Machine learning and return predictability across firms, time and portfolios

Fahiz Baba-Yara*
Nova SBE

November 20, 2020

Latest Version

Abstract

Previous research finds that machine learning methods predict short-term return variation in the cross-section of stocks, even when these methods do not impose strict economic restrictions. However, without such restrictions, the models' predictions fail to generalize in a number of important ways, such as predicting time-series variation in returns to the market portfolio and long-short characteristic sorted portfolios. I show that this shortfall can be remedied by imposing restrictions, that reflect findings in the financial economics literature, in the architectural design of a neural network model and provide recommendations for using machine learning methods in asset pricing. Additionally, I study return predictability over multiple future horizons, thus shedding light on the dynamics of intermediate and long-run conditional expected returns.

Keywords: Return Predictability, Long-run Stock Returns, Machine Learning, Neural Networks.
JEL Classification: E44, G10, G11, G12, G17

*Nova School of Business and Economics, Campus de Carcavelos, 2775-405 Carcavelos, Portugal. E-mail: 25795@novasbe.pt Web: https://babayara.com/

☆ I am especially grateful to Martijn Boons and Andrea Tamoni for their guidance and comments. I thank Melissa Prado, Giorgio Ottonello, Fernando Anjos, Irem Demirci, Miguel Ferreira, Andre Castro Silva and participants at Nova SBE research seminar for helpful comments. This work was funded by FCT Portugal. The author has no conflicts of interest to disclose and all remaining errors are mine.

Introduction

In this paper, I study how incorporating economic priors in the specification of a machine learning model improves the resulting model's predictive accuracy with respect to predicting equity returns. I do this by comparing the forecasting accuracy of a neural network model that imposes several restrictions reflecting recent findings in the financial economics literature to an alternative neural network model that is much simpler in its structure. The simple model I benchmark against is the best performing neural network model, as specified in Gu et al. (2020b). I study the predictability of returns to individual US equities, 56 long-short characteristic sorted-portfolios, and the value-weighted market portfolio over multiple horizons.

This paper shows that the stylized facts from the financial economics literature have an integral role to play in guiding the application of machine learning to finance. This conclusion stems from the number and nature of improvements one observes when comparing my proposed model (the economically restricted model) to the benchmark model. First, in predicting individual equity returns, I find that forecasts from the benchmark model explain about 0.58% (out-of-sample R^2) of the variation in next month's returns, whereas forecasts from the economically restricted model can explain about 0.99%—close to a two-fold increase.[1]

Second, I show investors who employ return forecasts from the economically restricted neural network model enjoy large and robust economic gains. For instance, using individual equity return predictions' for the following month, a long-short portfolio that buys (sells) the 10% highest (lowest) expected return stocks has an annualized average return of 15.82%. This estimate corresponds to a Sharpe ratio of 0.78, a certainty equivalent of 11.73, and a Fama and French (2018) 6 factor-alpha of 11.05%, all annualized. This strategy only trades the 500 largest market-capitalized firms each month, thus generating the gains from the most liquid stocks in the cross-section. The annualized average return of the strategy falls to about 1% when one uses forecasts from the benchmark model.

This result is particularly interesting because it shows that the economically restricted model's improved predictive accuracy is not concentrated among small stocks. Given the

[1] All results pertain to the out-of-sample period; January 1995 to December 2018.

small annualized return the strategy produces when one condition on forecasts from the benchmark model, it is evident that the benchmark model extracts a non-trivial fraction of its predictive accuracy from small and difficult to arbitrage stocks (see Avramov et al. (2020)).

Third, I find that to produce forecasts that robustly generalize beyond individual equities, restrictions implied by findings from the financial economics literature are crucial. When predicting returns to the value-weighted market portfolio, forecasts from the economically restricted model predict time-series variation in monthly returns as far out as three years in the future. Additionally, when predicting returns to 56 long-short characteristic sorted portfolios, the aggregate forecasts predict time-series variation in next month's returns for 53 of the 56 portfolios. On the other hand, forecasts from the benchmark model fail to robustly predict time-series variation in returns to both the aggregate market and long-short characteristic sorted portfolios.

A natural question one would ask at this point is, "Along which dimension does the economically restricted model help improve stock return forecasts?" This question has so far received little attention in this emerging literature. I shed light on this by decomposing stock return forecasts into two components; a fraction explaining variations in a level factor (the equally-weighted market return) and a fraction explaining variations in returns over and above the cross-sectional mean (relative stock returns).[2]

I find that the improvement in forecasting accuracy primarily comes from predicting better the relative stock return component. For this component, forecasts from the economically restricted model explain about 0.55% of the variations in next month's returns, while the benchmark model only explains 0.16%-a more than three-fold improvement. The models are comparable in their ability to explain variations in the level component. Forecasts from the economically restricted model explain about 0.43 % of the variations in the level factor, and the benchmark model explains 0.41 %. .

Most papers in the literature study cross-sectional return predictability over the next month or at most over the following year (see Kozak (2019), Gu et al. (2020b), and Freyberger et al. (2020)). Papers that study returns further into the future, conditional

[2]I decompose returns ($r_{i,t}$) into a level factor; captured by the cross-sectional average return, $(N_t^{-1})\sum_{i\epsilon t} r_{i,t}$ and a slope factor; captured by the cross-sectional dispersion around the mean, $r_{i,t}^{RR} = r_{i,t} - (N_t^{-1})\sum_{i\epsilon t} r_{i,t}$.

on what we know today, tend to study exclusively the time-series properties of returns with annual holding periods. Therefore, this paper is among the first to study monthly stock return predictability as far as ten years into the future and documents new evidence on the cross-sectional and time-series properties of conditional expected stock returns across horizons. The two main results are that 1) stock return predictability decreases over the horizon and 2) the nature of stock predictability in the short run is very different from the long-run.

First, stock returns are much more predictable in the short-run than in the long-run. Forecasts from the economically restricted neural network model can explain about 0.99% (out-of-sample R^2) of the variations in next month's stock returns. However, this estimate falls to about 0.28% when predicting returns five years into the future and about 0.13% when predicting returns ten years into the future.

Second, the nature of short-run stock return predictability is very different from long-run stock return predictability. Accounting for the inherent factor structure underpinning stock returns, I find that a large fraction of the observed stock return predictability across horizons comes from predicting variations in the equally-weighted market return or level component in the pool of stocks. When predicting next month's return, about 43% of the variation explained (0.43% out of 0.99%) comes from explaining variations in the component factor. For forecasts that pertain to months that are at least one year in the future, over 95% of the variation explained comes from explaining variations in the level factor.

To summarize, I find that the forecasts' ability to explain cross-sectional variation in returns is only present in the short-run. While in the long-run, stock return predictability entirely comes from predicting variations in the level component in the pool of stocks.

The empirical asset pricing literature has shown that firm characteristics are correlated with subsequent stock returns, but evidence on how well these characteristics or combinations thereof proxy for conditional expected returns is scarce (see for example, Basu (1977), Jegadeesh and Titman (1993), and Sloan (1996)). Measuring relative stock returns as the stock return in excess of the cross-sectional mean, I find that a one percent relative return forecast on average predicts a 0.97 percentage point increase in next month's relative stock return. Similar to the conclusions drawn from the out-of-sample

R^2 analysis, I find that this estimate decreases as the horizon increases. In predicting the monthly relative stock return realized one-year in the future, this estimate falls to 0.79 and further down to 0.24 when predicting monthly relative returns two years in the future.

I find similarly robust estimates for the value-weighted market portfolio and long-short characteristic sorted portfolio returns. On average, a one percent demeaned market forecast predicts a 1.60 percentage point increase in market return. The estimates are statistically significant for monthly forecasts for up to three years in the future. A one percent demeaned long-short portfolio forecast, on average, predicts about a 2.00 percentage point increase in long-short portfolio returns. The estimates are generally statistically significant for forecasts up to a year into the future.

The findings in this paper are both important and interesting for the following reasons. First, conditional expected returns for multiple future dates are analogous to a term structure of implied discount rates (cost of capital) conditional on an information set observed today. These discount rates are of particular importance to firms when evaluating investment opportunities with cash-flows maturing over multiple future dates. The proposed model in this paper is one way of using project-specific characteristics observed today as a basis for coming up with consistent (implied) discount rates to help evaluate such investment opportunities.

Second, the return predictability literature is only beginning to tackle the question of whether or not long-short characteristic sorted portfolio returns are predictable over time (see Baba-Yara, Boons, and Tamoni (2018) and Haddad, Kozak, and Santosh (2020)). Reporting the Sharpe ratio of such portfolios only tells us that the cross-sectional variation in the characteristic generates an unconditional spread in returns but not whether the time-series variation in the returns to such portfolios is predictable. I find that the economically restricted neural network forecast can predict time-series variation in next month's return for over 90% of the long-short portfolios I study. This result is important because the returns to factor portfolios can be low or negative for prolonged periods (see Israel et al. (2020)). Having access to conditional expected return estimates for these portfolios should aid investors in portfolio timing decisions. More generally, improving short- and long-run expected return estimates is essential because these estimates serve

as fundamental inputs in tactical and strategic portfolio decisions, respectively.

Finally, Martin and Nagel (2019) consider a world where agents have to condition on thousands of potentially relevant variables to forecast returns. If agents (investors) are uncertain about how exactly cash-flows relate to these predictors, then a factor zoo will naturally emerge. A world not too dissimilar from our own. Bryzgalova et al. (2019) show that complex non-linearities exist between firm characteristics and stock returns. Taking these two stlyized facts together, agents will need learning algorithms that can efficiently handle large dimensional predictors while simultaneously learning the non-linearities that exist therein. Gu et al. (2020b) show that neural networks are the best learning algorithm for this problem. This paper shows that incorporating economic restrictions in the neural network design robustly enhances their predictive ability.

