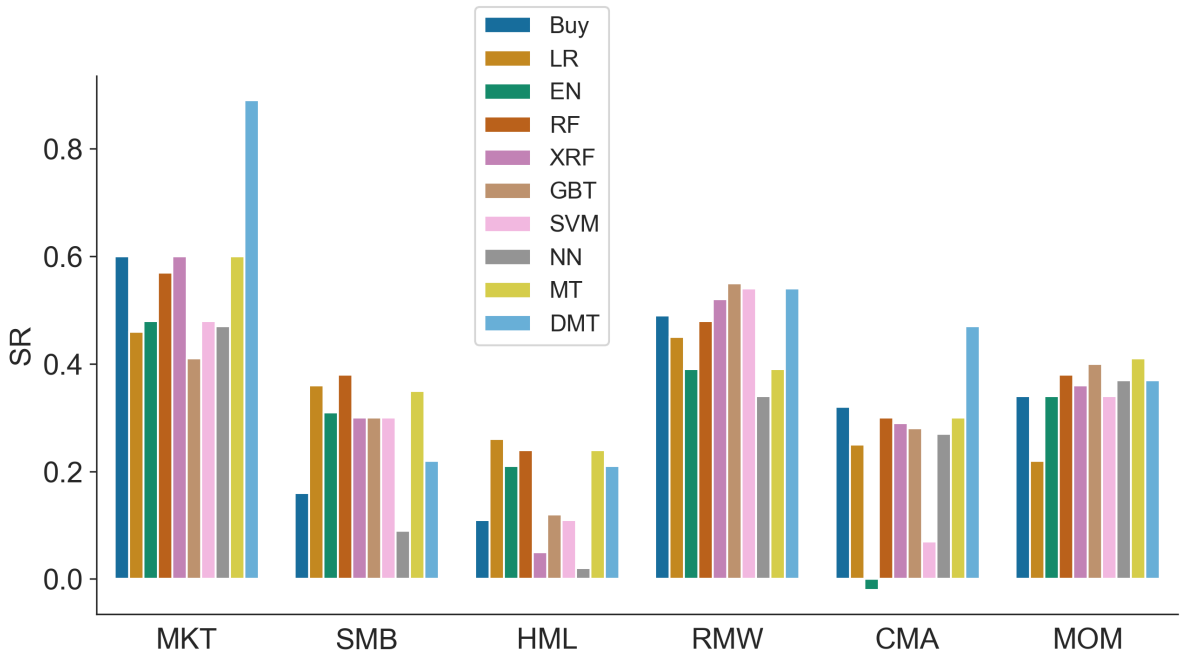
The fifth panel shows that DMT, with a Sharpe ratio of 0.47 and alpha of 1.17% (t-stat of 2.11), is the only model that outperforms CMA's Sharpe ratio of 0.32. MT and RF are the next best performers, which both earn Sharpe ratios of 0.3. In contrast, LR and EN exhibit lower performance compared to CMA, with Sharpe ratios of 0.25 and -0.02, respectively, highlighting the inadequacy of linear models for the investment factor.

The sixth panel reports that MT is the best performing model for MOM, reaching a Sharpe ratio of 0.41 and alpha of 1.16% (t-stat of 2.11). Tree-based models also perform well for the momentum factor, with RF, XRF, and GBT achieving Sharpe ratios of 0.38, 0.36, and 0.4, which surpasses MOM's Sharpe ratio of 0.34. Additionally, NN and DMT both record Sharpe ratios of 0.37. In contrast, LR and EN yield the lowest Sharpe ratios of 0.22 and 0.34, respectively.

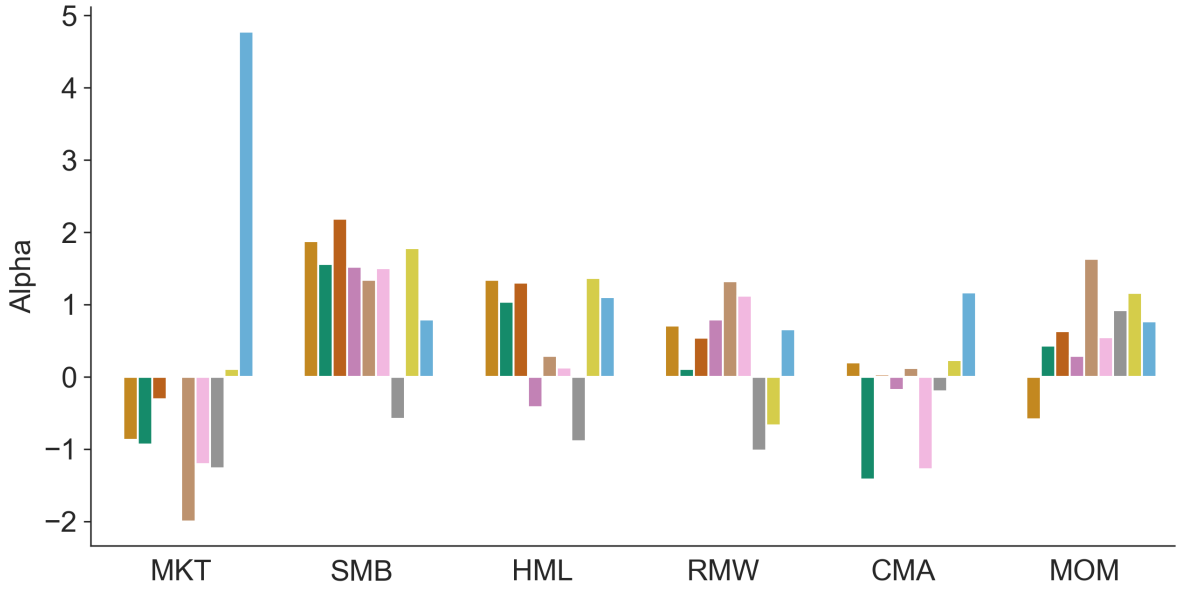
Notably, we find that nonlinear models outperform linear models for CMA and MOM, with the exception of SVM, which tends to be overly complex and often overfits to factor-specific noise. This finding underscores the importance of including nonlinear interactions for the investment and momentum factors. Such results indicate the presence of nonlinear economic mechanisms driving these factors, pointing to a complex functional form that warrants further investigation.

Figure 5 presents a comprehensive overview of the Sharpe ratios and alphas for each factor. Most models show enhanced performance in terms of Sharpe ratios and alphas compared to the single-factor Buy across various factors. Notably, there is considerable heterogeneity in off-the-shelf model performance across factors. In contrast, DMT consistently performs well, underscoring the advantages of incorporating MTL and time series dynamics. Remarkably, DMT is the only model that surpasses the single-factor Buy in every factor, showcasing

Figure 5: Risk-adjusted Returns by Factor and Model



(a) Sharpe Ratios



(b) Alphas

This figure presents the risk-adjusted returns of the single-factor timing strategies by model. Panel (a) shows the annualized Sharpe ratios (SR) for each model, while Panel (b) displays the alphas (α) with respect to the single-factor buy-and-hold benchmark (Buy).

exceptional performance in MKT and CMA.

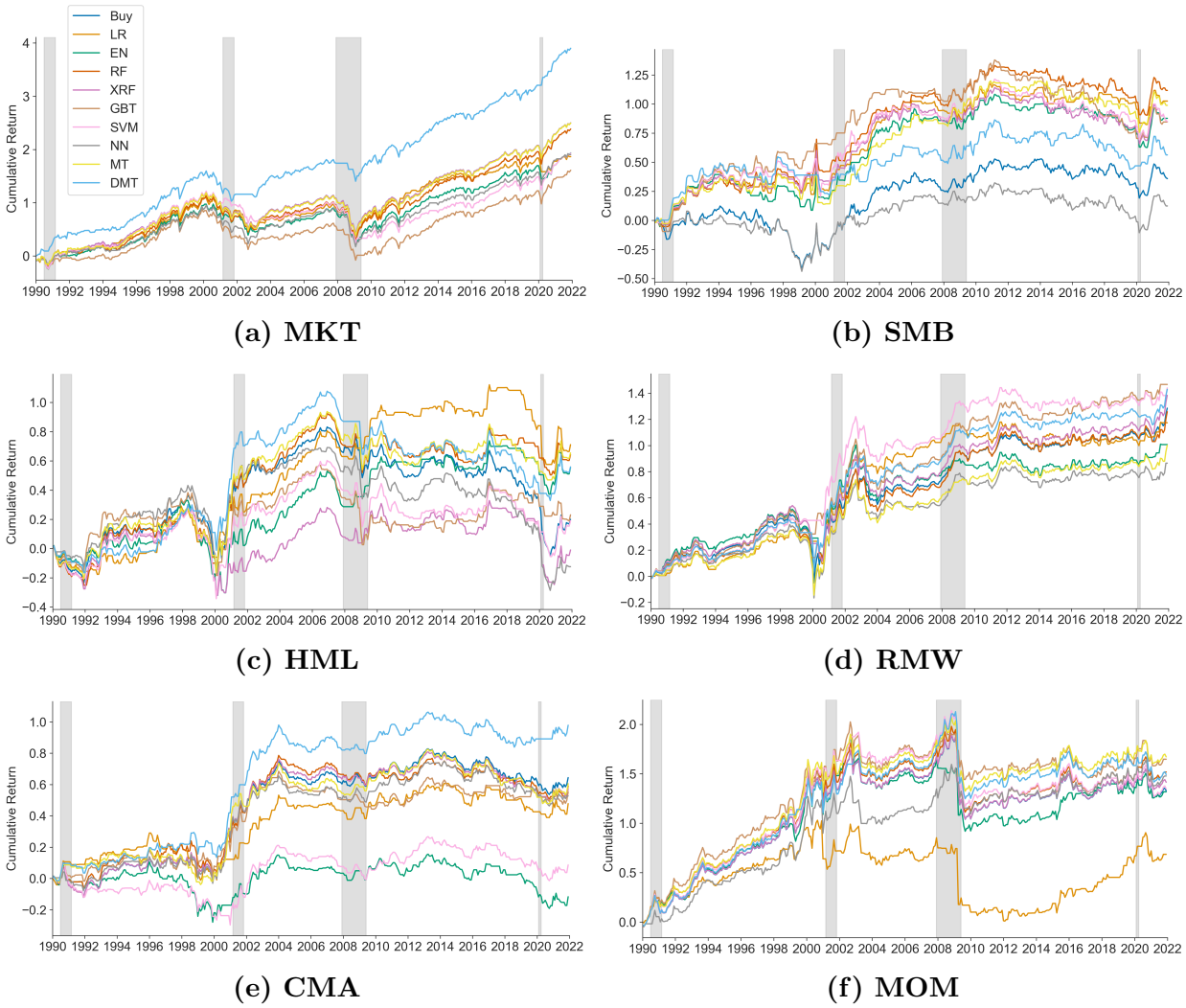
Quantitatively, DMT improves the Sharpe ratio relative to the single-factor Buy by 0.29, 0.06, 0.1, 0.05, 0.15, and 0.03 for the MKT, SMB, HML, RMW, CMA, and MOM factors, respectively. In comparison, the top-performing model from Gu et al. (2020) (NN3) enhances the Sharpe ratio over the single-factor Buy by 0.26, 0.15, 0.04, 0.01, 0.05, and -0.04 for these factors. This comparison underscores DMT's superior factor timing capabilities relative to the best model in Gu et al. (2020), except in the case of SMB. DMT's superior performance can be largely attributed to its incorporation of economic structure and time series dynamics, aspects not captured by NN3.

Figure 6 shows the cumulative returns of the models for each factor. DMT stands out by achieving higher cumulative returns compared to the single-factor Buy across all factors. Notably, some of DMT's superior risk-adjusted returns can be traced to its ability to avoid the drawdowns that certain factors have experienced. For example, during the GFC and Covid crisis, DMT registered considerably smaller drawdowns for MKT and CMA compared to other models. These results underscore the effectiveness of DMT in improving factor timing strategies.

4.4.1 Transaction Costs

Table 5 presents the profitability of multi-factor timing strategies after accounting for transaction costs incurred when deviating from the Buy. Motivated by Moreira and Muir (2017), we consider transaction costs of one, five, ten, and fourteen basis points, which are subtracted from returns in months when factors are sold, i.e., where Equation (3) is zero. Even with transaction costs of fourteen basis points in Panel (d), DMT still achieves an alpha of 1.33% (t-stat of 3.42) and Sharpe ratio of 1.17, surpassing the multi-factor Buy Sharpe ratio of 0.98. MT also performs well, earning an alpha of 0.68% (t-stat of 2.21) and Sharpe ratio of 1.06. In contrast, none of the off-the-shelf models earn statistically significant alphas, and the highest Sharpe ratio earned by an off-the-shelf model (RF) is 1.01.

Figure 6: Cumulative Returns of Single-factor Timing



(a) MKT **(b) SMB**

(c) HML **(d) RMW**

(e) CMA **(f) MOM**

This figure plots the cumulative log returns of single-factor timing strategies by model for each factor over the out-of-sample period from January 1990 to December 2021. Each model's returns are adjusted to have the same volatility as the buy-and-hold benchmark (Buy) of the respective factor. NBER recessions are shown in grey.

Table 6 presents the profitability of DMT for each factor. Even with transaction costs of fourteen basis points in Panel (d), DMT achieves a positive alpha for all factors. Notably, DMT's alpha for MKT is 4.5% (t-stat of 3.94). DMT also earns a Sharpe ratio that beats the single-factor Buy, displayed in Table 4, for every factor except SMB.

Table 5: Factor-timing Profitability with Transaction Costs

	Buy	LR	EN	RF	XRF	GBT	SVM	NN	MT	DMT
SR	0.98	0.83	0.82	1.07	1.00	0.87	0.87	0.77	1.13	1.25
α		0.52	0.17	0.78	0.42	0.49	0.09	-0.51	0.96	1.66
$t(\alpha)$		1.10	0.39	2.33	1.42	1.12	0.26	-1.84	3.09	4.23

(a) One Basis Point

	Buy	LR	EN	RF	XRF	GBT	SVM	NN	MT	DMT
SR	0.98	0.77	0.78	1.05	0.99	0.84	0.85	0.75	1.11	1.22
α		0.32	0.04	0.71	0.38	0.35	-0.00	-0.59	0.87	1.56
$t(\alpha)$		0.69	0.08	2.12	1.28	0.80	-0.00	-2.10	2.82	3.98

(b) Five Basis Points

	Buy	LR	EN	RF	XRF	GBT	SVM	NN	MT	DMT
SR	0.98	0.70	0.74	1.03	0.98	0.79	0.82	0.72	1.08	1.19
α		0.08	-0.13	0.62	0.33	0.18	-0.11	-0.70	0.77	1.43
$t(\alpha)$		0.16	-0.30	1.85	1.12	0.41	-0.32	-2.42	2.48	3.67

(c) Ten Basis Points

	Buy	LR	EN	RF	XRF	GBT	SVM	NN	MT	DMT
SR	0.98	0.65	0.70	1.01	0.97	0.75	0.79	0.69	1.06	1.17
α		-0.12	-0.26	0.55	0.29	0.04	-0.21	-0.78	0.68	1.33
$t(\alpha)$		-0.25	-0.60	1.64	0.99	0.10	-0.58	-2.67	2.21	3.42

(d) Fourteen Basis Points

This table presents the Sharpe ratio (SR), alpha (α), and alpha t-statistic (t(α)) of the factor timing strategy for each model after accounting for transaction costs. We consider one, five, ten, and fourteen basis points. Transaction costs are subtracted from strategy returns when factors are sold.

4.5 Which Predictors Matter?

We next study the importance of each predictor for factor timing. The ideal approach would be to re-estimate the model while sequentially excluding each predictor. However, this method is not feasible due to the computational demands posed by the large number of predictors (a total of 272) in our estimation. To circumvent this challenge, we employ Shapley values, a methodology rooted in cooperative game theory, to assess the impact of individual predictors on the model's prediction. Shapley values provide an approximation

Table 6: DMT Single-factor Timing Profitability with Transaction Costs

	MKT	SMB	HML	RMW	CMA	MOM
SR	0.89	0.21	0.21	0.53	0.46	0.37
α	4.76	0.75	1.07	0.64	1.14	0.76
$t(\alpha)$	4.15	0.80	1.25	1.26	2.05	0.87

(a) One Basis Point

	MKT	SMB	HML	RMW	CMA	MOM
SR	0.88	0.18	0.20	0.53	0.44	0.37
α	4.68	0.56	0.96	0.58	1.02	0.74
$t(\alpha)$	4.09	0.60	1.12	1.14	1.83	0.84

(b) Five Basis Points

	MKT	SMB	HML	RMW	CMA	MOM
SR	0.87	0.15	0.18	0.52	0.42	0.37
α	4.58	0.32	0.81	0.50	0.87	0.71
$t(\alpha)$	4.00	0.35	0.95	0.99	1.56	0.81

(c) Ten Basis Points

	MKT	SMB	HML	RMW	CMA	MOM
SR	0.87	0.12	0.17	0.51	0.40	0.37
α	4.50	0.13	0.70	0.44	0.74	0.68
$t(\alpha)$	3.94	0.14	0.81	0.86	1.34	0.78

(d) Fourteen Basis Points

This table presents the Sharpe ratio (SR), alpha (α), and alpha t-statistic (t(α)) of the single-factor timing strategies for DMT after accounting for transaction costs. We consider one, five, ten, and fourteen basis points. Transaction costs are subtracted from strategy returns when factors are sold.

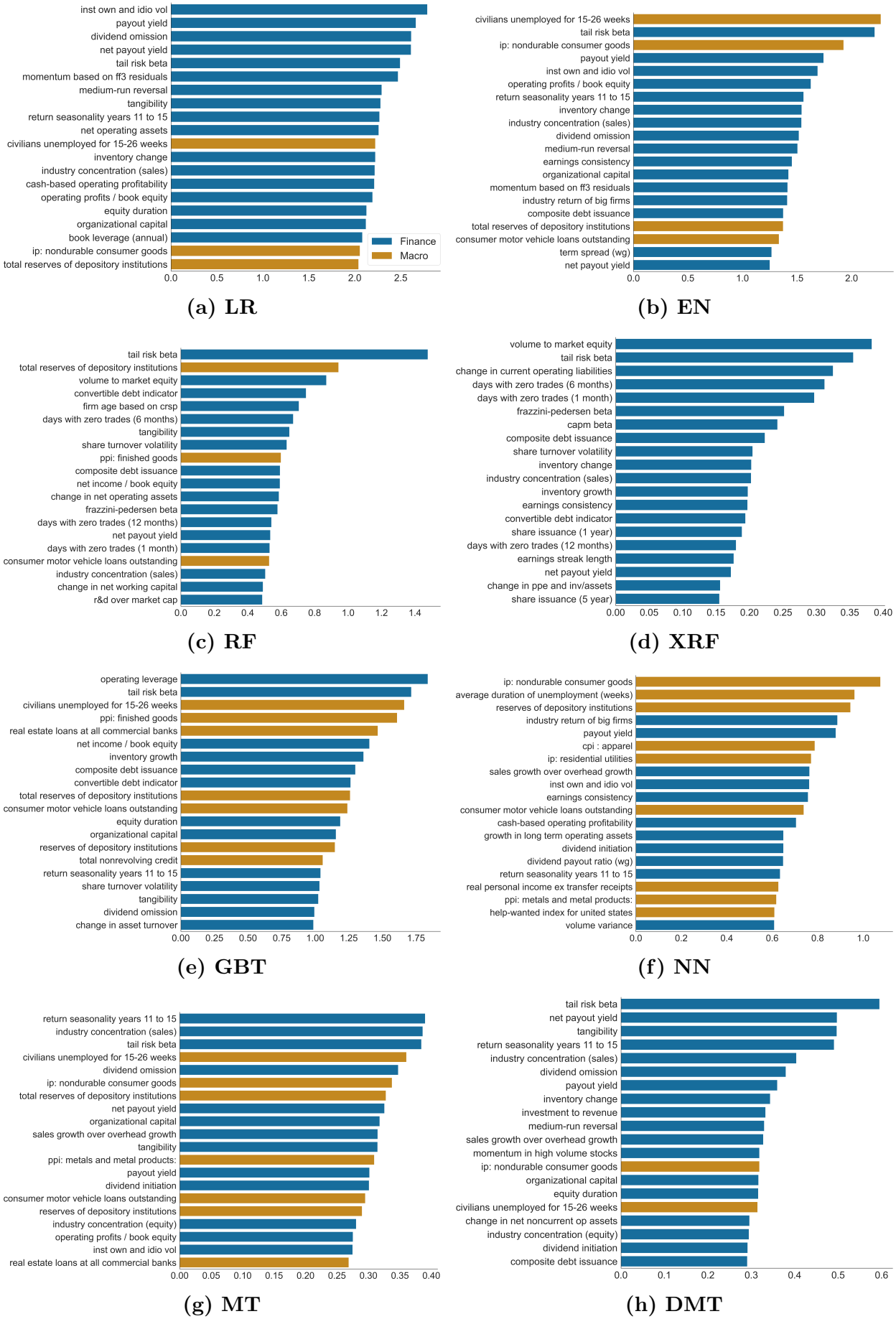
of how the model's probability predictions would have varied if certain predictors had been excluded in its estimation. For a detailed description of Shapley values, see Section IA2 in the Internet Appendix.

4.5.1 Multi-factor Variable Importance

We compute Shapley values for each model and factor for the out-of-sample period from January 1990 to December 2021. For each model and factor, we calculate the absolute Shapley value for each month and average these into a single importance measure for each predictor. Subsequently, we obtain a multi-factor importance measure for each model by averaging across the factors. Figure 7 displays the twenty most influential variables by model. Figures 8 and 9 present the rankings of financial and macroeconomic variables, respectively. We rank the importance of each variable for each model, then sum their ranks. The color gradient within each column shows the model-specific ranking of predictors from the most to least important (darkest to lightest).

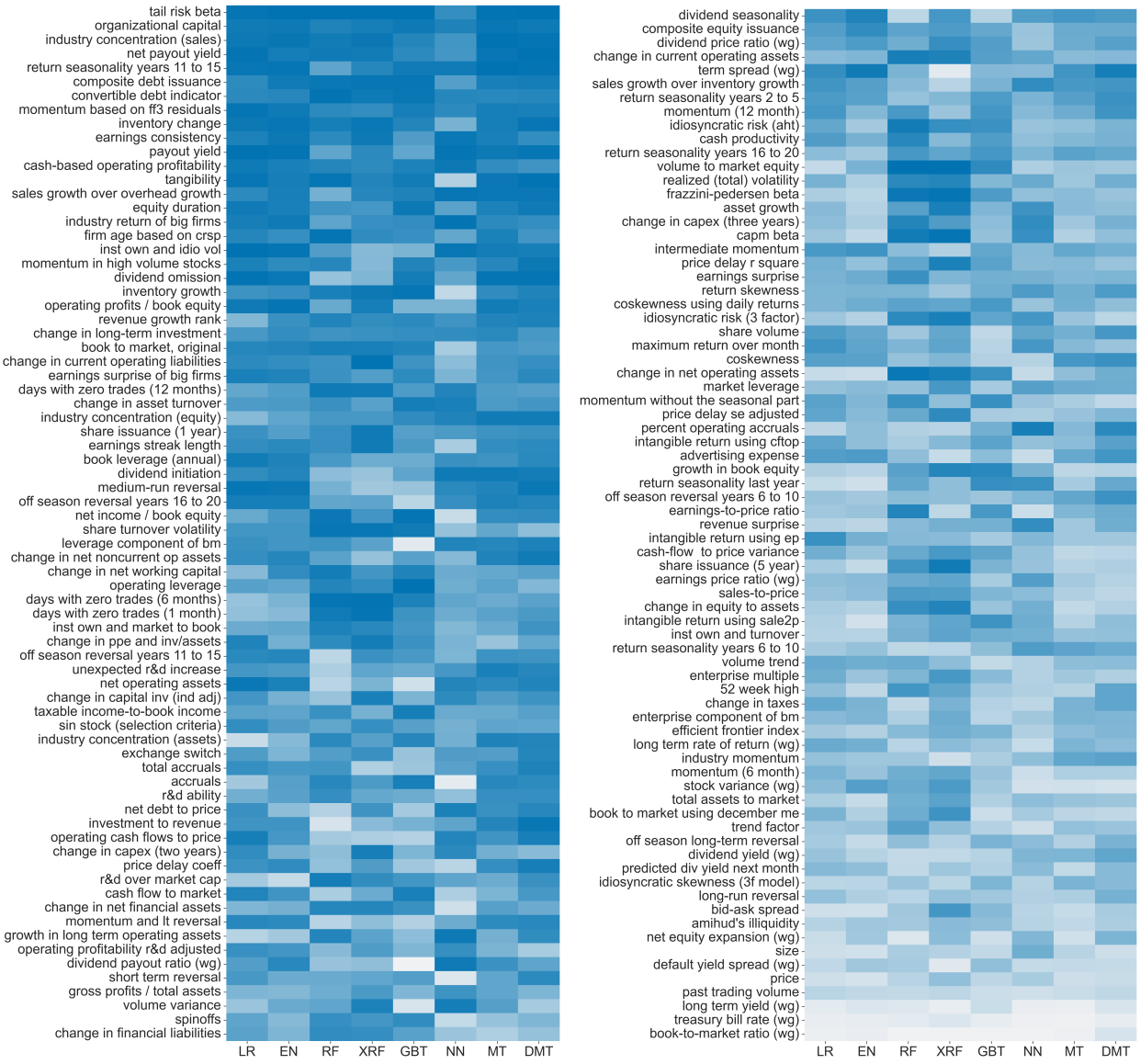
Figure 7 reveals a notable trend across all analyzed models: the number of influential financial variables exceeds that of macroeconomic variables. Delving deeper, Figure 8 reveals a consensus among models regarding the key financial variables. First, tail risk beta, belonging to the risk category, emerges as the most influential variable. This aligns with the findings in Kelly and Jiang (2014) and Liu (2020), which demonstrates the efficacy of tail risk, measured via a power law distribution, in predicting market and individual stock returns. However, it's interesting to note that other variables in the risk category, such as idiosyncratic risk, coskewness, and capm beta, are less influential, ranking lower among financial variables. Second, the leverage category contributes four of the top ten most influential variables, including organizational capital, industry concentration, composite debt issuance, and the convertible debt indicator. Third, price trends significantly impact model predictions, including return seasonality years 11 to 15, momentum based on ff3 residuals, industry return of big firms, and momentum in high volume stocks. Fourth, variables in the

Figure 7: Variable Importance by Model



This figure presents the twenty most influential variables by model. For each model and factor, we calculate the absolute Shapley value for each month and average these into a single importance measure for each predictor. Subsequently, we obtain a multi-factor importance measure for each predictor in each model by averaging across the factors.