Literature

This work is related to the emerging literature in economics and finance using machine learning methods to answer economic questions that are fundamentally predictive. Sirignano, Sadhwani, and Giesecke (2016) show that deep neural networks are strong predictors of mortgage repayment, delinquency, and foreclosures. Butaru et al. (2016) use regression trees to predict the probability of consumer credit card delinquencies and defaults. Freyberger, Neuhierl, and Weber (2020) use the adaptive group LASSO to study which subset of 62 characteristics provides incremental information about the cross-section of expected returns. The spline methodology the authors use cannot easily accommodate higher-order interactions between covariates (characteristics). However, deep neural networks, the learning algorithm used in this paper, easily approximates higher-order non-linear interactions between covariates (see Goodfellow et al. (2016)). Chen, Pelger, and Zhu (2019) estimate the stochastic discount factor using neural networks and find that it outperforms all other benchmarks in an out-of-sample setting. Like these authors, I show that designing neural network models using financial economic priors does generate robust forecasts, although the nature of my proposed model is very different. The economically restricted neural network model I propose is similar to the autoencoder model of Gu et al. (2020a). Gu et al. (2020a) primarily study the asset pricing implica-

tions of their model for next month returns, I study return predictability across time and portfolios.

This work primarily extends the literature on stock return predictability. I show that a neural network architecture design that imposes restrictions reflecting findings in the financial economics literature improves stock return forecasts out-of-sample. Lewellen (2015) studies expected returns across stocks as a linear function of firm-level characteristics and finds that the forecasts generated by the linear model explain some variation in returns. The proposed framework in this paper allows for high-dimensional non-linear interactions between characteristics and also imposes a Lasso penalty to remove non-essential return predictors in the information set I condition on. Gu, Kelly, and Xiu (2020b) show that allowing for non-linear interactions between characteristics help improve the forecasting accuracy of ML models. Specifically, the authors show that firm-level characteristics can be combined with macroeconomic variables using different machine learning methods to predict returns better. I show that the information set we condition on is not only informative of return realizations for the next month but extends much further out into the future. This finding is important because Van Binsbergen and Opp (2019) argue only characteristics that predict persistently generate substantial economic distortions. Finally, I show that relative stock return predictability is short-lived. Specifically, machine learning forecasts for return realizations beyond one year into the future are no better than a zero forecast in discriminating between high and low expected return firms; a result that suggests that longer-run discount rates converge across firms (see Keloharju, Linnainmaa, and Nyberg (2019)).

The results in this paper also contribute to the literature that studies aggregate market return predictability. Cochrane (2008) studies market return predictability and provides evidence that the dividend-yield predicts time-series variation in the equity risk premium. Goyal and Welch (2008) study market return predictability in the time-series using macroeconomic variables and show that the historical average market return is a challenging benchmark to beat. I show that a neural network model that adheres to economic theory robustly out-performs the historical equity return in predicting time-series variation in monthly market returns as far as three years into the future. Engelberg et al. (2019) aggregate 140 individual firm-characteristics, including the dividend-yield,

and ask how many of these aggregates can predict market returns. The authors find that the aggregated cross-sectional variables that appear to be statistically significant in predicting market returns when examined in isolation are no longer significant in a multiple testing framework. I find that we can distill the predictive information in individual firm-characteristics into a single measure of expected stock return using machine learning methods. Aggregating this single variable into a market forecast predicts time-series variation in market returns as far as three years (statistically significant at the 5% level) into the future.

My results also contribute to the stream of literature that studies time-series predictability of returns to characteristic sorted portfolios. Cohen et al. (2003) predict returns to the value portfolio. Cooper et al. (2004) and Daniel and Moskowitz (2016) both study time-series predictability of the returns to the momentum portfolio. Similar to Haddad et al. (2020), my framework allows me to study a much larger cross-section of long-short portfolios while entertaining a large dimensional conditioning information set. Specifically, I contribute to the literature by showing that long-short portfolio forecasts formed from stock return forecasts generated by a neural network model can predict time-series variation in 53 of 56 long-short portfolios (32 of 56 are statistically significant at the 5% level). I also show that imposing economic restrictions on the corresponding machine learning model is essential in producing the forecasts that generalize to the cross-section of long-short characteristic sorted portfolios.

1 Empirical Framework and Data

In this section, I detail the assumptions underlying the empirical exercise in this paper.

1.1 Factor Model

I assume that stock returns are conditionally priced by a linear combination of J factors, $F_{t+1} = [f_{1,t+1}, f_{2,t+1}, ..., f_{J,t+1}]$.

Assumption 1. *A conditional factor model holds such that:*

$$r_{i,t+1} = \beta'_{i,t} F_{t+1} + \varepsilon_{i,t+1} \tag{1}$$

where $r_{i,t+1}$ is the stock return of firm i at time $t+1$, $\beta_{i,t}$ is a $J \times 1$ dimensional vector of conditional factor loadings and $\varepsilon_{i,t+1}$ is an independent identically distributed normal random process, $\mathcal{N}(0, \sigma_{i,\varepsilon})$.

My interest in this paper is to learn a set of expected return functions, $\mathbb{E}_{t-h+1}[r_{i,t+1}]$, where $h \in H = \{1, 2, 3, 13, 37, 61, 91, 121\}$, conditional on some information set, I_{t-h+1}. Supposing this is month t, I predict returns for the following month, $t+1$, by conditioning on the information set I_t and generate return forecasts with the function, $\mathbb{E}_t[r_{i,t+1}]$. To predict returns one year from next month, *1+13*, I condition on the information set observed today, I_t, and generate return forecasts with the function, $\mathbb{E}_t[r_{i,t+13}] = \mathbb{E}_{t-12}[r_{i,t+1}]$.

1.2 Economically restricted model

Guided by economic theory, I introduce the following assumptions to pin down the structural nature of the expectation functions.

Assumption 2. *Expected stock returns are linear in conditional betas and conditional price of risks:*

$$\mathbb{E}_{t-h+1}[\beta'_{i,t}]\mathbb{E}_{t-h+1}[F_{t+1}] \approx b^*_h(\cdot)' * f^*_h(\cdot) \tag{2}$$

where $b^*_h(\cdot)$ is a function that approximates the time $t+1$ expected conditional risk exposures of firm i and $f^*_h(\cdot)$ is a function that approximates the time $t+1$ expected conditional price of risk, all conditional on the information set, I_{t-h+1}. The crucial assumption here is that expected returns is the sum of the product of conditional risk loadings (betas) and the corresponding conditional price of risk. This restriction is standard in the literature and follows from assuming that the SDF is linear or approximately linear in a set of unknown parameters.

I can impose this linearity assumption only because I model the conditional price of risk and conditional beta exposures separately. This separation also allows me to treat conditioning information more in line with findings in the literature. Specifically, I treat characteristic realizations as being informative of risk loadings as in Cosemans et al. (2016), Chordia et al. (2017) and Kelly et al. (2019) and treat the conditional price of risk as arising from linear combinations of trade-able portfolios formed from sorts on characteristics similar to factor definitions in Fama and French (1996), Hou et al. (2015)

and Stambaugh and Yuan (2017).

1.2.1 The conditional price of risk function

The conditional price of risk function, $f_h^*(\cdot)$, is initialized with a $(P+2)$-dimensional column vector of portfolio average returns, $\bar{r}_{p,t-h+1}$, when predicting returns for time $t+1$. This vector comprises an expanding window average return of long-short portfolios formed from sorts on the P firm-level characteristics. I concatenate this vector with the expanding window average return of the equally-weighted market and the risk-free assets. I compute all expanding window averages using portfolio returns starting from January 1965 up to time $t-h+1$. I define the conditional price of risk function as:

$$\mathbb{E}_{t-h+1}[F_{t+1}]' = \bar{r}_{p,t-h+1} W_{0,h} + b_{0,h} \tag{3}$$

where $W_{0,h} \in \mathbb{R}^{58\times 3}$ and $b_{0,h} \in \mathbb{R}^{1\times 3}$ are unknown parameters to be estimated[3]. This parameterization allows for the pricing function to be dense in the space of portfolio and security returns (58 average returns) and simultaneously remain sparse in pricing factor (3 latent factors).

From Kozak et al. (2020), we know that a handful of latent factors are enough to explain a significant fraction of the variations observed in realized returns. Guided by this finding, I set the number of pricing factors to 3[4]. It is worth mentioning that the small number of factors I impose does not restrict the resulting approximator to the same space as a three principal component (PC) model. This is because factor loadings in Equation (2) are time-varying as opposed to the statistic loadings in a PC model. Similar to Kelly et al. (2019) and Gu et al. (2020a), I find that restricting the model to one or two latent factors is too restrictive.

I do not allow for non-linear interactions between portfolio returns in determining the factor returns because I require the factor returns to be spanned by the returns of the underlying 58 portfolios. I construct each long-short characteristic sorted portfolio by

[3]For each forecasting horizon in H, we estimate a different expectation function denoted by the subscript h.

[4]Picking J between 3 and 10 does not qualitatively change the results but increases the time it takes the models to converge

fixing portfolio weights as the rank-normalized characteristic realizations at some time t[5]. I then go long one dollar and short another dollar. All the long-short portfolios I consider are therefore spanned by the stocks in the cross-section.