Figure 8: Variable Importance Rankings of Financial Variables



(a) Top 74 Financial Variables **(b) Bottom 75 Financial Variables**

This figure displays the rankings of financial variables. We rank the multi-factor importance of each financial variable for each model, then sum their ranks. The color gradient within each column shows the model-specific ranking of predictors from the most to least important (darkest to lightest).

value category, such as net payout yield, payout yield, equity duration, and book to market, as well as those in the profitability category, including earnings consistency, dividend omission, operating profits/book equity, and earnings surprise of big firms, also have a notable influence.

Figure 9: Variable Importance Rankings of Macroeconomic Variables

(a) Top 57 Macroeconomic Variables **(b) Bottom 66 Macroeconomic Variables**

This figure displays the rankings of macroeconomic variables. We rank the multi-factor importance of each macroeconomic variable for each model, then sum their ranks. The color gradient within each column shows the model-specific ranking of predictors from the most to least important (darkest to lightest).

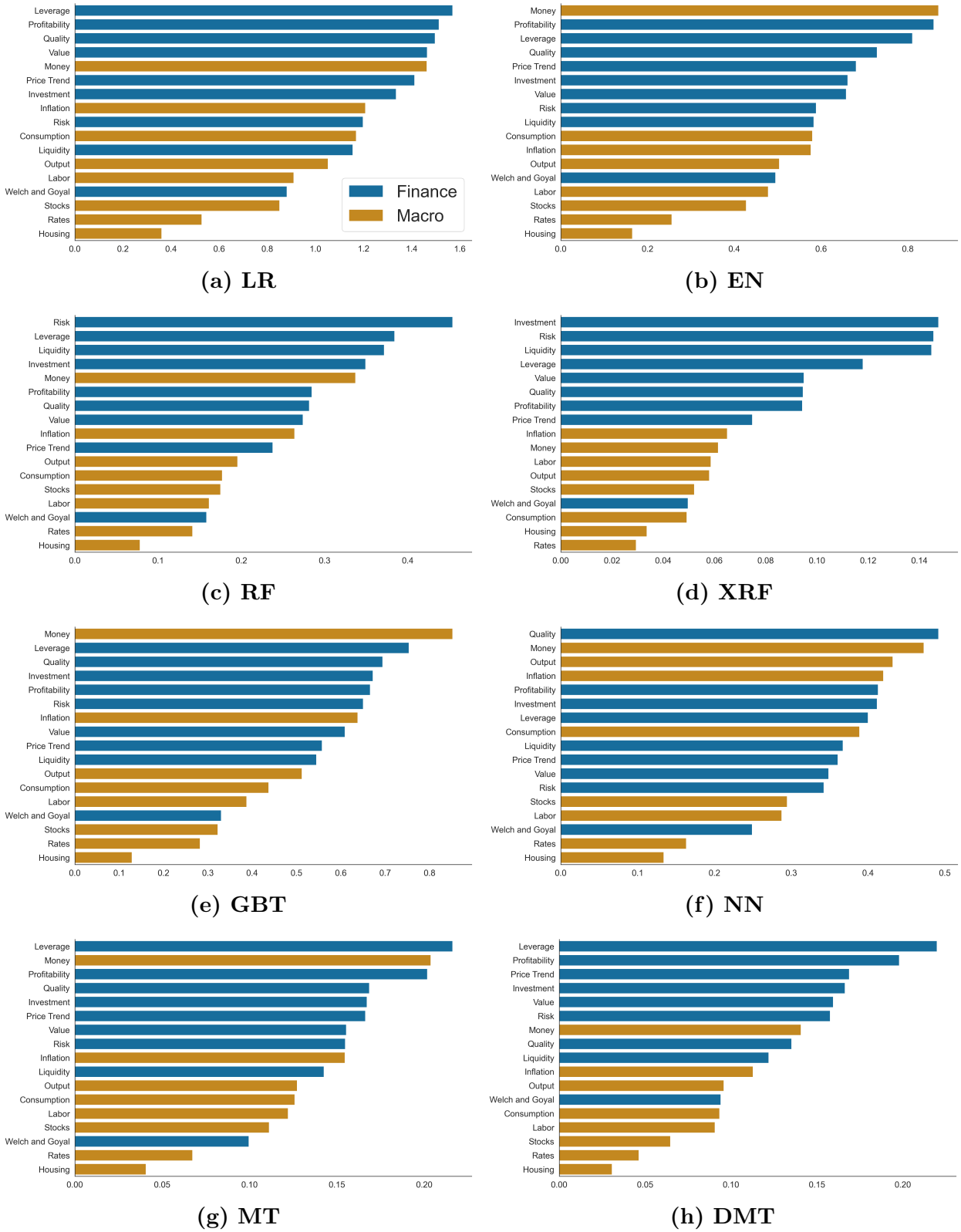
We find that numerous asset pricing anomalies significantly contribute to predictive accuracy in factor timing. This result aligns with the conclusions presented in Dong et al. (2022) but stands in contrast to the findings of Cakici et al. (2023) and Engelberg et al. (2023). Interestingly, variables identified by Gu et al. (2020) as important predictors of individual stock returns have a diminished influence for factor timing. These variables include short-term reversal, 6- and 12-month momentum, size, Amihud's illiquidity measure, and the bid-ask spread. This is likely because the predictions of machine learning models for

individual stocks are often driven by short-lived characteristics that are effective for small and illiquid stocks (Avramov et al. (2023), Jensen et al. (2022)). In contrast, our findings indicate that tail risk and accounting variables, particularly from the leverage, value, and profitability categories, are highly influential for factors, which are large and liquid. This underscores the substantial differences in the functional forms between individual stocks and factors.

Figure 9 presents the rankings of macroeconomic variables. Money is the most significant macroeconomic category, featuring key variables such as the total reserves of depository institutions, real estate loans at all commercial banks, consumer motor vehicle loans outstanding, and m1 money stock. Output ranks as the second most influential macroeconomic category, with a particular emphasis on IP pertaining to nondurable consumer goods and residential utilities. Labor emerges as the third most influential macroeconomic category, with key variables including the number of civilians unemployed for 15-26 weeks and the average duration of unemployment. Inflation ranks as the fourth most influential macroeconomic category, comprising indices such as the Consumer Price Index (CPI) for apparel and transportation; the Producer Price Index (PPI) for metals and finished goods; as well as Personal Consumption Expenditures (PCE) for durable goods. Lastly, while the canada/u.s. and u.s./u.k. foreign exchange rates have some influence, interest rate variables rank as some of the least influential macroeconomic variables.

Next, we derive an importance measure for each category by averaging the multi-factor importance measures within each respective category. Figure 10 shows that the category importance by model aligns with the the predictor importance results in Figure 7. There is a consensus among models that financial categories generally have a greater influence than macroeconomic categories. Among financial categories, risk, leverage, profitability, and price trends stand out in terms of influence. Additionally, the quality and investment categories demonstrate significant influence, whereas the value category shows limited importance, despite the presence of several important variables in the value category, as illustrated in

Figure 10: Category Importance by Model



This figure presents the most influential categories by model. For each model and factor, we calculate the absolute Shapley value for each month and average these into a single importance measure for each predictor. Subsequently, we obtain a multi-factor importance measure for each predictor in each model by averaging across the factors. Finally, we derive an importance measure for each category by averaging the multi-factor importance measures within each respective category.

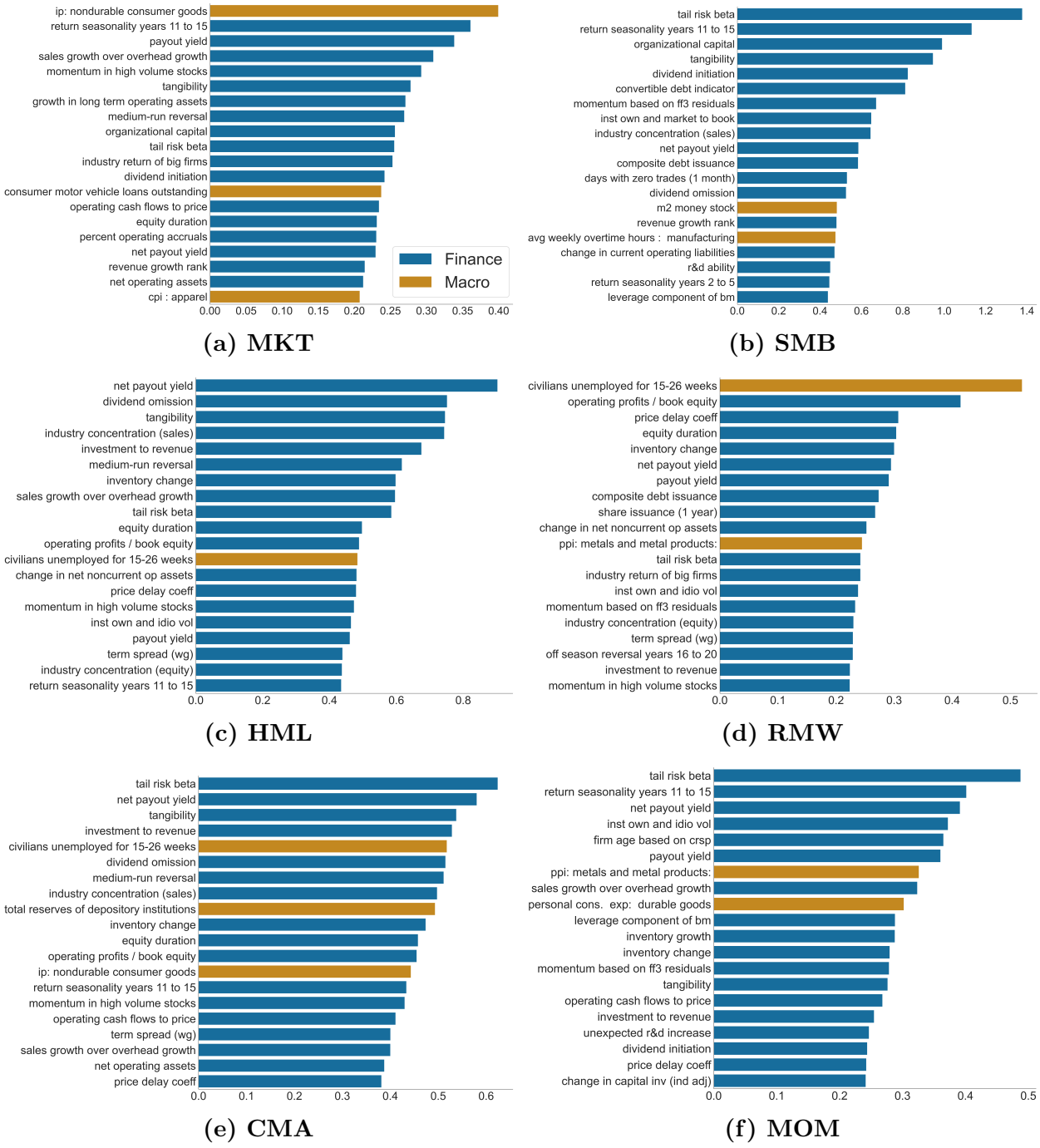
Figure 8. Within the macroeconomic categories, money, inflation, and output are particularly influential, while housing and rates rank as the least influential across all models. Notably, the variables from Welch and Goyal (2008) and the stock category rank lower in influence compared to those based on long-short portfolios, suggesting that long-short portfolios offer stronger predictive signals than aggregate variables. This aligns with the findings of strong predictive power for long-short portfolios as reported in Dong et al. (2022) and the weak predictive capability of aggregate variables documented in Engelberg et al. (2023) and Welch and Goyal (2008).

4.5.2 Single-factor Variable Importance

We now study the influential variables for each factor. Figure 11 presents the twenty most influential variables for each factor according to the DMT model. Panel (a) shows that for MKT, macroeconomic variables related to output (ip: nondurable consumer goods), money (consumer motor vehicle loans outstanding), and inflation (cpi: apparel) have significant influence. Our results, indicating that macroeconomic variables in the money and inflation categories are predictive of the market, are consistent with the findings in Flannery and Protopapadakis (2002). However, in contrast to their findings, we discover that IP related to nondurable consumer goods emerges as the most significant market predictor among all variables.

Panel (a) further reveals that price trends are a highly influential financial category for the market, encompassing variables such as return seasonality years 11 to 15, momentum in high volume stocks, medium-run reversal, and industry return of big firms. This is consistent with the recognized significance of momentum in time series market predictability, as detailed in Moskowitz et al. (2012). We also find that variables across all accounting categories hold significant influence, including value (payout yield, operating cash flows to price, equity duration), profitability (sales growth over overhead growth, dividend initiation, percent operating accruals), leverage (tangibility, organizational capital, net operating as-

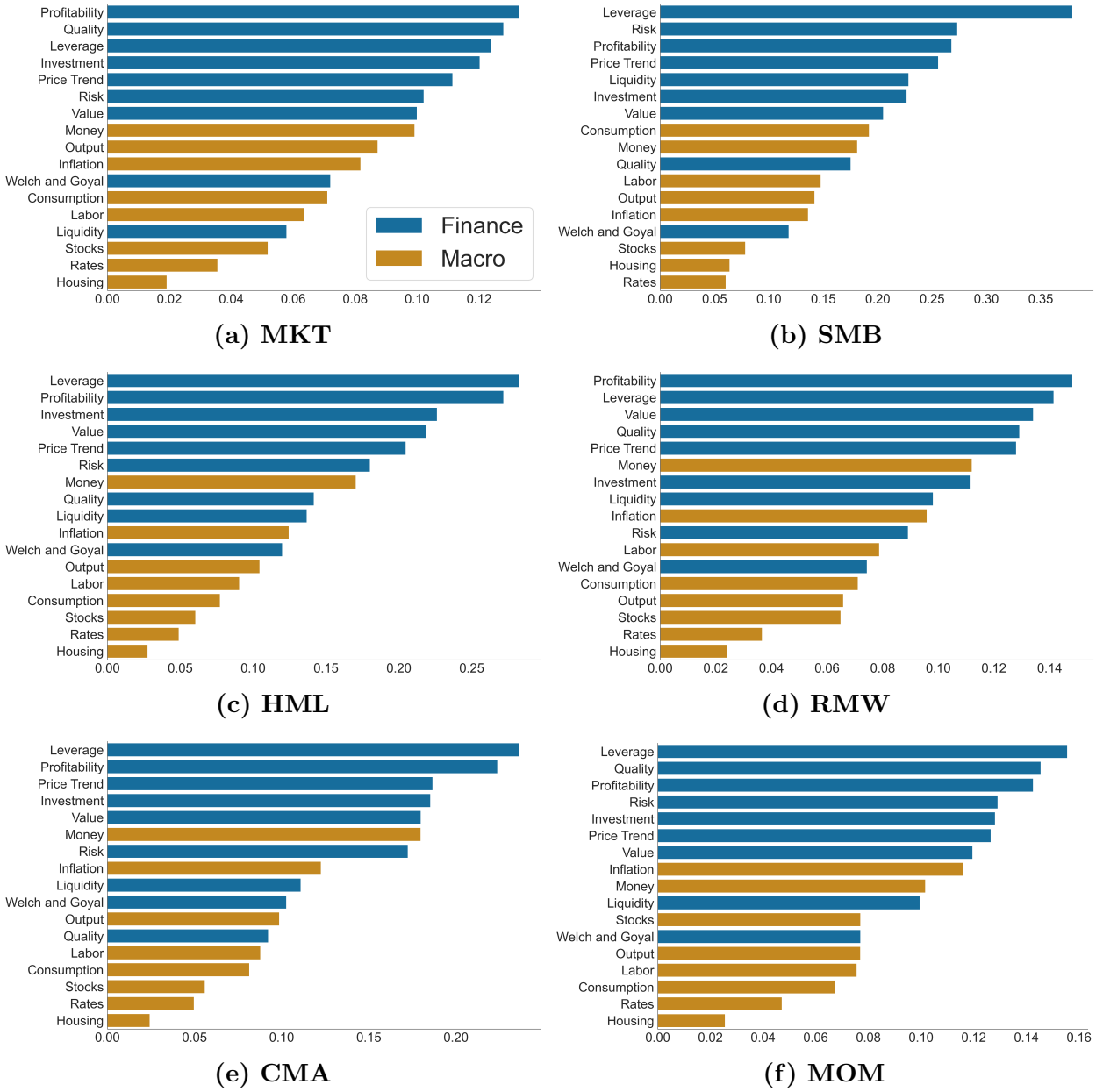
Figure 11: Variable Importance by Factor for the DMT Model



This figure presents the twenty most influential variables by factor for the DMT model. For each factor, we calculate the absolute Shapley value for each month and average these into a single importance measure for each predictor.

sets), and investment (growth in long term operating assets, revenue growth rank). The influential nature of accounting variables for market predictability aligns with the findings

Figure 12: Category Importance by Factor for the DMT Model



This figure presents the importance of variable categories by factor for the DMT model. For each factor, we calculate the absolute Shapley value for each month and average these into a single importance measure for each predictor. These importance measures are then averaged by category.

of Chen and Zhang (2007). Lastly, consistent with Kelly and Jiang (2014), we find that tail risk is a strong predictor of the market.

Panel (b) reports that financial variables have the largest influence on SMB, in particular tail risk beta. Variables in the price trend category are also highly influential, encompassing

return seasonality years 11 to 15, return seasonality years 2 to 5, and momentum based on ff3 residuals. Leverage also plays a pivotal role, with eight of the top twenty influential variables falling in this category, including organizational capital, tangibility, convertible debt indicator, industry concentration, composite debt issuance, r&d ability, and leverage component of bm. Several liquidity variables also demonstrate significant influence, including short-sale constraints (inst own and market to book) and days with zero trades. The only macroeconomic variables with a notable influence on the size factor fall within the money (m2 money stock) and labor (avg weekly overtime hours: manufacturing) categories. The significant influence of financial conditions on the size factor is intuitive, considering small firms' heightened sensitivity to tail risk, leverage, illiquidity, and fluctuations in credit and labor market conditions.

Panel (c) emphasizes the crucial role of accounting variables for HML, particularly within the categories of value (net payout yield, equity duration), profitability (dividend omission, sales growth over overhead growth), leverage (tangibility, industry concentration), and investment (investment to revenue, inventory change). Additionally, price trends (medium-run reversal, price delay coeff, momentum in high volume stocks) and tail risk are highly influential. The only influential macroeconomic variable for the value factor is civilians unemployed for 15-26 weeks in the labor category.

Interestingly, Panels (d) and (e), which focus on RMW and CMA, show a pattern similar to HML, suggesting a common set of variables influencing the value, profitability, and investment factors. These findings are intuitive considering that these factors are based on accounting data, making accounting variables highly predictive. Notably, industry concentration matters for all three factors, aligning with the findings of Hou and Robinson (2006). This might also shed light on the recent underperformance of these factors, especially considering the increase in industry concentration post-GFC (Grullon et al., 2019). A noteworthy discovery is the substantial influence of tail risk, price trends, and unemployment on the value, profitability, and investment factors, which warrants further investigation to uncover

the underlying economic mechanisms driving this phenomenon.

Lastly, Panel (f) demonstrates that tail risk is the most influential predictor for MOM, aligning with the tendency of momentum strategies to be vulnerable to significant crashes (Daniel and Moskowitz, 2016). Variables in the price trend category, notably return seasonality years 11 to 15, momentum based on ff3 residuals, and price delay coeff, are significantly influential for MOM. Surprisingly, accounting variables also hold significant influence over MOM, especially within the categories of value (net payout yield, payout yield, operating cash flows to price), leverage (firm age based on crsp, leverage component of bm, tangibility, investment to revenue, unexpected r&d increase), and profitability (sales growth over inventory growth, dividend initiation, change in capital inv). Furthermore, macroeconomic variables related to inflation, encompassing both producers (ppi: metals and metal products) and consumers (personal cons. exp: durable goods), exert substantial influence on MOM. These influential predictors complement the conventional view of momentum as mainly driven by sentiment (Antoniou et al., 2013), by also highlighting the significant role of fundamental macroeconomic and financial indicators in forecasting momentum.

We next analyze the category importance for each factor using the DMT model, as shown in Figure 12. First, leverage and profitability rank among the top three for every factor, indicating their critical role in factor timing. This suggests that the common structure across factors may be driven by key variables in these influential categories. Secondly, although certain macroeconomic predictors such as IP and unemployment are important for factor timing, financial categories tend to have greater influence than macroeconomic categories across all factors. Among the macroeconomic categories, money, output, and inflation are the most influential, whereas stocks, rates, and housing rank as the least influential categories.

To summarize, we find substantial variation in the rankings of the most influential variables across factors. Despite this heterogeneity, a small subset of variables consistently emerge as influential across factors. These include financial variables related to tail risk, price trends, and accounting categories (particularly in the leverage and profitability cate-

gories). They also include macroeconomic measures of money, output, labor, and inflation. Additionally, we uncover several new influential predictors for factor timing, which merits further exploration into the economic mechanisms of these predictive relationships.

4.5.3 Nonlinear Interactions

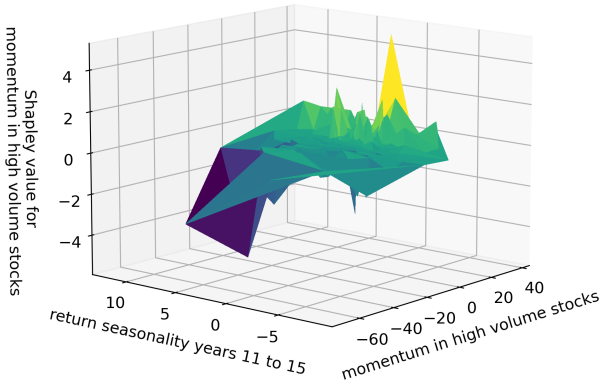
Finally, we study the contribution of nonlinear interactions to the superior performance of DMT. Although numerous potential interactions exist, we demonstrate a few to elucidate DMT's inner workings. Figure 13 displays three-dimensional plots, with predictor values on the horizontal axes and Shapley values on the vertical axis.

Panel (a) shows the effect of momentum in high volume stocks on DMT's MKT probability prediction and its interaction with return seasonality years 11 to 15. Notably, for moderate levels of momentum in high volume, changes in return seasonality years 11 to 15 have minimal impact on the probability predictions. However, when momentum in high volume stocks is extremely positive (negative), this can result in very high (low) probability predictions, conditional on the level of return seasonality years 11 to 15. This interaction introduces nonlinear characteristics to the surface, resulting in a cubic appearance. For example, when momentum in high volume stocks is close to zero, changes in return seasonality years 11 to 15 have almost no effect on probability predictions, as characterized by a flat surface. Conversely, when momentum in high-volume stocks is extremely positive (negative), values of return seasonality years 11 to 15 close to zero are associated with significantly high (low) probability predictions.

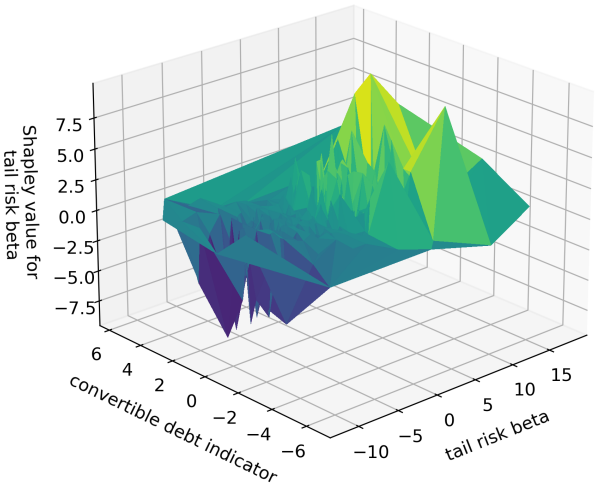
Panels (b) reveals that for SMB, there is a similar cubic relationship between tail risk beta and the convertible debt indicator.[5] Panels (c) through (f) reveal similar effects for the remaining factors, where pairs of influential variables generate complex nonlinear interactions. Although our analysis is limited to three dimensions for visualization purposes, it is plausible to infer that, in reality, there are highly complex nonlinear relationships in

[5]Our convertible debt indicator predictor, being a long-short portfolio sorted based on the respective binary anomaly, is represented as a continuous variable.

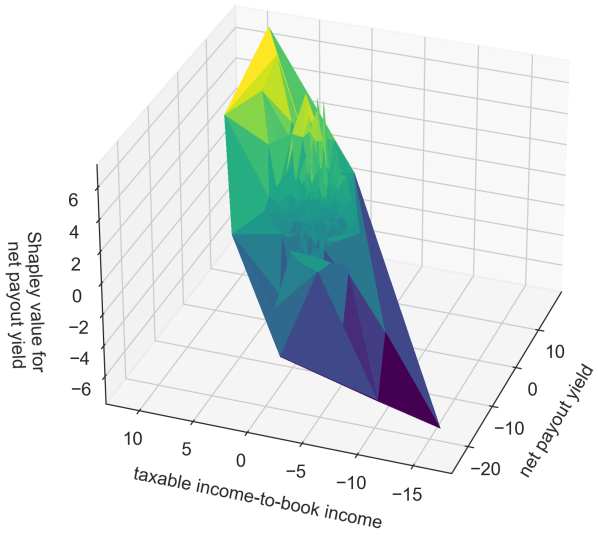
Figure 13: Interaction Plots by Factor



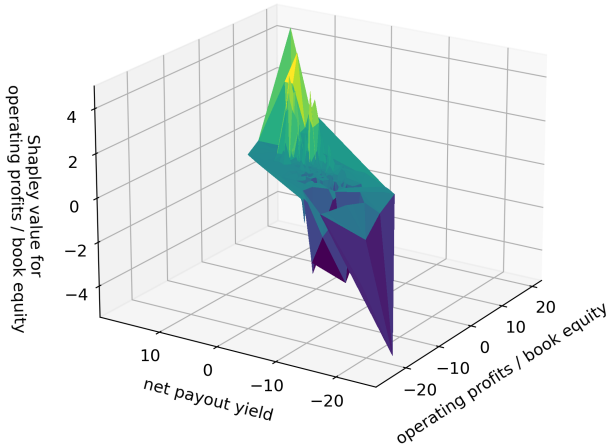
(a) MKT



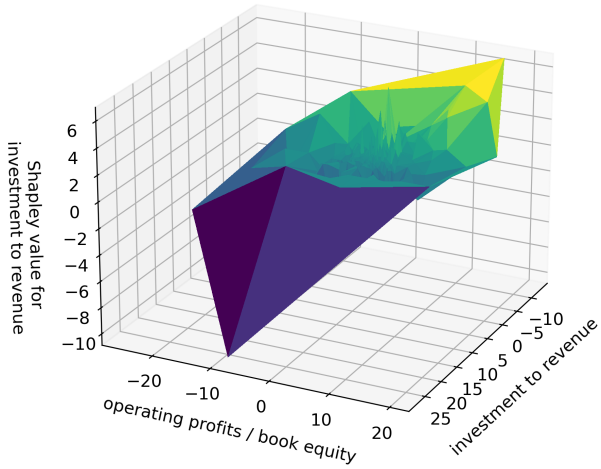
(b) SMB



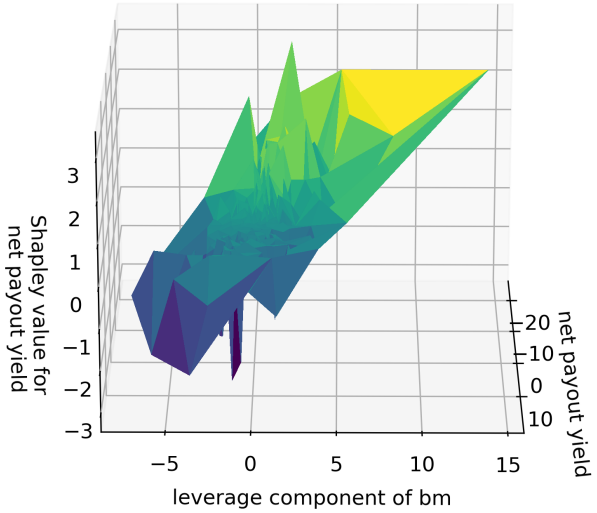
(c) HML



(d) RMW



(e) CMA



(f) MOM

Plot of three-dimensional surfaces. Each plot shows a variable plotted along the right horizon axis, its Shapley values for the DMT model along the vertical, and a variable it interacts with on the left horizontal axis.

even higher-dimensional spaces driving factor risk premia. More importantly, we illustrate the benefits of DMT in capturing these sophisticated nonlinear interactions, which result in superior out-of-sample performance. This reinforces the effectiveness of DMT in leveraging these multidimensional relationships to optimize predictions, providing a holistic understanding of the intricate web of variables that influence factor risk premia.