1.2.2 The expected conditional beta function

The expected conditional beta exposure function, $b_h^*(\cdot)$, is initialized with a P-dimensional vector of rank-normalized firm characteristics, $p_{i,t}$, when predicting returns for time $t+h$. I assume that characteristic realizations at time t are informative of their time $t+h$ realizations.[6]. I approximate the beta exposures as:

$$Y_{1,h} = \psi(p_{i,t-h+1}W_{0,h} + b_{0,h}) \tag{4}$$

$$Y_{2,h} = \psi(Y_{1,h}W_{1,h} + b_{1,h}) \tag{5}$$

$$\mathbb{E}_{t-h+1}[\beta_{i,t}]' = Y_{2,h}W_{2,h} + b_{2,h} \tag{6}$$

where $W_{0,h} \in \mathbb{R}^{56\times1024}$, $W_{1,h} \in \mathbb{R}^{1024\times1024}$, $W_{2,h} \in \mathbb{R}^{1024\times3}$, $b_{0,h} \in \mathbb{R}^{1\times1024}$, $b_{1,h} \in \mathbb{R}^{1\times1024}$ and $b_{2,h} \in \mathbb{R}^{1\times3}$ are unknown parameters to be estimated. ψ is the relu non-linearity; $\psi(\cdot) = max(y, 0)$. This parameterization of the beta exposure function allows me to project the 56 firm-characteristics into a higher dimensional (1024-dimensional) feature space where new features are easier to learn and project the resulting feature set back to the 3-dimensional latent pricing factor space (see Recanatesi et al. (2019)). By allowing the nodes in the first layer of the model to be greater than the size of the input vector, I also maintain the universal approximation property of the deep neural network model (see Johnson (2019)).

Even though I initialize all conditional beta exposure functions with the same characteristic vector, the resulting $J \times 1$ vector of conditional betas can differ across horizons. To see this, consider the relation between the momentum characteristic and expected returns. Momentum is positively related to realized returns for time period $t+1$ but negatively related to realized returns for time period $t+13$ (the reversal characteristic). Therefore, the learned relationship between the same characteristic and realized returns

[5] I rank-normalize all firm characteristics in a cross-section at time t to the interval [-1,1]

[6] Given that some characteristics are highly persistent, this is not a controversial claim (see Baba-Yara et al. (2020)) Replacing the time t realizations with rolling window means does not change the results.

at different horizons by the neural network model will be different.

In the asset pricing literature, betas (risk loadings) are mostly specified as unconditional scaling functions that load on factor portfolio returns. Although this parameterization restricts the resulting model, it is still preferred to the conditional alternative because it is easier to estimate. Given that I estimate most of the unknown parameters of the model using stochastic gradient descent, I do not pay a steep estimation cost by preferring a conditional beta model to an unconditional model.

Additionally, by allowing for beta to be time-varying, the resulting predictive model is much more general in that beta responds to evolving firm characteristics. Consider a growth firm in the initial part of our sample transitioning to a value firm by the end of the sample. By allowing firm characteristics to inform conditional betas, the firm's risk loading (beta) on a particular factor can similarly transition from a low value to a high value across these two distinct regimes. Compare this to the unconditional beta model, which would have to be a scaler that captures the average risk loading of both the growth and value phases of the firm.

Besides the beta conditionality, I also allow for nonlinear interactions between firm characteristics via the ψ non-linearities. This specification is motivated by recent findings in the literature that shows that non-linearities between firm characteristics matter in explaining variations in firm returns. Bryzgalova et al. (2019) find that allowing for non-linearities through conditional sorting improves the resulting mean-variance frontier in the space of characteristic sorted portfolios. Gu et al. (2020a) find that allowing for non-linearities results in an autoencoder asset pricing model that prices 87 out of 95 factor portfolios they consider.

1.3 A simple neural network model

I consider a simpler forecasting model that approximates the product of expected conditional price of risk and expected risk loadings with minimal assumptions coming from economic theory. Specifically, I estimate:

$$\mathbb{E}_{t-h+1}[r_{i,t+1}] \approx g_h^*(z_{i,t-h+1}) \tag{7}$$

where $g_h^*(\cdot)$ is some real-valued deterministic function of $P + M$ real variables, $z_{i,t-h+1}$. $z_{i,t-h+1} = [p_{i,t-h+1} : q_{t-h+1}]$, where $p_{i,t-h+1}$ is firm specific and q_{t-h+1} is the same across firms. I specify $p_{i,t-h+1}$ as a 56-vector of firm level characteristics, the same as in the expected conditional beta exposures function, and concatenate it with an M-dimensional, q_{t-h+1}, aggregate variables as in Gu et al. (2020b).

The difference between this forecasting model and the one I propose is that it does not model the conditional beta exposures and conditional price of risk functions separately. It approximates the expected return function directly while skipping all intermediary restrictions. This is the best performing machine learning model in Gu et al. (2020b) and so serves as a natural benchmark for the more restricted model I propose. It is simpler in that it makes very little structural assumptions about how the different constituents of the information set interact in informing return expectations.

Following Gu et al. (2020b), I approximate Equation (7) using a three-layer feedforward neural network, which is defined as[7]:

$$Y_{1,h} = \psi(z_{i,t-h+1}W_{0,h} + b_{0,h}) \tag{8}$$

$$Y_{2,h} = \psi(Y_{1,h}W_{1,h} + b_{1,h}) \tag{9}$$

$$Y_{3,h} = \psi(Y_{2,h}W_{2,h} + b_{2,h}) \tag{10}$$

$$\mathbb{E}_{t-h+1}[r_{i,t+1}] = Y_{3,h}W_{3,h} + b_{3,h} \tag{11}$$

where $W_{0,h} \in \mathbb{R}^{64\times32}$, $W_{1,h} \in \mathbb{R}^{32\times16}$, $W_{2,h} \in \mathbb{R}^{16\times8}$, $W_{3,h} \in \mathbb{R}^{8\times1}$, $b_{0,h} \in \mathbb{R}^{1\times32}$, $b_{1,h} \in \mathbb{R}^{1\times16}$, $b_{2,h} \in \mathbb{R}^{1\times8}$ and $b_{3,h} \in \mathbb{R}$ are unknown parameters, θ, to be estimated. ψ is a non-linear function (relu) applied element-wise after linearly transforming an input vector, either $z_{i,t-h+1}$ or $Y_{k,h}$.

Despite its flexibility, this simple forecasting model imposes some important restrictions on the estimation problem. The function, $g_h^*(\cdot)$, depends neither on i nor t but only h. By maintaining the same functional form over time and across firms for some time-period h, the model leverages information from the entire firm-month panel. This restriction significantly reduces the number of parameters I need to estimate and increases

[7]Feedforward networks are the main building blocks of much more complicated neural networks. Among the five feedforward neural network models that Gu et al. (2020b) study, the three-layer deep neural network out-performs along several dimensions.

the resulting estimates' stability. This restriction is loose in that I re-estimate $g_h^*(\cdot)$ every two years, which means that each subsequent 24 month set of stock forecasts for some particular horizon h comes from a slightly different approximation of $g_h^*(\cdot)$. Finally, the specification also assumes that the same information set is I_t is relevant for making predictions for all horizons in H.

1.4 Loss Function

I estimate Equation (2) and Equation (7) by minimizing the mean squared error loss function with an l_1 penalty:

$$\mathcal{L}(\theta) = (N_t T)^{-1} \sum_{i=1}^{N_t} \sum_{t=1}^{T} \left(R_{t+1} - \hat{R}_{t+1} \right)^2 + \lambda_1 \left|\left|\theta\right|\right|_1 \tag{12}$$

where R_{t+1} is a vector of stock returns for time t, $\tilde{R}_{t+1}$ is a vector of predicted returns for all N_t firms in the cross section at time t, θ is the vector of model parameters. I minimize the empirical loss function over a pool of firm-month observations. I choose hyper-parameters such as λ_1 via a validation set. All hyper-parameters are detailed in in Appendix C.

1.5 Estimation

I use the AdaBound learning algorithm from Luo et al. (2019) to estimate the unknown parameters (θ)[8].

In addition to the l_1 penalty, I use batch normalization to help prevent internal covariate shifts across layers during training, (see Ioffe and Szegedy (2015)). I train the model on a batch size of randomly sampled 10000 firm-month observations per iteration. I estimate the model over 100 epochs, where an epoch represents a complete cycle through all of the training data. I stop training before the 100th epoch if the validation set does not increase after five subsequent epochs. Further details of the learning algorithm are provided in Appendix C.

[8]AdaBound leverages the rapid training process of the more popular adaptive optimizers such as Adam, (Kingma and Ba, 2014), and generalizes like the classic stochastic gradient descent optimizer. Also, AdaBound has theoretical convergence guarantees which other optimizers such as ADAM lack.

1.5.1 Sample Splitting

The dataset starts from January 1965 and ends in December 2018. I employ a rolling window estimation scheme by splitting the dataset into three parts; training, validation, and testing.

[Insert Figure 1 about here]

In predicting returns for month $t+1$ using information available up to time t, I estimate the model using 15 years of data starting from January 1975 and ending in December 1989. I choose hyper-parameters by comparing estimated model performance over a validation dataset starting from January 1990 to December 1994. I use the optimal model to make one-month ahead return predictions from January 1995 to December 1996. Figure 1 illustrates this exercise. I move the training, validation, and testing set forward by two years and repeat the process.

In predicting returns for month $t+2$ using information available up to time t, I estimate the model using 15 years of data starting from December 1974 and ending s in November 1989. I choose optimal hyper-parameters by comparing estimated model performance over a validation dataset from December 1989 to November 1994. I use the optimal model to make two-month ahead predictions starting from December 1994 to November 1996. This ensures that when comparing model performance across horizons, I am always comparing returns realized between January 1995 to December 1996, thereby aligning return realization dates across prediction periods, H. Similar to $t+1$, I move the training, validation, and test set forward by two years and repeat the process.