5 Conclusion

Understanding factor risk premia is a central topic in financial economics and the burgeoning factor investing industry. An extensive literature studies factor predictability by employing static and linear single-factor models with a limited set of predictors. These approaches have yielded inconsistent outcomes, leaving the feasibility of factor timing as a topic of considerable debate. Furthermore, these traditional models are inadequate in handling a zoo of predictors that are influential for factor timing, leading to inconclusive conclusions about the key drivers of factor predictability.

In this paper, we introduce deep neural networks that incorporate economically motivated restrictions, tailored to address the main challenges of factor timing. We develop a dynamic multi-task deep learning model to forecast six well-known factors, using 123 macroeconomic and 149 financial predictors, in the 57 year period from January 1965 to December 2021. Empirically, we demonstrate several results that enhance our knowledge of factor timing. First, we show that incorporating economic structure and time series dynamics significantly improves the predictive accuracy of factor timing models. Second, our results reveal that deep learning models with economic structure produce significant economic gains in a multi-factor portfolio, which is further enhanced by incorporating time series dynamics. Third, we find that integrating multi-task learning with time series dynamics can yield consistent economic gains across all factors relative to the buy-and-hold benchmark. Fourth, we document that the most important variables for factor timing, include tail risk and variables

belonging to the price trends, leverage, and profitability categories. Fifth, we demonstrate that nonlinear interactions among these influential variables are important for effective factor timing. Overall, the improved factor timing yielded from our dynamic multi-factor deep learning approach paves the way for a more reliable investigation of the economic mechanisms driving factor risk premia, and underscores the value of incorporating economically motivated restrictions into deep learning models for factor investing.

References

Antoniou, C., Doukas, J. A., and Subrahmanyam, A. (2013). Cognitive dissonance, sentiment, and momentum. *Journal of Financial and Quantitative Analysis*, 48(1):245–275.

Arnott, R., Harvey, C. R., Kalesnik, V., and Linnainmaa, J. (2019). Alice's adventures in factorland: Three blunders that plague factor investing. *Journal of Portfolio Management*, 45(4):18–36.

Asness, C., Chandra, S., Ilmanen, A., and Israel, R. (2017). Contrarian factor timing is deceptively difficult. *The Journal of Portfolio Management*, 43(5):72–87.

Asness, C. S. (2016). Invited editorial comment: The siren song of factor timing aka "smart beta timing" aka "style timing". *The Journal of Portfolio Management*, 42(5):1–6.

Avramov, D., Cheng, S., and Metzker, L. (2023). Machine learning vs. economic restrictions: Evidence from stock return predictability. *Management Science*, 69(5):2587–2619.

Baba Yara, F., Boons, M., and Tamoni, A. (2021). Value return predictability across asset classes and commonalities in risk premia. *Review of Finance*, 25(2):449–484.

Bender, J., Sun, X., Thomas, R., and Zdorovtsov, V. (2018). The promises and pitfalls of factor timing. *The Journal of Portfolio Management*, 44(4):79–92.

Bianchi, D., Büchner, M., and Tamoni, A. (2021). Bond risk premiums with machine learning. *The Review of Financial Studies*, 34(2):1046–1089.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth.

Bryzgalova, S., Pelger, M., and Zhu, J. (2023). Forest through the trees: Building cross-sections of stock returns. *Journal of Finance, forthcoming*.

Cakici, N., Fieberg, C., Metko, D., and Zaremba, A. (2023). Do anomalies really predict market returns? new data and new evidence. *Review of Finance*, page rfad025.

Campbell, J. Y. and Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4):1509–1531.

Caruana, R. (1997). Multitask learning. *Machine Learning*, 28:41–75.

Chen, A. Y. and Zimmermann, T. (2022). Open source cross-sectional asset pricing. *Critical Finance Review*, 11(2):207–264.

Chen, L., Pelger, M., and Zhu, J. (2023). Deep learning in asset pricing. *Management Science*.

Chen, P. and Zhang, G. (2007). How do accounting variables explain stock price movements? theory and evidence. *Journal of Accounting and Economics*, 43(2-3):219–244.

Cohen, R. B., Polk, C., and Vuolteenaho, T. (2003). The value spread. *The Journal of Finance*, 58(2):609–641.

Cooper, M. J., Gutierrez Jr, R. C., and Hameed, A. (2004). Market states and momentum. *The Journal of Finance*, 59(3):1345–1365.

Daniel, K. and Moskowitz, T. J. (2016). Momentum crashes. *Journal of Financial Economics*, 122(2):221–247.

Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in Neural Information Processing Systems*, 27.

Dichtl, H., Drobetz, W., Lohre, H., Rother, C., and Vosskamp, P. (2019). Optimal timing and tilting of equity factors. *Financial Analysts Journal*, 75(4):84–102.

Didisheim, A., Ke, S. B., Kelly, B. T., and Malamud, S. (2023). Complexity in factor pricing models.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Dong, X., Li, Y., Rapach, D. E., and Zhou, G. (2022). Anomalies and the expected market return. *The Journal of Finance*, 77(1):639–681.

Engelberg, J., McLean, R. D., Pontiff, J., and Ringgenberg, M. C. (2023). Do cross-sectional predictors contain systematic information? *Journal of Financial and Quantitative Analysis*, 58(3):1172–1201.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Fama, E. F. and French, K. R. (1996). Multifactor explanations of asset pricing anomalies. *The Journal of Finance*, 51(1):55–84.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.

Feng, G., He, J., Polson, N. G., and Xu, J. (2018). Deep learning in characteristics-sorted factor models. *arXiv preprint arXiv:1805.01104*.

Flannery, M. J. and Protopapadakis, A. A. (2002). Macroeconomic factors do influence aggregate stock returns. *The Review of Financial Studies*, 15(3):751–782.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63:3–42.

Greenwood, R. and Hanson, S. G. (2012). Share issuance and factor timing. *The Journal of Finance*, 67(2):761–798.

Grullon, G., Larkin, Y., and Michaely, R. (2019). Are us industries becoming more concentrated? *Review of Finance*, 23(4):697–743.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Gu, S., Kelly, B., and Xiu, D. (2021). Autoencoder asset pricing models. *Journal of Econometrics*, 222(1):429–450.

Guijarro-Ordonez, J., Pelger, M., and Zanotti, G. (2021). Deep learning statistical arbitrage. *arXiv preprint arXiv:2106.04028*.

Gupta, T. and Kelly, B. (2019). Factor momentum everywhere. *The Journal of Portfolio Management*, 45(3):13–36.

Haddad, V., Kozak, S., and Santosh, S. (2020). Factor timing. *The Review of Financial Studies*, 33(5):1980–2018.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hodges, P., Hogan, K., Peterson, J. R., and Ang, A. (2017). Factor timing with cross-sectional and time-series predictors. *The Journal of Portfolio Management*, 44(1):30–43.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Hou, K. and Robinson, D. T. (2006). Industry concentration and average stock returns. *The Journal of Finance*, 61(4):1927–1956.

Ilmanen, A., Israel, R., Lee, R., Moskowitz, T. J., and Thapar, A. (2021). How do factor premia vary over time? a century of evidence. *Journal Of Investment Management*, 19(4):15–57.

Jensen, T. I., Kelly, B., and Pedersen, L. H. (2023). Is there a replication crisis in finance? *The Journal of Finance*, 78(5):2465–2518.

Jensen, T. I., Kelly, B. T., Malamud, S., and Pedersen, L. H. (2022). Machine learning and the implementable efficient frontier. *Available at SSRN 4187217*.

Kagkadis, A., Nolte, I., Nolte, S., and Vasilas, N. (2023). Factor timing with portfolio characteristics. *The Review of Asset Pricing Studies*, page raad010.

Kelly, B. and Jiang, H. (2014). Tail risk and asset prices. *The Review of Financial Studies*, 27(10):2841–2871.

Kelly, B. T., Malamud, S., and Zhou, K. (2023). The virtue of complexity in return prediction. *The Journal of Finance, forthcoming*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koijen, R. S. and Van Nieuwerburgh, S. (2011). Predictability of returns and cash flows. *Annu. Rev. Financ. Econ.*, 3(1):467–491.

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

Liu, F. (2020). Can the premium for idiosyncratic tail risk be explained by exposures to its common factor? *Available at SSRN 3711215*.

Liu, F. (2023). Quantile machine learning and the cross-section of stock returns. *Available at SSRN 4491887*.

McCracken, M. W. and Ng, S. (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589.

Moreira, A. and Muir, T. (2017). Volatility-managed portfolios. *The Journal of Finance*, 72(4):1611–1644.

Moskowitz, T. J., Ooi, Y. H., and Pedersen, L. H. (2012). Time series momentum. *Journal of Financial Economics*, 104(2):228–250.

Polk, C., Haghbin, M., and de Longis, A. (2020). Time-series variation in factor premia: The influence of the business cycle. *Journal Of Investment Management*, 18(1):69–89.

Proner, R. (2023). A multi-task deep learning model for inflation forecasting: Dynamic phillips curve neural network. *Available at SSRN 4454118*.

Rapach, D. and Zhou, G. (2013). Forecasting stock returns. 2:328–383.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Štrumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665.

Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508.

Multi-Factor Timing with Deep Learning

Internet Appendix

Table of Contents:

- Section IA1 provides details on the **machine learning models** used in this paper.
- Section IA2 provides details our estimation of variable importance using **Shapley values**.
- Section IA3 provides details on the **setup of the hyperparameter optimization**.
- Table IA1 presents the **set of hyperparameters** used in this paper.
- Table IA2 provides a detailed list of the **financial predictors**.
- Table IA3 provides a detail list of the **macroeconomic predictors**.

IA1 Machine Learning Models

This section provides detailed descriptions for the off-the-shelf models used in this paper. We omit i from the notation in this section to emphasize that these models are estimated separately for each factor, rather than jointly as in the case of MT and DMT.

IA1.1 Linear Models

IA1.1.1 Logistic Regression

Logistic regression is a type of generalized linear model used for modeling binary variables. Unlike linear regression, which models continuous variables directly, logistic regression models the log-odds of a binary response variable. Specifically, it employs the log "link" function to connect the probability of a positive excess return at time $t+1$, denoted as π_{t+1}, with the P linear predictors at time t. Mathematically, this relationship models the log odds as a linear function of the predictors, expressed as follows:

$$\log\left(\frac{\pi_{t+1}}{1-\pi_{t+1}}\right) = x_t'\beta + \beta_0, \tag{4}$$

where $\beta = (\beta_1, ..., \beta_P)$ is a P-dimensional parameter vector. This expression can be transformed to obtain the desired probability

$$\pi_{t+1} = \frac{e^{x_t'\beta+\beta_0}}{1+e^{x_t'\beta+\beta_0}}, \tag{5}$$

using the sigmoid function

$$\sigma(z) = \frac{e^z}{1+e^z}. \tag{6}$$

Parameters β and β_0 can be estimated by minimizing objective function

$$\mathcal{L}(\beta, \beta_0) = -\sum_{t=1}^{T}(y_{t+1}\log(\pi_{t+1}) + (1 - y_{t+1})\log(1 - \pi_{t+1})), \tag{7}$$

where $y_{t+1} \equiv I(r_{t+1} > 0)$ is an indicator function that takes a value of one if the excess return is positive and zero otherwise. To achieve this, we utilize the SAGA algorithm (Defazio et al., 2014), which serves as the primary estimation technique in scikit-learn's implementation of logistic regression.

IA1.1.2 Penalized Logistic Regression

Logistic regression can suffer from high variance in the presence of many predictors, since the model may overfit to noise rather than extracting signal. This problem is compounded by the low signal-to-noise ratio of factors. To mitigate overfitting, we append an elastic net (EN) penalty term to the original objective function:

$$\mathcal{L}(\beta, \beta_0, \cdot) = \mathcal{L}(\beta, \beta_0) + \lambda \sum_{j=1}^{P} \frac{1 - \rho}{2}\beta_j^2 + \rho|\beta_j|. \tag{8}$$

The EN penalty involves two non-negative hyperparameters, λ and ρ. When $\rho = 1$, the EN penalty corresponds to the ℓ_1 penalty, capable of setting coefficients to exactly 0, thus enforcing sparsity and acting as a variable selection method. When $\rho = 0$, the EN penalty corresponds to the ℓ_2 penalty, which shrinks all coefficients towards zero without enforcing exact zeros. For intermediate values of ρ, the EN benefits from both the shrinkage effect and sparsity.