I always predict returns for the out-of-sample period; January 1995 to December 2018. As discussed above, I do this by shifting the conditioning information further into the past. This allows me to maintain the same training, validation and testing data size (in months) across horizons. Although this allows me to compare forecasts from different horizons for the same out-of-sample period, the subset of firms I am comparing across horizons is different. This is because firms enter and exit the CRSP file over time. Consider two different horizon forecasts for the month January 1995. The one month ahead forecast will condition on firms alive in December 1994. Whereas, the five year-ahead monthly forecast will condition on firms alive in December 1989. The trade-off I

make is to align my setup more with a real-time setting, where agents form expectations for all future horizons in H, conditional on what they observe at the time.

I choose to estimate monthly forecasts because this allows us to bring the standard financial econometric tools to the problem and side step the econometric issues inherent in using compounded returns.

1.6 Data

I obtain monthly market data for US common stocks traded on AMEX, NASDAQ, and NYSE stock exchanges from CRSP. I match market data with annual and quarterly fundamental data from COMPUSTAT. I build a set of 56 firm-level characteristics from this panel.[9] The characteristic definitions are from Freyberger et al. (2020) and Green et al. (2017). I obtain the one-month risk-free rate from Kenneth French's website. To avoid forward-looking bias, I follow the standard practice in the literature and delay monthly, quarterly and annual characteristics, by a month, four months, and six months respectively (similar to Green et al., 2017; Gu et al., 2020b). To be included in the sample for some month t, a firm must have at least 30 non-missing characteristic observations. I rank-normalize the characteristics to the interval [-1,1] and replace missing values with zero.

The aggregate variable set, q_t, I use is from Goyal and Welch (2008), namely the S&P 500 dividend-to-price ratio, the S&P 12-month earnings-to-price ratio, the S&P 500 book-to-market ratio, net equity expansion, stock variance, the term spread, the default spread, and the treasury-bill rate[10]. I condition on this set of aggregate variables to keep the simple model in-line with the specification in Gu et al. (2020b). Conditioning the simpler model on the same aggregate variables in as in Equation (3) leads to qualitatively poorer results.

[9] The details of the characteristics are provided in Table B.1.

[10] I would like to thank Amit Goyal for making this series available on his website.

2 Neural network forecasts in the cross-section of stocks

This section examines how incorporating economic theory in designing a neural network forecasting model helps improve return forecasts. I do this by comparing the forecasting accuracy of the economically restricted neural network model to that of the simple model in the cross-section of stocks across horizons. Additionally, I decompose the forecasts of both models to shed light on the cross-sectional and time-series prediction properties of the models.

The standard statistic I use to assess the predictive performance of these forecasts is the out-of-sample R Squared (R^2_{OOS}), which is defined as:

$$R^2_{OOS} = 1 - \frac{\sum_{(t) \in oss}(R_t - \tilde{R}_{t,1})^2}{\sum_{(t) \in oss}(R_t - \tilde{R}_{t,2})^2} \tag{13}$$

where R_t is the time t vector of realized stock returns, $\tilde{R}_{t,1}$ is a vector of forecasts from model 1 and $\tilde{R}_{t,2}$ is a vector of forecasts from model 2. Intuitively, the statistic compares the forecasting error of model 1 ($(R_t - \tilde{R}_{t,1})^2$), to that of model 2 ($(R_t - \tilde{R}_{t,2})^2$). If the forecasting error of model 1 is smaller than that of model 2, then R^2_{OOS} will be positive. A positive R^2_{OOS} therefore means that forecasts from model 1 improve upon forecasts from model 2.

I formally test the null hypothesis that forecasts from model 1 are no different from forecasts from model 2 in explaining variations in stock returns using the Clark-West (2007) test with Newey-West (1987) adjusted standard errors.

2.1 Can neural networks predict stock returns across horizons?

To answer this question, I define forecasts for model 1 as forecasts from the neural network models. I compare each models forecast to a zero prediction benchmark; $\tilde{R}_{t,2} = 0$. The results from this exercise answer the question, "How much variation in realized returns are explained by the neural network forecasts?"

[Insert Table 1 about here]

Panel A of Table 1 reports results for both the economically restricted model and the simple model. All the R^2_{OOS} estimates are positive and statistically significant across horizons. In general, both models' ability to explain variations in stock returns monotonically decrease the further into the future the forecasts pertain. Whereas the economically restricted model can explain about 0.99% of the variation in next month's return, it can only explain about 0.13% of the variation in ten-year returns. Similarly, the simple model can explain about 0.58% of the variation in next month's return, and this falls to 0.18% of the variations in return ten years in the future.

Comparing the models on the variations in returns they explain in next month's return, the economically restricted model explains close to twice the variation explained by the simple model; 0.99% against 0.58%. In explaining variations in stock returns further in the future, the simple model explains a slightly larger fraction; 0.18% against 0.13%.

2.2 Disentangling the composite R^2_{OOS}

The R^2_{OOS} tells us how much variation in returns the forecasts from model 1 explain when the benchmark model (model 2) is a zero prediction model. The results show that both models can predict stock returns across horizons. However, the R^2_{OOS}, as defined above, fails to tell us along which dimension of stock returns these estimates forecast well. The forecast may be predicting stock returns well because they predict the level factor in stocks. Or they could additionally be predicting time-series variation in the cross-sectional dispersion in stock returns. Given that a strong factor structure holds in the pool of stocks, it is instructive that we disentangle the R^2_{OOS} to shed light on this.

I assume a two-factor structure holds for the stock return forecasts. I fix the first factor as the equally-weighted market forecast and allow the second factor to subsume all other priced factors in the cross-section[11]. This parameterization allows me to decompose return forecasts from some model m for a firm i at some time t into two parts:

$$r_{m,i,t} = \left(N_t^{-1}\right) \sum_{k \in t} r_{1,k,t} + r^{RR}_{m,i,t} \tag{14}$$

[11] Kozak et al. (2020) show that an asset pricing model of a similar form explains a significant fraction of the variations in returns

where $(N_t^{-1}) \sum_{k\epsilon t} r_{1,k,t}$ captures the cross-sectional mean forecast of model m and $r^{RR}_{m,i,t}$ captures the cross-sectional variation in forecasts across firms. The return to each firm i ($r_{m,i,t}$) is therefore made up of the cross-sectional level factor ($(N_t^{-1}) \sum_{k\epsilon t} r_{1,k,t}$) and firm specific relative return ($r^{RR}_{m,i,t}$).

I further decompose the relative forecast ($r^{RR}_{m,i,t}$) into an unconditional component ($\mu^{RR}_{1,i}$) and a conditional component ($\tilde{r}^{RR}_{1,i,t}$). Specifically, I decompose $r^{RR}_{m,i,t}$ as follows:

$$r^{RR}_{m,i,t} = \mu^{RR}_{m,i} + \tilde{r}^{RR}_{m,i,t} \tag{15}$$

where $\tilde{r}^{RR}_{1,i,t}$ is mean zero (by construction) and captures the relative (residual) time-series forecasts of model m. $\mu^{RR}_{1,i}$ is the average firm i forecast over the out-of-sample period and captures the unconditional relative (residual) stock forecast. This parameterization allows me to study the time-series predictability of relative stock returns absent the unconditional component. See section D in the Appendix for more details on the decomposition.

Panel B of Table 1 reports the results for the decomposition of the R^2_{OOS} against a zero prediction benchmark. For both models, the ability of their forecasts to explain time-series variation in relative stock return is only present for short-run months. Neither model can explain time-series variation in relative stock returns realized beyond one year in the future.

However, the amount of time-series variation in relative stock returns the models can explain is very different. Whereas the simple model explains about 0.20% of time-series variation in next month's relative stock return, the economically restricted model explains about 0.71%, a more than three-fold improvement. In predicting monthly relative stock returns one year in the future, the simple model explains about 0.03% of time-series variation in relative stock returns against 0.08% for the economically restricted model.

For both models, a large fraction of the reported composite R^2_{OOS} comes from explaining variations in the level factor in stock returns. For the economically restricted model, about 40% of the composite R^2_{OOS} (0.43% out of 0.99%) comes from explaining variations in next month's level factor. For the simple model, this figure is 70% (0.41 out of 0.58%). For all other future forecasting periods, more than 90% of the composite R^2_{OOS}

comes from the models' ability to explain variations in level factor, with little to negative (R^2_{OOS}) contributions coming from explaining variations in relative stock returns.

The results show that intermediate and long-run forecasts from the neural network models are very different from short-run forecasts. Whereas short-run predictions can discriminate between high and low expected return stocks (relative stock returns) in addition to forecasting the level factor, intermediate and longer-run forecasts only explain variations in the cross-sectional average return (level factor).

2.3 An alternative benchmark

Results from the decomposition of the R^2_{OOS} with respect to the zero prediction benchmark show that the dominant factor that the forecasts are predicting is the equally-weighted market return. This result suggests that an alternative benchmark that does reasonably well along this particular dimension of returns should be tougher for the neural network forecasts to beat. From Goyal and Welch (2008), we know that one such example is the historical average market return. I define this benchmark's $t+h$ stock return forecast as the time t average equally-weighted market return computed using data from 1926.[12]

[Insert Table 2 about here]

The results are reported in Table 2. For short-run months, I find a more than 30% reduction in the composite R^2_{OOS} compared to the zero-prediction model. From this result, we can conclude that the historical average market return is a challenging benchmark, even in the pool of individual stocks. In the long-run, I find an increase in the composite R^2_{OOS} compared to the zero-prediction model. This result means that the zero-prediction model remains the tougher benchmark for longer run returns. This finding is explained by the fact that more than 40% of firms alive at any period t fall out of the sample by $t+60$. Thus, the historical average market return computed as a function of firms alive at some t, will be a poor estimate of the longer-run unconditional average return.