IA1.2 Nonlinear Models

IA1.2.1 Random Forests and Extremely Randomized Trees

Classification Trees: A classification tree partitions the predictor space into M regions. For any observation landing in a specific region, we predict the majority class of that region. We select regions to minimize the Gini index, given by

$$H = \sum_{k=1}^{2} p_{mk}(1 - p_{mk}), \tag{9}$$

where p_{mk} denotes the proportion of observations in region m belonging to class k. The Gini index measures the total variance across all classes. By targeting node purity with this metric, we reduce the likelihood of classification errors within a region. Nodes dominated by observations of either mostly zero or mostly one exhibit a low Gini index.

Scikit-learn employs the Classification and Regression Tree (CART) algorithm to train decision trees (Breiman et al., 1984). The CART algorithm for classification operates as follows:

1. Let the data at node m be denoted by Q_m. For each split candidate $\theta = (j, s_m)$ comprising of a predictor j and threshold s_m, partition the data into two subsets

$$Q_m^{left}(\theta) = \{(x_t, y_{t+1}) : x_{t,j} \leq s_m\} \text{ and } Q_m^{right}(\theta) = \{(x_t, y_{t+1}) : x_{t,j} > s_m\}.$$

2. To assess the quality of the candidate split for node m, utilize the Gini index from Equation (9) as the impurity metric:

$$G(Q_m, \theta) = \frac{T_m^{left}}{T_m} H(Q_m^{left}(\theta)) + \frac{T_m^{right}}{T_m} H(Q_m^{right}(\theta)),$$

   where T_m^{left} and T_m^{right} denote the number of observations in the left and right node, respectively.
3. Select the parameters θ^* that minimize $G(Q_m, \theta)$.
4. Repeat for subsets $Q_m^{left}(\theta^*)$ and $Q_m^{left}(\theta^*)$ until a desired tree depth is reached, until the number of samples in each leaf (end node) reaches some set minimum, or until there is only one sample remaining in each leaf.

Individual decision trees tend to overfit, leading to high variance. In our analysis, we study various ensemble tree models that aggregate forecasts from multiple trees into a single

prediction. These methods include random forests and extremely randomized trees, which construct trees independently, and gradient boosting, which builds trees sequentially.

Random forest is a collection of decision trees (Breiman, 2001). Each decision tree is fit on a bootstrap sample of the training data. At each splitting step, only a random subset of predictors is considered. The probability prediction from a random forest is the average of the predicted probabilities from the individual classification trees within the forest.

Extremely Randomized Trees, similar to random forest, add an extra layer of randomization by not only varying the number of predictors considered at each split but also randomizing the splitting thresholds (Geurts et al., 2006). In contrast to random forest, each decision tree is typically fitted on the original sample without employing bootstrapping.

IA1.2.2 Gradient Boosted Trees

Algorithm 1 Gradient Boosted Trees for Binary Classification

1: Initialize $f_0(x) = log\left(\frac{\sum_{t=1}^{T} y_{t+1}}{\sum_{t=1}^{T}(1-y_{t+1})}\right)$

2: For $b = 1, 2, ..., B$:

3: Set

$$\pi_{t+1} = \frac{e^{f_{b-1}(x)}}{1 + e^{f_{b-1}(x)}}$$

4: For $t = 1, 2, ..., T$ compute

$$u_{t+1,b} = -\left[\frac{\partial L(y_{t+1}, f(x_t))}{\partial f(x_t)}\right]_{f=f_{b-1}} = y_{t+1} - \pi_{t+1}$$

5: Fit a regression tree on $u_{t+1,b}$, giving terminal regions $R_{j,b}$, $j = 1, 2, ..., J_b$

6: For $j = 1, 2, ..., J_b$ compute

$$\gamma_{j,b} = \arg\min_{\gamma} \sum_{x_t \in R_{j,b}} L(y_{t+1}, f_{b-1}(x_t) + \gamma) = \frac{\sum_{x_t \in R_{j,b}} u_{t+1,b}}{\sum_{x_t \in R_{j,b}} \pi_{t+1}(1 - \pi_{t+1})} \tag{10}$$

7: Update $f_b(x) = f_{b-1}(x) + \nu \sum_{j=1}^{J_b} \gamma_{j,b} I(x \in R_{j,b})$

8: Output $\hat{f}(x) = f_B(x)$

Gradient boosted trees also combines numerous decision trees, but they are constructed sequentially. Algorithm 1 outlines the gradient boosting algorithm for binary classification,

adapted from Hastie et al. (2009). The model begins with the log odds of the sample. At each new step b, the probability prediction is derived from the log odds via the sigmoid function in Equation (6). Subsequently, pseudo-residuals are computed as the difference between the actual observation and the predicted probability. As these pseudo-residuals are real numbers, a regression tree is then constructed on them. The output from each tree's leaf is determined by Equation (10). These outputs are added into the total model with a shrinkage weight ν, which controls the learning rate. This process continues until the ensemble comprises B trees. The final model is an additive combination of decision trees, whose output can be transformed into a probability prediction using the sigmoid function.

We also implement a regularization technique called stochastic gradient boosting, where we consider a random sub-sample of the data for each tree. This results in faster optimization, and a much lower variance at the expense of a slightly higher bias.

IA1.2.3 Support Vector Machines

Support Vector Machines (SVM) are popular nonlinear classifiers that have been shown to perform well for high-dimensional data with complex functional forms. Mathematically, SVM chooses parameters to solve the optimization problem:

$$\begin{aligned} \min_{\beta,\beta_0,\zeta} \frac{1}{2}||\beta||^2 + C\sum_{t=1}^{T}\zeta_t \\ \text{subject to } \tilde{y}_{t+1}(\beta'\phi(x_t)+\beta_0) \geq 1-\zeta_t, \\ \zeta_t \geq 0, t = 1,..,T, \end{aligned}$$

where $\tilde{y}_{t+1} = 2 \times y_{t+1} - 1$, $\tilde{y}_{t+1} \in \{-1, 1\}$, $\beta = (\beta_1, ..., \beta_P)$ is a P-dimensional parameter vector, and $\zeta = (\zeta_1, ..., \zeta_T)$ is a T-dimensional vector of slack variables. Intuitively, SVM seeks to find a hyperplane that maximizes the margin between classes (i.e., the distance separating the classes), which is equivalent to minimizing $\frac{1}{2}||\beta||^2$. Ideally, $\tilde{y}_{t+1}(\beta'\phi(x_t)+\beta_0)$ would be

greater than or equal to 1 for all observations, which would indicate a perfect classification for all samples. However, since the data is not always perfectly separable with a hyperplane, we can allow some of the observations to be on the wrong side of the hyperplane by distance ζ_t. Penalty term, C, is a hyperparameter that controls the strength of this penalty, and acts as an inverse regularization parameter.

Furthermore, the classes may not be linearly separable. To accommodate nonlinearities, we map x_t to a higher dimensional space with mapping function $\phi(\cdot)$. Notably, a kernel function $K(x_t, x_{t'}) = \langle \phi(x_t), \phi(x_{t'}) \rangle$ can be used to implicitly map the predictors to a higher-dimensional space in a computationally efficient manner, which is often referred to as the "kernel trick". We use the highly flexible radial basis function (RBF) kernel, which implicitly maps the predictors to an infinite-dimensional space, allowing SVM to be highly flexible.

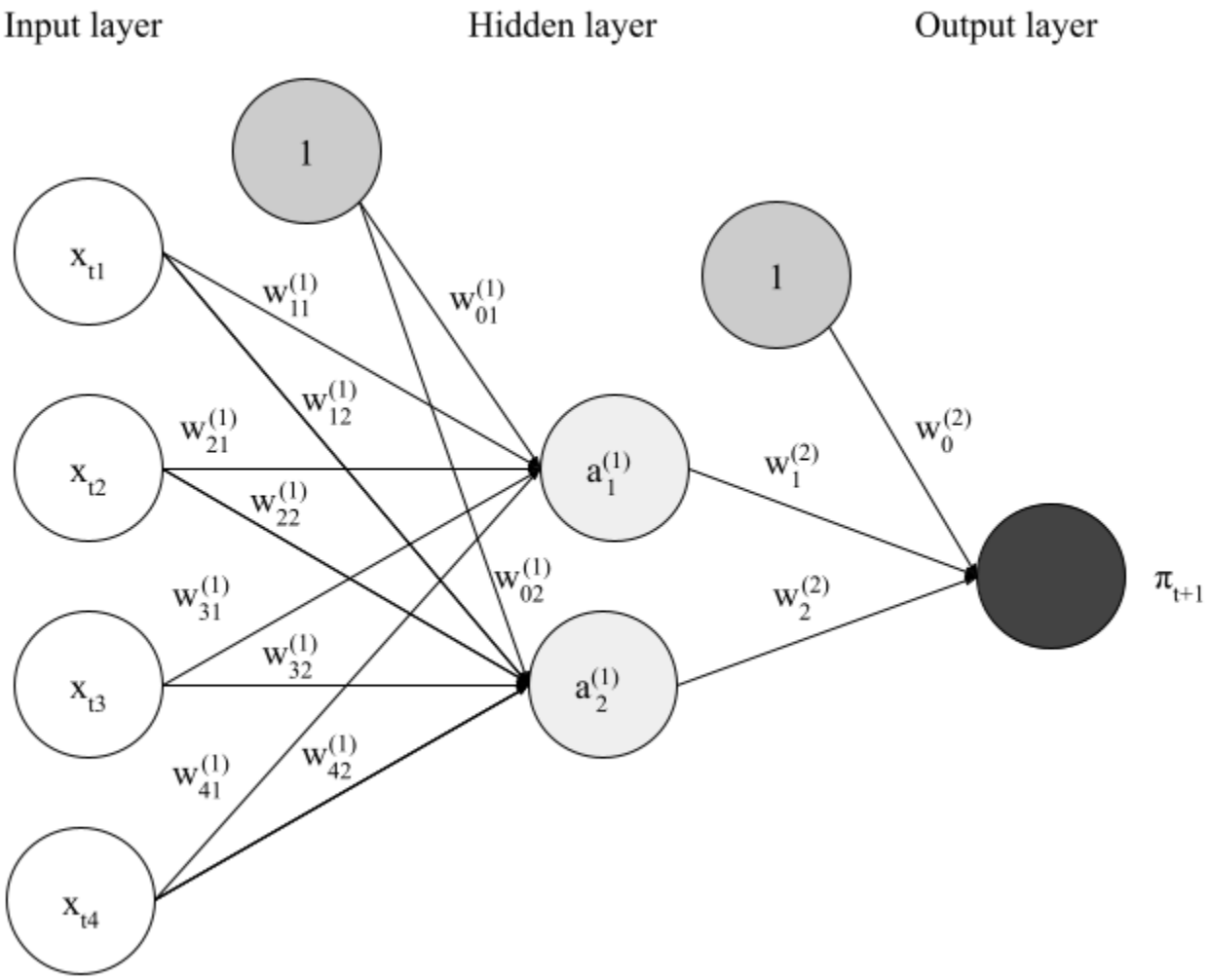
IA1.2.4 Feed-forward Neural Networks

Neural networks are powerful models that have theoretical underpinnings as universal approximators for any continuous function (Hornik et al., 1989). This section focuses on traditional feed-forward neural networks. Consider the simple feed-forward neural network with one input layer, one hidden layer, and one output layer depicted in Figure 14. Excluding the input layer, each node in each layer is a linear combination of predictors (similar to linear regression), which is then transformed by some nonlinear activation function. Each node $a_l^{(1)}$ in the hidden layer is given by

$$a_l^{(1)} = f\left(w_{0l}^{(1)} + \sum_{j=1}^{4} w_{jl}^{(1)} x_{tj}\right), \tag{11}$$

where $f(\cdot)$ is some nonlinear function (e.g. ReLU, tanh, ELU), and $w_{0l}^{(1)}$ is an intercept term. In our implementation we choose f to be the ReLU activation function. Then, the output node is given by

Figure 14: Example feed-forward neural network with one hidden layer



$$\pi_{t+1} = g\left(w_0^{(2)} + \sum_{l=1}^{2} w_l^{(2)} a_l^{(1)}\right), \tag{12}$$

where g is some activation function. For g, we use the sigmoid function in Equation (6) to forecast the probability of next month's excess return being positive.

Neural networks are trained with one of many algorithms that perform some version of stochastic gradient descent to minimize a loss function. We opt for the ADAM algorithm (Kingma and Ba, 2014) for its speed and efficiency. For classification we use the log loss function and minimize the following objective function:

$$\mathcal{L}(\theta) = -\frac{1}{T}\sum_{t=1}^{T}(y_{t+1}\log(\pi_{t+1}) + (1-y_{t+1})\log(1-\pi_{t+1})). \tag{13}$$

Parameter estimates $\hat{\theta}$ for the neural network are the solution to

$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}(\theta). \tag{14}$$

Estimating MT and DMT is similar to NN, except the objective function is the sum of the individual log losses across factors.

IA1.2.5 Long Short-Term Memory Neural Networks

Factors are time-varying and depend on short- and long-term financial and macroeconomic conditions, which cannot be captured by using predictors with only one lag. To handle this problem, we use LSTMs, which are capable of learning these long-term dependencies.