Comparing the R^2_{OOS} estimates of the simple model to the economically restricted model across horizons and benchmarks, it is evident that economic restrictions generally

[12]Results for other alternative models are in the Internet Appendix **??** of the Internet Appendix.

improve the forecasts. In predicting next month's return, the economically restricted model has an R^2_{OOS} of about 0.64%, whereas the simple model has an R^2_{OOS} of about 0.20%. In predicting returns ten-years into the future, the economically restricted model has an R^2_{OOS} of 0.62%, and the simple model has an R^2_{OOS} of about 0.55%.

Taken together, the results in this section show that incorporating economic restrictions improves the ability of a neural network model to predict stock returns. This improvement is most evident in the ability of the forecasts to explain time-series variations in relative stock returns over the short-run.

3 Predicting market and long-short portfolio returns

The previous section shows that incorporating economic theory in designing a neural network architecture improves return forecasts in the cross-section of stocks. Since individual stock forecasts can easily be aggregated to forecast market returns and returns to long-short characteristic sorted portfolios, it is natural to ask if the model that incorporates economic theory generalizes better along these dimensions than the simple model. That is the central question I answer in this section.

3.1 Can the forecasts predict market returns?

To answer this question, I define the market forecast as the value-weighted monthly stock forecast for period $t+h$ and define the market return as the value-weighted monthly cross-sectional average stock return of firms in the CRSP file at time $t+h$.To capture the pure effect of different forecasts, I always use market-caps from time t but allow the forecasts to vary across horizons. I compute the R^2_{OOS} with respect to two benchmarks; a zero-prediction model and the historical average-market return.

[Insert Table 3 about here]

The R^2_{OOS} of the neural network against a zero prediction benchmark tells us how much variation in market returns the forecasts explain. The results in Table 3 show that both models can robustly explain market returns across all horizons, I consider. The economically restricted model can explain a larger fraction of the variation in market

returns compared to the simple model, especially for short-run to intermediate horizons (up to three years). For example, in predicting next returns, the economically restricted model explains about 5.35% of the variation in market returns while the simple model explains about 2.05%.

Decomposing the R^2_{OOS} into a time-series variation and an unconditional return component shows that less than 35% of the variation explained in market returns across horizons pertains to the ability of both models to explain time-series variations in returns. For instance, in predicting market returns one year into the future, 1.51% of the 4.90% composite R^2_{OOS} comes from the ability of the economically restricted model forecasts' to explain time-series variation in market returns. The rest comes from matching the unconditional market return in the out-of-sample period.

Focusing on the more challenging historical average market return benchmark, we see that the simple model's market return forecasts offer no improvements. For all horizons and dimensions of market returns, this model fails to improve upon the historical average market forecast. The story is different for the economically restricted model. This model fails to improve upon the historical average market forecast in predicting the unconditional market return in the out-of-sample period. However, its ability to out-perform the historical average market return forecast in predicting time-series variation in market returns is large and statistically significant at the 5% level up to three years in the future.

3.2 Can forecasts predict long-short portfolios returns?

The positive and statistically significant R^2_{OOS} in rows 1 and 3 in panel B of Table 1 suggest that both neural network forecasts should be able to forecast returns to long-short portfolios. This is because this dimension of the decomposed R^2_{OOS} is related to predicting time-series variation in relative stock returns. And this translates into returns of long-short portfolios. However, we can not make conclusive statements from the results in Table 1 because the R^2_{OOS} are computed with respect to the entire cross-section of stocks, whereas long-short portfolios only buy and sell a fraction of stocks that are most of the time in the tails of the return distribution. Additionally, long-short portfolios are mostly value-weighted as and not equally-weighted as in Table 1.

To answer the question, I sort stocks on the five characteristics in the Fama and French (2018) factor model; book-to-market, investment, size, operating profit, and momentum[13]. For characteristics computed from balance sheet or income statement variables, I update them at the end of June of year s using the characteristic observations from the fiscal year-end $s-1$. For characteristics computed only from CRSP variables, I update them at the end of each month and re-balance accordingly. I form decile portfolios from the sorts and value-weight to reduce the effect of small stocks. The return (forecast) to the long-short portfolio is the value-weighted return (forecast) of portfolio ten minus the value-weighted return (forecast) to portfolio one.

Similar to analyzing market return predictability, I decompose the (R^2_{OOS}) to investigate time-series and unconditional forecasting accuracy of the long-short characteristic portfolio forecasts.

[Insert Table 4 about here]

Results for the simple neural network model are reported in Table 5. The alternative model is the zero-prediction model. Even against this much weaker benchmark, the simple model fails to robustly explain any variation in returns to long-short characteristic sorted portfolios. For almost all reported horizons and across all five long-short portfolios, the R^2_{OOS} is negative. For the few horizons and portfolios where the estimate is positive, it is seldom statistically significant.

[Insert Table 5 about here]

Table 5 reports results for the economically restricted model. For this model, the benchmark is the historical average long-short portfolio return computed using data from 1964. The model does a much better job predicting returns to long-short portfolios than the simple model, despite the more challenging benchmark. Focusing on forecasts for time $t+1$, I find positive R^2_{OOS} for all five long-short portfolios as against two for the simple model. For four of these five portfolios, the R^2_{OOS} is statistically significant at the 5 % level. The decomposed R^2_{OOS} shows that the forecasting power of the economically restricted model is driven by its ability to better predict time-series variation in returns to these long-short portfolios.

[13]To be included in a sort, a firm must have a neural network forecast and non-missing observations for return and characteristic being sorted on.

To show how pervasive this finding is, I expand the universe of long-short portfolios to all 56 characteristics that I condition on in the beta function (Equation (4)) and focus on forecasts for month $t+1$. Figure 2 reports the results.

[Insert Figure 2 about here]

In panel A, I find that 53 (32) of the 56 long-short portfolios have positive (and statistically significant) composite R^2_{OOS}. Similar to the results above, most of the composite R^2_{OOS} is driven by the forecasts' ability to predict time-series variation in returns to long-short portfolios. In panel B, I find that for 53 of the 56 long-short portfolios the neural network forecasts improve upon the the benchmarks ability to predict time-series variation in returns. For 31 portfolios, this improvement is statistically significant at the 5%.

[Insert Figure 3 about here]

Moving beyond month $t+1$ forecasts, I report results for other horizons in Figure 3. To keep things compact, I only report the fraction of long-short portfolios with positive composite R^2_{OOS}, positive contributions coming from forecasting time-series variation in returns, and positive contributions coming from predicting the unconditional long-short portfolio return. Panel B reports the fractions that are both positive and significant. I observe that although the neural network forecasts predict a majority of long-short portfolio returns in short-run months, the fraction that is significant precipitously drops to zero when I use forecasts older than three months. From this, I conclude that the more timely the information set I condition, the more accurate the forecasts predict time-series variations in returns to long-short portfolios.

The results in this section show that machine learning guided by economic theory can lead to significant improvements in predicting returns that robustly generalize beyond the cross-section of stocks. Specifically, such a model can predict time-series variation in monthly market returns up to three years into the future. Additionally, the model can predict time-series variation in next month returns for 53 (32 are statistically significant) out of 56 long-short portfolios.

4 Neural network forecasts and conditional expected returns

The previous sections show that the economically restricted model explains significant variation in stock returns. This ability generalizes to market returns and long-short portfolio returns. This section analyzes how well the economically restricted model forecasts line up with conditional expected returns across firms, portfolios, and time.

The standard tool in the literature used in this specific analysis is time-series predictive regressions (see among others Cochrane (2008), and Lewellen (2015)). The slope coefficient from regressing demeaned forecasts on returns is informative of how well the forecasts line up with conditional expected returns. We are interested in predictions that get the conditional direction of returns right. If the slope coefficient is positive and statistically different from zero, then it fulfills this requirement. Additionally, we are interested in unbiased return forecasts, that is, models for which the slope coefficient is indistinguishable from one. For such models, a one percent forecast on average translates into a one percent return.

[Insert Table 6 about here]

Panel A of Table 6 report results from regressing demeaned relative stock returns on realized stock returns. The results generally confirm the conclusions from the decomposed out-of-sample R^2 analysis. For the short-run months, $t+1$ up to $t+13$, I can reject the null hypothesis that the forecasts fail to predict time-series variation in relative stock returns. This is because the 95% confidence interval of the slope coefficient is strictly positive. For $t+1$, the forecasts are unbiased because the 95% confidence interval of the slope coefficient includes one. Specifically, a one percent relative stock forecast on average translates into a 0.97 percentage point increase in next month's realized relative stock return. The model over predicts time-series variation in relative stock returns for all other short-run months because the confidence intervals are strictly less than one but positive.

From these results, we can conclude that the model's forecasts line up well with expected stock returns for the next month's returns but over-predict stock returns for all

other months.

Panel B of Table 6 shows that the market forecasts, up to intermediate-term months, on average, do line up with expected market returns. For months $t+1$ up to $t+37$, the slope coefficients from regressing demeaned market forecasts on market returns are positive and statistically different from zero. The estimates are around 1.50, meaning a one percent market return forecast translates into a 1.50 percentage point increase in market return. And the 95% confidence interval of the slope coefficient includes one for these specific monthly forecasts.

Panel C of Table 6 reports results for long-short portfolios. Slope coefficients for months $t+1$, $t+3$, $t+13$ are positive and statistically different from zero. This means the aggregate neural network forecasts from the economically restricted model can predict time-series variation in returns to long-short portfolios for short-run months. These forecasts for long-short portfolios do not generally line up well with conditional expected returns. A one percent forecast on average translates into about a 2 percentage point realized return to the typical long-short Fama and French (2018) 5 model characteristic sorted portfolio. We cannot reject the null hypothesis that the slope coefficients for $t+1$ and $t+3$ are unbiased. However, we can reject this null for $t+13$. Forecasts for this month on average under-predict conditional expected returns to long-short portfolios.

5 Optimal Portfolios

This section introduces several optimal trading strategies that highlight the practical usefulness of the neural network forecasts. We show that an investor using these forecasts in a pseudo-real-time setting over the out-of-sample period enjoys significant improvements measured by average returns, Sharpe ratios, risk-adjusted returns, and certainty equivalents.

I define the certainty equivalent with respect to an investor with a mean-variance utility function and a risk aversion parameter of 2. Specifically, I compute the certainty equivalent return of a strategy as:

$$CE = \bar{r}_h^p - \frac{\gamma}{2}\sigma_{p,h}^2 \tag{16}$$

where $\sigma_{p,h}$ is the sample standard deviation of the strategy. The certainty equivalent can be interpreted as the risk-free return that a mean-variance investor with a risk-aversion coefficient of γ would consider equivalent to employing this strategy. Alternatively, it can be viewed as a fee that an investor is willing to pay to use the information inherent in our forecast. I report the certainty equivalent annualized and in percentages.

5.1 Optimal timing strategies

I consider a strategy that times a risky security by levering up and down the position in the security based on whether conditional expected returns are high or low. The previous section showed that the forecasts from the economically restricted neural network model explain time-series variation in returns for most of the portfolios we consider. Therefore, we should expect these forecasts to be informative of when to lever up and down based on return expectations for the future.

For each month, t, I use the conditional expected return forecast from the economically restricted neural network model to calculate the Markowitz optimal weight to be invested in the risky asset as:

$$w_{t,h} = \frac{\tilde{r}_{t,h} - r^f_{t+1}}{\gamma \sigma(\tilde{r}_{1:t-1,h})} \tag{17}$$

where γ is the risk aversion coefficient, which I set to 2. I fix the conditional standard deviation estimate ($\sigma(\tilde{r}_{1:t-1,h})$) at an annualized value of 15 % across securities because of two main reasons; 1) to remove the impact of volatility timing from the exercise (see **?**) and 2) because the forecasting model does not produce a conditional standard deviation estimate. At the end of each month, I compute the timing portfolio return as:

$$r^p_{t,h} = w_{t,h} r_{t+1,h} - (1 - w_{t,h}) r^f_{t+1} \tag{18}$$

and iterate until the end of the out-of-sample period, December 2018.

5.1.1 The optimal market timing portfolio

The first trading strategy I consider tries to time the value-weighted market return by deciding how much to invest between the market and a risk-free asset using the aggregated forecast for the market. I restrict the market to the 500 largest market capitalized firms at each time t. For each month t, the strategy invests $w_{t,h}$ in the value-weighted market and $1 - w_{t,h}$ in the risk-free asset.

[Insert Table ?? about here]

Panel A of Table 7 reports the results. A buy and hold strategy that is fully invested in the market over the sample period makes a annualized average return of 10.39 % with a certainty equivalent of 8.18. The return to this strategy is fully explained by the CAPM and the Fama and French (2018) 5 factor model. A timing strategy that uses the most recent market forecasts, $t + 1$, earns an annualized average return of 17.21 % with a certainty equivalent of 12.86. Timing the market with predictions that are a month old, $t+2$ to two years old, $t+25$ all out-perform the buy and hold strategy. Generally, the more timely the forecasts are, the higher their accuracy in predicting time-series variation in market returns. We can see this from the higher certainty equivalents and average returns for periods $t + 1$ and lower estimates for much older forecasts such as $t + 121$.

5.1.2 The optimal characteristic timing portfolio

The second timing strategy I consider tries to time an equally-weighted portfolio of book-to-market, size, investment, profitability, and momentum long-short portfolios. For each month t, the strategy invests $w_{t,h}$ in this equally-weighted portfolio and $1 - w_{t,h}$ in the risk-free asset[14].

Panel B of Table 7 reports the result for the timing strategy. Similar to previous results, I find that the spread in returns generated by the timing strategy is strongest when the forecast is much closer to the re-balancing month t. Forecasts that are older than two years to the date of re-balancing generate negative spreads. Characteristic timing, like all other strategies considered in this section, requires timely information. The most

[14]Although the portfolio we are timing is equally-weighted, the individual long-short characteristic sorted portfolios are all value-weighted.

timely forecast $t+1$ almost triples the buy and hold certainty equivalent, average returns, and alphas.

5.1.3 Timing individual characteristic sorted portfolios

The third timing strategy I consider tries to time individual value-weighted portfolios formed from sorts on the characteristics I condition on. The time t weight $w_{t,h}$ invested in each portfolio is determined as in the other timing strategies. I only report results for the $t+1$ forecasts for the sake of space..

[Insert Figure 4 about here]

Panel A of Figure 4 reports the annualized average return of the difference in returns between the timing portfolio and a buy and hold variant. For 54 of the 56 long-short portfolios, timing leads to an improvement in average returns. For 23 of the 56, this improvement is statistically significant at the 5% level. 37 of the 56 portfolios show a greater than 5% improvement in annualized returns. Panel B reports the Sharpe ratio of the difference in returns between the timing portfolio and the buy and hold. For 46 of the 56 long-short portfolios, the increase in the annualized Sharpe ratio is at least 0.20 units. Finally, in Panel C, I report the alpha from regressing the timing portfolio returns on the Fama and French (2018) 6 factor model. For 34 of the 56 portfolios, the timing portfolio returns are not fully explained by this asset pricing model.

The results from this timing exercise show that the neural network forecasts predict well time-series variation in relative stock returns for short-run months. And this predictive ability strongly generalizes to predicting time-series variation in long-short portfolios. Remember that long-short portfolios zero out the level factor in stocks by construction. So, forecasts that predict stock returns well only because they predict the level factor cannot predict return variations in long-short portfolios.

5.2 Optimal rotation strategies

The alternate set of strategies that I consider rotates across securities in the cross-section of stocks and long-short characteristic sorted portfolios. Strategies of this type allow us

to gauge how accurately the relative stock return forecasts' discriminate between high expected return stocks and low expected return stocks in the cross-section.

For each period t, I sort all securities in a particular cross-section on their forecasted return, buy (sell) the top (bottom) 10%. I repeat these sorts H times for each forecasting period. If the timeliness of the information set is important for accurately discriminating between high and low expected return securities in a cross-section, then the strategies that use forecasts for short-run periods (h $\in \{1, 2, 3, 13, \}$) should out-perform strategies that use much longer-run forecasts.

[Insert Table 8 about here]

5.2.1 Long-short stocks strategy

This is a long-short strategy that buys (sells) the value-weighted portfolio of the 10% highest (lowest) expected return stocks within the 500 largest market capitalized firms in the cross-section of stocks at some time t.

The results are presented in panel A of Table 8. Re-balancing the long-short stock portfolio each month using the $t+1$ forecasts produces a certainty equivalent of 11.73, larger than the 8.13 for the buy and hold investor. This re-balanced portfolio also generates returns that are not explained by the CAPM or the Fama and French (2018) 5 factor model. Therefore, the long-short portfolio's improved performance does not come from loading on the fundamental factors in either asset pricing model. Generally, the certainty equivalents, average returns, and alphas fall with the horizon. The right way to think about this is that using older forecasts to re-balance the long-short portfolio comes at a cost. Remember, the portfolios are re-balanced monthly, and the horizon dimension is captured by how old the forecast are. A monthly re-balanced long-short portfolio using forecasts for period $t+120$ means re-balancing with forecasts that are ten years old.

5.2.2 Characteristics rotation strategy

This cross-section is made up of the five characteristics in the Fama and French (2018) 6 factor model. The strategy buys (sells) an equally-weighted portfolio of the two long-short characteristic sorted portfolios with the highest (lowest) expected returns. Individual

long-short portfolios are value-weighted.

The results are presented in panel B of Table 8. The rotation strategy using forecasts for period $t+1$ almost double the certainty equivalent, average returns, and alphas of the benchmark strategy that buys and holds all five portfolios. The gains from using the forecasts fall as I use older forecasts. Whereas forecasts for period $t+1$ generate an annualized average return of 17.06 %, forecasts for period $t+120$ generate an annualized average return of -3.22 %. Expanding this cross-section to the 56 long-short portfolios does not change the conclusions.

6 Predictability decay over time

Ben-Rephael et al. (2015) show that the characteristic liquidity premium has significantly declined over time. Chordia et al. (2013) study the predictive accuracy of a broader set of characteristics in two sub-periods and find significant attenuation in the predictive accuracy of the characteristics in the second sub-period. The decreasing predictive accuracy of characteristics is not much challenged in the literature. However, the hypothesis put forward to explain the phenomena are many and varied. Chordia et al. (2011) argue this result may be the effect of institutional activity, bringing about more efficient price formation. McLean and Pontiff (2016) argue that the result may be due to popularization coming from academic research.

Given that I condition the forecasting model on an overlap of these characteristic realizations, it is essential to answer the question, "How has return predictability evolved over the sample period?" If the information set has become less informative about expected returns, we should expect to see decreasing out-of-sample R^2 over time. To answer this question, I compute a two-year rolling out-of-sample R^2 from January 1997 to December 2018. I present the results for forecasts generated for month $t+1$.