The LSTM, proposed by Hochreiter and Schmidhuber (1997) is composed of a cell state and three "gates", which control the flow of information inside the LSTM unit. The LSTM accommodates short-range and long-range dependence through three gate functions that control information flow from the previous hidden state, h_{t-1}, to h_t:

$$\begin{aligned} input_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i), \\ forget_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f), \\ output_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o), \end{aligned}$$

where W_a, U_a, and b_a are unknown parameters for $a \in \{i, f, o\}$, and the sigmoid function is an element-wise nonlinear transformation defined in Equation (6). Gate functions $input_t$, $forget_t$, and $output_t$ are sigmoid functions of the predictors x_t and previous hidden state

h_{t-1}.

Intuitively, the cell state, c_t, works like a conveyor belt, representing the long-term memory part of the LSTM unit, and is responsible for remembering the dependencies between the elements in the input sequence, allowing the network to memorize long-range information. Fresh information from h_{t-1} and the x_t arrive through $\tilde{c}_t$:

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c),$$

where W_c, U_c, and b_c are unknown parameters. Next, cell state, c_t, is updated as an additive function of c_{t-1} and $\tilde{c}_t$ given by

$$c_t = forget_t \odot c_{t-1} + input_t \odot \tilde{c}_t,$$

where $\odot$ denotes element-wise multiplication. Forget gate, $forget_t$, controls how much of the information from the past to forget. The input gate, $input_t$, controls the extent to which a new value flows into the cell state. Finally, the next hidden state is given by

$$h_t = output_t \odot \tanh(c_t),$$

where output gate, $output_t$, controls the extent to which new information is passed onto the next hidden state. Unlike MT, which directly uses x_t as inputs into the hard sharing layer, DMT first transforms x_t into the state processes h_t using LSTMs. These h_t are low-dimensional and shared across factors. Following this, DMT employs h_t as inputs for the hard sharing layer, effecting reducing dimensionality while simultaneously extracting nonlinear dynamics.

IA2 Shapley Values

The Shapley values for observation x_t are the weighted average marginal contribution of each predictor to the payoff $g(x_t) - E[g(X)]$ over all subset of predictors, where $g(x_t)$ is the output of a machine learning model and $E[g(X)]$ is the mean output of the machine learning model across all observations. Specifically, the Shapley value for an observation x_t and predictor j is given by

$$\phi_j(x_t) = \sum_{Q \subseteq S \setminus \{j\}} \underbrace{\frac{|Q|!(|S| - |Q| - 1)!}{|S|!}}_{\text{weight}} \underbrace{(\Delta_{Q \cup \{j\}}(x_t) - \Delta_Q(x_t))}_{\text{marginal contribution}}$$

where Q is a subset of predictors that does not contain predictor j, Δ_Q denotes the payoff under Q, and S is the complete set of predictors.

It can be shown that

$$\sum_{j=1}^{p} \phi_j(x_t) = g(x_t) - E[g(X)],$$

resulting in the natural linear interpretation of

$$g(x_t) = E[g(X)] + \phi_1(x_t) + ... + \phi_p(x_t).$$

Thus, we can interpret any prediction of machine learning model in a linear fashion, wherein the sum of the Shapley values across predictors equals the difference between the model's probability prediction for a specific observation and the average probability prediction across all observations.

Time complexity for computing exact Shapley values for p predictors is $O(2^p)$, so approximations are necessary. We adopt a Monte Carlo (MC) approach proposed by Štrumbelj and Kononenko (2014), whose details we provide in Algorithm 2.

Algorithm 2 Shapley Value Monte Carlo Algorithm

Input estimator g, predictors matrix X, predictor of interest j, observation x_{tj}, number of MC iterations M.
Initialize Shapley value $\phi(x_{tj}) = 0$
for $m = 1, 2, ..., M$:
    Generate a random permutation of feature indices, $\mathcal{O}$
    Sample at random another observation x_r
    Construct two new observations:

$$b_1 = \overbrace{\boxed{\text{preceding } j \text{ in } \mathcal{O}}\,\boxed{j}}^{\text{take values from } x_t}\,\overbrace{\boxed{\text{succeeding } j \text{ in } \mathcal{O}}}^{\text{take values from } x_r}$$

$$b_2 = \overbrace{\boxed{\text{preceding } j \text{ in } \mathcal{O}}}^{\text{take values from } x_t}\,\overbrace{\boxed{j}\,\boxed{\text{succeeding } j \text{ in } \mathcal{O}}}^{\text{take values from } x_r}$$

    $\phi(x_{tj}) = \phi(x_{tj}) + g(b_1) - g(b_2)$
$\phi(x_{tj}) = \frac{\phi(x_{tj})}{M}$
Output $\phi(x_{tj})$

IA3 Hyperparameters

Table IA1 shows the grids employed in the hyperparameter optimization for each model.

For EN, we tune λ and ρ. λ determines the strength of the elastic net penalty, with higher values corresponding to larger penalties. ρ manages the balance between ℓ_1 and ℓ_2 penalties.

For RF and XRF, we tune the number of trees and the maximum depth of each tree. Increasing the number of trees in the ensemble helps to lower variance. Conversely, increasing the maximum depth of the trees tends to reduce bias. Deeper trees result in a more finely partitioned predictor space, leading to fewer samples in each leaf node and consequently, lower bias. Bias in a specific tree would be minimized if it perfectly matched the training data, with one sample in each leaf node. However, this approach leads to higher variance, highlighting a trade-off between bias and variance. Following the approach in Gu et al. (2020), we explore tree depths ranging from 1 to 6. Additionally, we set the number of randomly selected predictors at each split to $\sqrt{p}$, a strategy empirically found to be effective.

For GBT, we tune the learning rate ν, the number of trees, the sub-sample size, and the maximum depth of trees. Unlike RF, constructs trees sequentially, with each new tree refining the predictions of its predecessor. The learning rate acts as a shrinkage factor for new trees, limiting the extent of correction each tree contributes, akin to the function of the learning rate in gradient descent. The sub-sample size represents the fraction of the training data that is randomly selected for sampling. Constructing subsequent trees on random sub-samples of the training data reduces the likelihood of the ensemble model overfitting on the training set and also decreases computational time. Maximum depth in GBT serves a similar purpose as in RF and XRF. Aligning with the approach detailed in Gu et al. (2020), we opt for shallow trees, focusing on depths between 1 and 2.

For SVM, we tune the C and γ. C acts as an inverse regularization hyperparameter. Lower values of C force the decision surface to be smooth by by permitting some misclassifications, whereas higher values of C strive to classify all training samples correctly. γ is a parameter in the RBF kernel function, dictating the extent of influence a single training example exerts.

Our neural networks follow similar regularization approaches as Gu et al. (2020). We utilize early stopping, which ends the parameter estimation when the validation set error begins to increase. Early stopping is also a popular substitute to ℓ_2 penalization, because it performs shrinkage at a much lower computational cost. We also include ℓ_1 penalization to encourage sparsity. Next, we use batch normalization to control the variability of predictors across different regions in the network. We also adopt an ensemble approach, by combining the predictions of multiple neural networks estimated with different random seeds.

Architecturally, NN comprises three hidden layers, each with 32 units. MT's structure includes a single fully connected hard sharing layer with 32 units, followed by two factor-specific layers for each factor, each containing eight units. DMT incorporates two separate LSTMs for macroeconomic and financial variables, both with 16 units, which then feed into a fully connected hard sharing layer with 32 units. Subsequently, each factor in DMT has

Table IA1: Hyperparameters for all mModels

Model	Hyperparameters
EN	$\lambda \in (10^{-4}, 10^{3})$ $\alpha \in (0.0, 1.0)$
RF and XRF	#Trees $\in \{50, 100, 200, 500, 1000\}$ #Features $= \sqrt{p}$ Depth $= 1 \sim 6$
GBT	Learning Rate $\in \{10^{-3}, 10^{-2}, 10^{-1}\}$ #Trees $\in \{50, 100, 200\}$ Subsample $\in \{0.25, 0.5, 1\}$ Depth $= 1 \sim 2$
SVM	C $\in (10^{-3}, 10^{3})$ $\gamma \in (10^{-3}, 10^{3})$ Kernel = rbf
NN, MT, and DMT	$l_1 \in (0, 10^{-2})$ Batch size = 4 Adam Params. = default Learning Rate $\in \{10^{-4}, 10^{-3}\}$ Epochs = 500 Patience = 10 Ensemble = 10

two factor-specific layers, each with eight units.

IA4 Data

Table IA2: Financial Variables

Predictor	Authors	Year	Journal	Category
asset growth	Cooper, Gulen and Schill	2008	JF	Investment
change in capex (two years)	Anderson and Garcia-Feijoo	2006	JF	Investment
change in capex (three years)	Anderson and Garcia-Feijoo	2006	JF	Investment
change in current operating assets	Richardson et al.	2005	JAE	Investment
change in current operating liabilities	Richardson et al.	2005	JAE	Investment
change in equity to assets	Richardson et al.	2005	JAE	Investment
change in long-term investment	Richardson et al.	2005	JAE	Investment
exchange switch	Dharan and Ikenberry	1995	JF	Investment
growth in book equity	Lockwood and Prombutr	2010	JFR	Investment
growth in long term operating assets	Fairfield, Whisenant and Yohn	2003	AR	Investment
inventory change	Thomas and Zhang	2002	RAS	Investment
inventory growth	Belo and Lin	2012	RFS	Investment
revenue growth rank	Lakonishok, Shleifer, Vishny	1994	JF	Investment
advertising expense	Chan, Lakonishok and Sougiannis	2001	JF	Leverage
book leverage (annual)	Fama and French	1992	JF	Leverage
change in financial liabilities	Richardson et al.	2005	JAE	Leverage
change in net financial assets	Richardson et al.	2005	JAE	Leverage
change in net operating assets	Hirshleifer, Hou, Teoh, Zhang	2004	JAE	Leverage
change in ppe and inv/assets	Lyandres, Sun and Zhang	2008	RFS	Leverage
composite debt issuance	Lyandres, Sun and Zhang	2008	RFS	Leverage
convertible debt indicator	Valta	2016	JFQA	Leverage
firm age based on crsp	Barry and Brown	1984	JFE	Leverage
industry concentration (assets)	Hou and Robinson	2006	JF	Leverage
industry concentration (equity)	Hou and Robinson	2006	JF	Leverage
industry concentration (sales)	Hou and Robinson	2006	JF	Leverage
investment to revenue	Titman, Wei and Xie	2004	JFQA	Leverage
leverage component of bm	Penman, Richardson and Tuna	2007	JAR	Leverage
market leverage	Bhandari	1988	JF	Leverage
net debt to price	Penman, Richardson and Tuna	2007	JAR	Leverage
net operating assets	Hirshleifer et al.	2004	JAE	Leverage
operating leverage	Novy-Marx	2011	ROF	Leverage
organizational capital	Eisfeldt and Papanikolaou	2013	JF	Leverage
r&d ability	Cohen, Diether and Malloy	2013	RFS	Leverage
r&d over market cap	Chan, Lakonishok and Sougiannis	2001	JF	Leverage
sin stock (selection criteria)	Hong and Kacperczyk	2009	JFE	Leverage
tangibility	Hahn and Lee	2009	JF	Leverage
unexpected r&d increase	Eberhart, Maxwell and Siddique	2004	JF	Leverage
amihud's illiquidity	Amihud	2002	JFM	Liquidity
bid-ask spread	Amihud and Mendelsohn	1986	JFE	Liquidity
days with zero trades (1 month)	Liu	2006	JFE	Liquidity
days with zero trades (6 months)	Liu	2006	JFE	Liquidity
days with zero trades (12 months)	Liu	2006	JFE	Liquidity
inst own and idio vol	Nagel	2005	JFE	Liquidity
inst own and market to book	Nagel	2005	JFE	Liquidity
inst own and turnover	Nagel	2005	JFE	Liquidity
past trading volume	Brennan, Chordia, Subra	1998	JFE	Liquidity
share turnover volatility	Chordia, Subra, Anshuman	2001	JFE	Liquidity
share volume	Datar, Naik and Radcliffe	1998	JFM	Liquidity
size	Banz	1981	JFE	Liquidity
volume to market equity	Haugen and Baker	1996	JFE	Liquidity

Table IA2: Financial Variables (continued)