[Insert Figure 5 about here]

Panel A of Figure 5 reports the rolling out-of-sample R^2 with respect to predicting time-series variation in relative stock returns. The figure shows that a large fraction of the forecasts' ability to explain time-series variation in relative stock returns comes from the

initial part of the sample. The average out-of-sample R^2 in the first half of the sample is about twice as large in the second half. From this, I can conclude that the predictive ability of the information set has waned over time. However, I cannot conclude that the forecasts have entirely lost their ability to predict time-series variation in relative stock returns even though the estimates are negative at the end of the sample. This is because a similar negative streak is present between 2007 and 2009, after which positive estimates remerged.

The results for the market show that the rolling estimates are much more volatile. This may be because the time-series variation in market returns is tougher to predict or because the sample from which I compute the rolling estimate is smaller. Similar to the case of individual stocks, the forecasts' ability to predict time-series variation in returns is much stronger in the first half of the sample than the second. At the end of the sample, the rolling estimates are negative.

7 Variable importance across horizons

In this section, I investigate which covariates matter the most in generating return forecasts across horizons. I use the notion of Shapley values from Lundberg and Lee (2017). The authors show that Shapley values generalize many competing measures of model explainability with respect to neural networks.

Shapley regression values are measures of covariate importance for linear models that are robust to multicollinearity (Lundberg and Lee (2017)). It is a feature importance measure that requires re-estimation of a model on all possible covariate subsets $S \subseteq F$, where F is the set of all covariates. It assigns a value to each covariate representing the marginal effect on the model prediction of including that feature.

To compute this marginal effect for some model f, consider the model $f_{S\cup\{k\}}$ trained with all covariates present and model f_S with feature k withheld. The predictions from the two models are then compared for some observation x_S, $f_{S\cup\{k\}}(x_{S\cup\{k\}}) - f_{S\cup\{k\}}(x_S)$, to compute the marginal effect for that observation. This effect is computed for all possible subsets $S \subseteq F\backslash k$. Shapley values are then computed as the weighted average of all possible

differences:

$$\phi_i = \sum_{S \subseteq F \setminus k} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \Big[f_{S \cup \{k\}}(x_{S \cup \{k\}}) - f_{S \cup \{k\}}(x_S) \Big] \tag{19}$$

Shapley additive explanation (SHAP) values represent an easily computable approximation of Eq. (19). These values provide the unique covariate importance measure I use[15].

[Insert Figure 6 about here]

Figure 6 reports the overall ranking of characteristics for select horizons. I estimate SHAP values for each observation in our sample and then average across characteristics. Characteristics are ordered so that the highest-ranked is at the bottom and lowest-ranked characteristics (out of the top four) at the top. Blue represents a negative contribution to the forecast and red a positive contribution. In analyzing the results in this section, it is important to remember the previous section's results. In the cross-section of stocks, the dominant factor we predict is the level factor.

The results show that the same three variables are the most important drivers of return forecasts across horizons. The most important variable is dividend-yield (DP). A unit increase in this variable positively predicts returns in the short and intermediate run. It is not surprising that the most dominant predictive variable in our conditioning information is the one variable that the literature has shown to predict market returns, the dividend-yield (see, Goyal and Welch (2008),Cochrane (2011) and Ferreira and Santa-Clara (2011)).

The second most important variable is debt-to-price. A unit increase in this variable leads to a reduction in return forecasts across horizons. Intuitively, this results suggests that as the average leverage in the cross-section of stocks increase, expected returns falls.

The third important variable is closeness of last month's price to last 52 week close (CL2HG), a trend related measure, is the third. The fourth measure is depreciation and amortization to total assets.

Although one may be tempted to interpret all these estimates and their impact on return forecasts with respect to only the cross-section of stocks, it is important to remember that there exists a factor structure inherent in this cross-section. However, it is

[15]See Lundberg and Lee (2017) for detail

not clear how to decompose variable importance along similar lines as in Eq. (14). One solution is to exclude a subset of characteristics and then compare the reduction in the decomposed out-of-sample R^2. However, systematically dropping and including characteristics is leads to a combinatorial explosion in the number of tests. And so without strong priors on what clusters of characteristics matter for which dimension of returns, this solution is computationally infeasible.

8 Conclusion

This paper primarily shows that incorporating economic theory in a neural network model architectural design significantly improves the resulting model's forecasting ability. I find that the improvements mainly come from the resulting model's ability to explain time-series variations in returns. The improved model explains almost twice as much of the variation in the following month's returns compared to a much simpler neural network model.

I show that the model that strictly adheres to economic theory produces forecasts that robustly generalize beyond the cross-section of stocks compared to the simpler model. Specifically, the aggregate market forecast robustly predicts time-series variations in market returns up to three years into the future. And, the aggregate long-short portfolio forecasts predict time-series variation in the following month's return for 53 out of the 56 long-short portfolios. The simpler neural network fails along these two dimensions.

I disentangle the nature of the stock forecasts and show that short-run stock return predictability is very different from long-run predictability. Monthly forecasts for up to one year into the future, predict cross-sectional variations in stock returns and time-series variation in relative stock returns, in addition to predicting the level factor in returns. In contrast, forecasts for periods beyond one year only predict variations in the level factor.

Studying the time-series and cross-sectional properties of conditional expected returns across multiple horizons is important for many reasons; 1) such studies naturally produce new test portfolios for examining asset-pricing models. 2) they uncover new stylized facts about expected returns that constitute a new set of moments for emerging theoretical models to match. 3) their findings are practically useful because more accurate stock return forecasts allow us to devise better trading strategies and find better costs of capital estimates for future cash-flows.

In light of this paper's findings, it would be interesting in future research to answer the question, "Why does cross-sectional return predictability decay so quickly along the horizon dimension?" One possible hypothesis is that most of the predictable cross-section variation in next month's return is due to mispricing, which is then corrected quickly. And so beyond the following month, there is very little cross-sectional mispricing information

in the information set we are conditioning on. Another hypothesis is that today's firms are very different from their future selves if we compare them based on their characteristic realizations'. If this is true, then characteristic realizations today will have a lot less to say about future cross-sectional dispersion in returns the further in the future we want to predict.

References

Abarbanell, J. S., Bushee, B. J., 1997. Fundamental analysis, future earnings, and stock prices. Journal of accounting research 35, 1–24.

Anderson, A.-M., Dyl, E. A., 2005. Market structure and trading volume. Journal of Financial Research 28, 115–131.

Ang, A., Hodrick, R. J., Xing, Y., Zhang, X., 2006. The cross-section of volatility and expected returns. Journal of Finance .

Avramov, D., Cheng, S., Metzker, L., 2020. Machine learning versus economic restrictions: Evidence from stock return predictability. Available at SSRN 3450322 .

Baba-Yara, F., Boons, M., Tamoni, A., 2018. Value return predictability across asset classes and commonalities in risk premia. Review of Finance, (forthcoming) .

Baba-Yara, F., Boons, M., Tamoni, A., 2020. New and Old Sorts: Implications for Asset Pricing. Available at SSRN 3529140 p. 69.

Balakrishnan, K., Bartov, E., Faurel, L., 2010. Post loss/profit announcement drift. Journal of Accounting and Economics 50, 20–41.

Bali, T. G., Cakici, N., Whitelaw, R. F., 2011. Maxing out: Stocks as lotteries and the cross-section of expected returns. Journal of Financial Economics 99, 427–446.

Ball, R., Gerakos, J., Linnainmaa, J. T., Nikolaev, V. V., 2015. Deflating profitability. Journal of Financial Economics 117, 225–248.

Bandyopadhyay, S. P., Huang, A. G., Wirjanto, T. S., 2010. The accrual volatility anomaly. Unpublished Manuscript, University of Waterloo .

Basu, S., 1977. Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. The journal of Finance 32, 663–682.

Basu, S., 1983. The relationship between earnings' yield, market value and return for nyse common stocks: Further evidence. Journal of financial economics 12, 129–156.

Ben-Rephael, A., Kadan, O., Wohl, A., 2015. The diminishing liquidity premium. Journal of Financial and Quantitative Analysis pp. 197–229.

Bhandari, L. C., 1988. Debt/equity ratio and expected common stock returns: Empirical evidence. The journal of finance 43, 507–528.

Brown, D. P., Rowe, B., 2007. The productivity premium in equity returns. Available at SSRN 993467 .

Bryzgalova, S., Pelger, M., Zhu, J., 2019. Forest through the trees: Building cross-sections of stock returns. Available at SSRN 3493458 .

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., Siddique, A., 2016. Risk and risk management in the credit card industry. Journal of Banking & Finance 72, 218–239.

Chandrashekar, S., Rao, R. K., 2009. The productivity of corporate cash holdings and the cross-section of expected stock returns. McCombs Research Paper Series No. FIN-03-09 .

Chen, L., Pelger, M., Zhu, J., 2019. Deep learning in asset pricing. Available at SSRN 3350138 .

Chordia, T., Goyal, A., Shanken, J. A., 2017. Cross-sectional asset pricing with individual stocks: betas versus characteristics. Available at SSRN 2549578 .

Chordia, T., Roll, R., Subrahmanyam, A., 2011. Recent trends in trading activity and market quality. Journal of Financial Economics 101, 243–263.

Chordia, T., Subrahmanyam, A., Anshuman, V. R., 2001. Trading activity and expected stock returns. Journal of Financial Economics 59, 3–32.

Chordia, T., Subrahmanyam, A., Tong, Q., 2013. Trends in the cross-section of expected stock returns. SSRN eLibrary .

Chung, K. H., Zhang, H., 2014. A simple approximation of intraday spreads using daily data. Journal of Financial Markets 17, 94–120.

Clark, T. E., West, K. D., 2007. Approximately normal tests for equal predictive accuracy in nested models. Journal of econometrics 138, 291–311.

Cochrane, J. H., 2008. The dog that did not bark: A defense of return predictability. The Review of Financial Studies 21, 1533–1575.