Predictor	Authors	Year	Journal	Category
volume trend	Haugen and Baker	1996	JFE	Liquidity
volume variance	Chordia, Subra, Anshuman	2001	JFE	Liquidity
52 week high	George and Hwang	2004	JF	Price Trend
industry momentum	Grinblatt and Moskowitz	1999	JF	Price Trend
industry return of big firms	Hou	2007	RFS	Price Trend
intermediate momentum	Novy-Marx	2012	JFE	Price Trend
long-run reversal	De Bondt and Thaler	1985	JF	Price Trend
maximum return over month	Bali, Cakici, and Whitelaw	2011	JFE	Price Trend
medium-run reversal	De Bondt and Thaler	1985	JF	Price Trend
momentum (6 month)	Jegadeesh and Titman	1993	JF	Price Trend
momentum (12 month)	Jegadeesh and Titman	1993	JF	Price Trend
momentum and lt reversal	Chan and Ko	2006	JOIM	Price Trend
momentum based on ff3 residuals	Blitz, Huij and Martens	2011	JEmpFin	Price Trend
momentum in high volume stocks	Lee and Swaminathan	2000	JF	Price Trend
momentum without the seasonal part	Heston and Sadka	2008	JFE	Price Trend
off season long-term reversal	Heston and Sadka	2008	JFE	Price Trend
off season reversal years 6 to 10	Heston and Sadka	2008	JFE	Price Trend
off season reversal years 11 to 15	Heston and Sadka	2008	JFE	Price Trend
off season reversal years 16 to 20	Heston and Sadka	2008	JFE	Price Trend
price	Blume and Husic	1973	JF	Price Trend
price delay coeff	Hou and Moskowitz	2005	RFS	Price Trend
price delay r square	Hou and Moskowitz	2005	RFS	Price Trend
price delay se adjusted	Hou and Moskowitz	2005	RFS	Price Trend
return seasonality last year	Heston and Sadka	2008	JFE	Price Trend
return seasonality years 2 to 5	Heston and Sadka	2008	JFE	Price Trend
return seasonality years 6 to 10	Heston and Sadka	2008	JFE	Price Trend
return seasonality years 11 to 15	Heston and Sadka	2008	JFE	Price Trend
return seasonality years 16 to 20	Heston and Sadka	2008	JFE	Price Trend
short term reversal	Jegadeesh	1990	JF	Price Trend
trend factor	Han, Zhou, Zhu	2016	JFE	Price Trend
accruals	Sloan	1996	AR	Profitability
cash productivity	Chandrashekar and Rao	2009	WP	Profitability
cash-flow to price variance	Haugen and Baker	1996	JFE	Profitability
change in capital inv (ind adj)	Abarbanell and Bushee	1998	AR	Profitability
change in net noncurrent op assets	Soliman	2008	AR	Profitability
change in net working capital	Soliman	2008	AR	Profitability
change in taxes	Thomas and Zhang	2011	JAR	Profitability
dividend initiation	Michaely, Thaler and Womack	1995	JF	Profitability
dividend omission	Michaely, Thaler and Womack	1995	JF	Profitability
dividend seasonality	Hartzmark and Salomon	2013	JFE	Profitability
earnings consistency	Alwathainani	2009	BAR	Profitability
earnings streak length	Loh and Warachka	2012	MS	Profitability
earnings surprise	Foster, Olsen and Shevlin	1984	AR	Profitability
earnings surprise of big firms	Hou	2007	RFS	Profitability
net income / book equity	Haugen and Baker	1996	JFE	Profitability
operating profits / book equity	Fama and French	2006	JFE	Profitability
percent operating accruals	Hafzalla, Lundholm, Van Winkle	2011	AR	Profitability
revenue surprise	Jegadeesh and Livnat	2006	JAE	Profitability
sales growth over inventory growth	Abarbanell and Bushee	1998	AR	Profitability
sales growth over overhead growth	Abarbanell and Bushee	1998	AR	Profitability

Table IA2: Financial Variables (continued)

Predictor	Authors	Year	Journal	Category
total accruals	Richardson et al.	2005	JAE	Profitability
cash-based operating profitability	Ball et al.	2016	JFE	Quality
change in asset turnover	Soliman	2008	AR	Quality
efficient frontier index	Nguyen and Swanson	2009	JFQA	Quality
gross profits / total assets	Novy-Marx	2013	JFE	Quality
operating profitability r&d adjusted	Ball et al.	2016	JFE	Quality
capm beta	Fama and MacBeth	1973	JPE	Risk
coskewness	Harvey and Siddique	2000	JF	Risk
coskewness using daily returns	Ang, Chen and Xing	2006	RFS	Risk
frazzini-pedersen beta	Frazzini and Pedersen	2014	JFE	Risk
idiosyncratic risk (3 factor)	Ang et al.	2006	JF	Risk
idiosyncratic risk (aht)	Ali, Hwang, and Trombley	2003	JFE	Risk
idiosyncratic skewness (3f model)	Bali, Engle and Murray	2015	Book	Risk
realized (total) volatility	Ang et al.	2006	JF	Risk
return skewness	Bali, Engle and Murray	2015	Book	Risk
tail risk beta	Kelly and Jiang	2014	RFS	Risk
book to market using december me	Fama and French	1992	JPM	Value
book to market, original	Stattman	1980	Other	Value
cash flow to market	Lakonishok, Shleifer, Vishny	1994	JF	Value
composite equity issuance	Daniel and Titman	2006	JF	Value
earnings-to-price ratio	Basu	1977	JF	Value
enterprise component of bm	Penman, Richardson and Tuna	2007	JAR	Value
enterprise multiple	Loughran and Wellman	2011	JFQA	Value
equity duration	Dechow, Sloan and Soliman	2004	RAS	Value
intangible return using cftop	Daniel and Titman	2006	JF	Value
intangible return using ep	Daniel and Titman	2006	JF	Value
intangible return using sale2p	Daniel and Titman	2006	JF	Value
net payout yield	Boudoukh et al.	2007	JF	Value
operating cash flows to price	Desai, Rajgopal, Venkatachalam	2004	AR	Value
payout yield	Boudoukh et al.	2007	JF	Value
predicted div yield next month	Litzenberger and Ramaswamy	1979	JF	Value
sales-to-price	Barbee, Mukherji and Raines	1996	FAJ	Value
share issuance (1 year)	Pontiff and Woodgate	2008	JF	Value
share issuance (5 year)	Daniel and Titman	2006	JF	Value
spinoffs	Cusatis, Miles and Woolridge	1993	JFE	Value
taxable income-to-book income	Lev and Nissim	2004	AR	Value
total assets to market	Fama and French	1992	JF	Value
book-to-market ratio (wg)	Welch and Goyal	2008	RFS	Welch and Goyal
default yield spread (wg)	Welch and Goyal	2008	RFS	Welch and Goyal
dividend payout ratio (wg)	Welch and Goyal	2008	RFS	Welch and Goyal
dividend price ratio (wg)	Welch and Goyal	2008	RFS	Welch and Goyal
dividend yield (wg)	Welch and Goyal	2008	RFS	Welch and Goyal
earnings price ratio (wg)	Welch and Goyal	2008	RFS	Welch and Goyal
long term rate of return (wg)	Welch and Goyal	2008	RFS	Welch and Goyal
long term yield (wg)	Welch and Goyal	2008	RFS	Welch and Goyal
net equity expansion (wg)	Welch and Goyal	2008	RFS	Welch and Goyal
stock variance (wg)	Welch and Goyal	2008	RFS	Welch and Goyal
term spread (wg)	Welch and Goyal	2008	RFS	Welch and Goyal
treasury bill rate (wg)	Welch and Goyal	2008	RFS	Welch and Goyal

Table IA3: Macroeconomic Variables

Predictor	FRED-MD Acronym	Category
new orders for durable goods	AMDMNOx	Consumption
real manu. and trade industries sales	CMRMTSPLx	Consumption
real personal consumption expenditures	DPCERA3M086SBEA	Consumption
retail and food services sales	RETAILx	Consumption
total business inventories	BUSINVx	Consumption
total business: inventories to sales ratio	ISRATIOx	Consumption
unfilled orders for durable goods	AMDMUOx	Consumption
housing starts, midwest	HOUSTMW	Housing
housing starts, northeast	HOUSTNE	Housing
housing starts, south	HOUSTS	Housing
housing starts, west	HOUSTW	Housing
housing starts: total new privately owned	HOUST	Housing
new private housing permits (saar)	PERMIT	Housing
new private housing permits, midwest (saar)	PERMITMW	Housing
new private housing permits, northeast (saar)	PERMITNE	Housing
new private housing permits, south (saar)	PERMITS	Housing
new private housing permits, west (saar)	PERMITW	Housing
all employees: construction	USCONS	Labor
all employees: durable goods	DMANEMP	Labor
all employees: financial activities	USFIRE	Labor
all employees: goods-producing industries	USGOOD	Labor
all employees: government	USGOVT	Labor
all employees: manufacturing	MANEMP	Labor
all employees: mining and logging: mining	CES1021000001	Labor
all employees: nondurable goods	NDMANEMP	Labor
all employees: retail trade	USTRADE	Labor
all employees: service-providing industries	SRVPRD	Labor
all employees: total nonfarm	PAYEMS	Labor
all employees: trade, transportation & utilities	USTPU	Labor
all employees: wholesale trade	USWTRADE	Labor
average duration of unemployment (weeks)	UEMPMEAN	Labor
avg hourly earnings : construction	CES2000000008	Labor
avg hourly earnings : goods-producing	CES0600000008	Labor
avg hourly earnings : manufacturing	CES3000000008	Labor
avg weekly hours : goods-producing	CES0600000007	Labor
avg weekly hours : manufacturing	AWHMAN	Labor
avg weekly overtime hours : manufacturing	AWOTMAN	Labor
civilian employment	CE16OV	Labor
civilian labor force	CLF16OV	Labor
civilian unemployment rate	UNRATE	Labor

Table IA3: Macroeconomic Variables (continued)

Predictor	FRED-MD Acronym	Category
civilians unemployed - 15 weeks & over	UEMP15OV	Labor
civilians unemployed - less than 5 weeks	UEMPLT5	Labor
civilians unemployed for 15-26 weeks	UEMP15T26	Labor
civilians unemployed for 27 weeks and over	UEMP27OV	Labor
civilians unemployed for 5-14 weeks	UEMP5TO14	Labor
help-wanted index for united states	HWI	Labor
initial claims	CLAIMSx	Labor
ratio of help wanted/no. unemployed	HWIURATIO	Labor
commercial and industrial loans	BUSLOANS	Money
consumer motor vehicle loans outstanding	DTCOLNVHFNM	Money
m1 money stock	M1SL	Money
m2 money stock	M2SL	Money
monetary base	BOGMBASE	Money
nonrevolving consumer credit to personal income	CONSPI	Money
real estate loans at all commercial banks	REALLN	Money
real m2 money stock	M2REAL	Money
reserves of depository institutions	NONBORRES	Money
securities in bank credit at all commercial banks	INVEST	Money
total consumer loans and leases outstanding	DTCTHFNM	Money
total nonrevolving credit	NONREVSL	Money
total reserves of depository institutions	TOTRESNS	Money
capacity utilization: manufacturing	CUMFNS	Output
ip index	INDPRO	Output
ip: business equipment	IPBUSEQ	Output
ip: consumer goods	IPCONGD	Output
ip: durable consumer goods	IPDCONGD	Output
ip: durable materials	IPDMAT	Output
ip: final products (market group)	IPFINAL	Output
ip: final products and nonindustrial supplies	IPFPNSS	Output
ip: fuels	IPFUELS	Output
ip: manufacturing (sic)	IPMANSICS	Output
ip: materials	IPMAT	Output
ip: nondurable consumer goods	IPNCONGD	Output
ip: nondurable materials	IPNMAT	Output
ip: residential utilities	IPB51222S	Output
real personal income	RPI	Output
real personal income ex transfer receipts	W875RX1	Output
cpi : all items	CPIAUCSL	Inflation
cpi : all items less food	CPIULFSL	Inflation
cpi : all items less medical care	CUSR0000SA0L5	Inflation

Table IA3: Macroeconomic Variables (continued)

Predictor	FRED-MD Acronym	Category
cpi : all items less shelter	CUSR0000SA0L2	Inflation
cpi : apparel	CPIAPPSL	Inflation
cpi : commodities	CUSR0000SAC	Inflation
cpi : durables	CUSR0000SAD	Inflation
cpi : medical care	CPIMEDSL	Inflation
cpi : services	CUSR0000SAS	Inflation
cpi : transportation	CPITRNSL	Inflation
crude oil, spliced wti and cushing	OILPRICEx	Inflation
ppi: crude materials	WPSID62	Inflation
ppi: finished consumer goods	WPSFD49502	Inflation
ppi: finished goods	WPSFD49207	Inflation
ppi: intermediate materials	WPSID61	Inflation
ppi: metals and metal products:	PPICMM	Inflation
personal cons. exp: durable goods	DDURRG3M086SBEA	Inflation
personal cons. exp: nondurable goods	DNDGRG3M086SBEA	Inflation
personal cons. exp: services	DSERRG3M086SBEA	Inflation
personal cons. expend.: chain index	PCEPI	Inflation
1-year treasury c minus fedfunds	T1YFFM	Rates
1-year treasury rate	GS1	Rates
10-year treasury c minus fedfunds	T10YFFM	Rates
10-year treasury rate	GS10	Rates
3-month aa financial commercial paper rate	CP3Mx	Rates
3-month commercial paper minus fedfunds	COMPAPFFx	Rates
3-month treasury bill:	TB3MS	Rates
3-month treasury c minus fedfunds	TB3SMFFM	Rates
5-year treasury c minus fedfunds	T5YFFM	Rates
5-year treasury rate	GS5	Rates
6-month treasury bill:	TB6MS	Rates
6-month treasury c minus fedfunds	TB6SMFFM	Rates
canada / u.s. foreign exchange rate	EXCAUSx	Rates
effective federal funds rate	FEDFUNDS	Rates
japan / u.s. foreign exchange rate	EXJPUSx	Rates
moody's baa corporate bond minus fedfunds	BAAFFM	Rates
moody's aaa corporate bond minus fedfunds	AAAFFM	Rates
moody's seasoned aaa corporate bond yield	AAA	Rates
moody's seasoned baa corporate bond yield	BAA	Rates
switzerland / u.s. foreign exchange rate	EXSZUSx	Rates
u.s. / u.k. foreign exchange rate	EXUSUKx	Rates
s&p's common stock price index: composite	S&P 500	Stocks
s&p's common stock price index: industrials	S&P: indust	Stocks
s&p's composite common stock: dividend yield	S&P div yield	Stocks
s&p's composite common stock: price-earnings ratio	S&P PE ratio	Stocks
vix	VIXCLSx	Stocks