Cochrane, J. H., 2011. Presidential address: Discount rates. The Journal of finance 66, 1047–1108.

Cohen, R. B., Polk, C., Vuolteenaho, T., 2003. The value spread. The Journal of Finance 58, 609–641.

Cooper, M. J., Gulen, H., Schill, M. J., 2008. Asset growth and the cross-section of stock returns. The Journal of Finance 63, 1609–1651.

Cooper, M. J., Gutierrez Jr, R. C., Hameed, A., 2004. Market states and momentum. The journal of Finance 59, 1345–1365.

Cosemans, M., Frehen, R., Schotman, P. C., Bauer, R., 2016. Estimating security betas using prior information based on firm fundamentals. The Review of Financial Studies 29, 1072–1112.

D'Acunto, F., Liu, R., Pflueger, C., Weber, M., 2018. Flexible prices and leverage. Journal of Financial Economics 129, 46–68.

Daniel, K., Moskowitz, T. J., 2016. Momentum crashes. Journal of Financial Economics 122, 221–247.

Datar, V. T., Naik, N. Y., Radcliffe, R., 1998. Liquidity and stock returns: An alternative test. Journal of Financial Markets 1, 203–219.

Davis, J. L., Fama, E. F., French, K. R., 2000. Characteristics, covariances, and average returns: 1929 to 1997. The Journal of Finance 55, 389–406.

De Bondt, W. F., Thaler, R., 1985. Does the stock market overreact? The Journal of finance 40, 793–805.

Desai, H., Rajgopal, S., Venkatachalam, M., 2004. Value-glamour and accruals mispricing: One anomaly or two? The Accounting Review 79, 355–385.

Engelberg, J., McLean, R. D., Pontiff, J., Ringgenberg, M., 2019. Are cross-sectional predictors good market-level predictors? In: *American Finance Association Annual Meeting Paper*.

Fama, E. F., French, K. R., 1992. The cross-section of expected stock returns. the Journal of Finance 47, 427–465.

Fama, E. F., French, K. R., 1993. Common risk factors in the returns on stocks and bonds. Journal of financial economics 33, 3–56.

Fama, E. F., French, K. R., 1996. Multifactor explanations of asset pricing anomalies. The journal of finance 51, 55–84.

Fama, E. F., French, K. R., 2008. Dissecting anomalies. The Journal of Finance 63, 1653–1678.

Fama, E. F., French, K. R., 2018. Choosing factors. Journal of Financial Economics 128, 234–252.

Ferreira, M. A., Santa-Clara, P., 2011. Forecasting stock market returns: The sum of the parts is more than the whole. Journal of Financial Economics 100, 514–537.

Frazzini, A., Pedersen, L. H., 2014. Betting against beta. Journal of Financial Economics 111, 1–25.

Freyberger, J., Neuhierl, A., Weber, M., 2020. Dissecting characteristics nonparametrically. Tech. Rep. 5.

Gandhi, P., Lustig, H., 2015. Size anomalies in us bank stock returns. The Journal of Finance 70, 733–768.

Garfinkel, J. A., 2009. Measuring investors' opinion divergence. Journal of Accounting Research 47, 1317–1348.

George, T. J., Hwang, C.-Y., 2004. The 52-week high and momentum investing. The Journal of Finance 59, 2145–2176.

Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., 2016. Deep learning, vol. 1. MIT press Cambridge.

Gorodnichenko, Y., Weber, M., 2016. Are sticky prices costly? evidence from the stock market. American Economic Review 106, 165–99.

Goyal, A., Welch, I., 2008. A comprehensive look at the empirical performance of equity premium prediction. Review of Financial Studies 21, 1455–1508.

Green, J., Hand, J. R., Zhang, X. F., 2017. The characteristics that provide independent information about average us monthly stock returns. The Review of Financial Studies 30, 4389–4436.

Gu, S., Kelly, B., Xiu, D., 2020a. Autoencoder asset pricing models. Journal of Econometrics .

Gu, S., Kelly, B., Xiu, D., 2020b. Empirical asset pricing via machine learning. The Review of Financial Studies 33, 2223–2273.

Haddad, V., Kozak, S., Santosh, S., 2020. Factor timing. Tech. Rep. 5.

Haugen, R. A., Baker, N. L., 1996. Commonality in the determinants of expected stock returns. Journal of financial economics 41, 401–439.

Hirshleifer, D., Hou, K., Teoh, S. H., Zhang, Y., 2004. Do investors overvalue firms with bloated balance sheets? Journal of Accounting and Economics 38, 297–331.

Hou, K., Karolyi, G. A., Kho, B.-C., 2011. What factors drive global stock returns? The Review of Financial Studies 24, 2527–2574.

Hou, K., Xue, C., Zhang, L., 2015. Digesting anomalies: An investment approach. The Review of Financial Studies 28, 650–705.

Hou, K., Xue, C., Zhang, L., 2020. Replicating anomalies. Tech. Rep. 5.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 .

Israel, R., Laursen, K., Richardson, S. A., 2020. Is (systematic) value investing dead? Available at SSRN .

Jegadeesh, N., 1990. Evidence of predictable behavior of security returns. The Journal of finance 45, 881–898.

Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. The Journal of finance 48, 65–91.

Johnson, J., 2019. Deep, skinny neural networks are not universal approximators. In: *7th International Conference on Learning Representations, ICLR 2019*.

Kelly, B. T., Pruitt, S., Su, Y., 2019. Characteristics are covariances: A unified model of risk and return. Journal of Financial Economics .

Keloharju, M., Linnainmaa, J. T., Nyberg, P. M., 2019. Long-term discount rates do not vary across firms. Available at SSRN 3125502 .

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

Kozak, S., 2019. Kernel trick for the cross-section. Available at SSRN 3307895 .

Kozak, S., Nagel, S., Santosh, S., 2020. Shrinking the cross-section. Journal of Financial Economics 135, 271–292.

Lakonishok, J., Shleifer, A., Vishny, R. W., 1994. Contrarian investment, extrapolation, and risk. The journal of finance 49, 1541–1578.

Lewellen, J., 2015. The cross section of expected stock returns. Critical Finance Review .

Lewellen, J., Nagel, S., 2006. The conditional capm does not explain asset-pricing anomalies. Journal of financial economics 82, 289–314.

Litzenberger, R. H., Ramaswamy, K., 1979. The effect of personal taxes and dividends on capital asset prices: Theory and empirical evidence. Journal of financial economics 7, 163–195.

Lundberg, S. M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: *Advances in neural information processing systems*, pp. 4765–4774.

Luo, L., Xiong, Y., Liu, Y., Sun, X., 2019. Adaptive gradient methods with dynamic bound of learning rate. arXiv preprint arXiv:1902.09843 .

Lyandres, E., Sun, L., Zhang, L., 2008. The new issues puzzle: Testing the investment-based explanation. The Review of Financial Studies 21, 2825–2855.

Martin, I., Nagel, S., 2019. Market efficiency in the age of big data. Tech. rep., National Bureau of Economic Research.

McLean, R. D., Pontiff, J., 2016. Does academic research destroy stock return predictability? The Journal of Finance 71, 5–32.

Newey, W. K., West, K. D., 1987. Hypothesis testing with efficient method of moments estimation. International Economic Review pp. 777–787.

Novy-Marx, R., 2011. Operating leverage. Review of Finance 15, 103–134.

Novy-Marx, R., 2012. Is momentum really momentum? Journal of Financial Economics 103, 429–453.

Ou, J. A., Penman, S. H., 1989. Financial statement analysis and the prediction of stock returns. Journal of accounting and economics 11, 295–329.

Palazzo, B., 2012. Cash holdings, risk, and expected returns. Journal of Financial Economics 104, 162–185.

Pontiff, J., Woodgate, A., 2008. Share issuance and cross-sectional returns. The Journal of Finance 63, 921–945.

Recanatesi, S., Farrell, M., Advani, M., Moore, T., Lajoie, G., Shea-Brown, E., 2019. Dimensionality compression and expansion in deep neural networks. arXiv preprint arXiv:1906.00443 .

Richardson, S. A., Sloan, R. G., Soliman, M. T., Tuna, I., 2005. Accrual reliability, earnings persistence and stock prices. Journal of accounting and economics 39, 437–485.

Sirignano, J., Sadhwani, A., Giesecke, K., 2016. Deep learning for mortgage risk. arXiv preprint arXiv:1607.02470 .

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

Soliman, M. T., 2008. The use of dupont analysis by market participants. The Accounting Review 83, 823–853.

Stambaugh, R. F., Yuan, Y., 2017. Mispricing factors. The Review of Financial Studies 30, 1270–1315.

Thomas, J. K., Zhang, H., 2002. Inventory changes and future returns. Review of Accounting Studies 7, 163–187.

Van Binsbergen, J. H., Opp, C. C., 2019. Real anomalies. The Journal of Finance 74, 1659–1706.

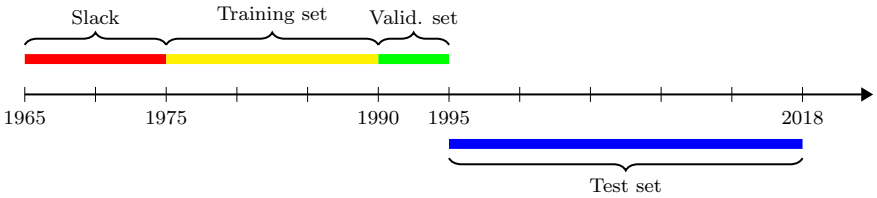

Figure 1: **Sample Splitting time-line**
This figure presents a time-line for sample splitting scheme in the paper.

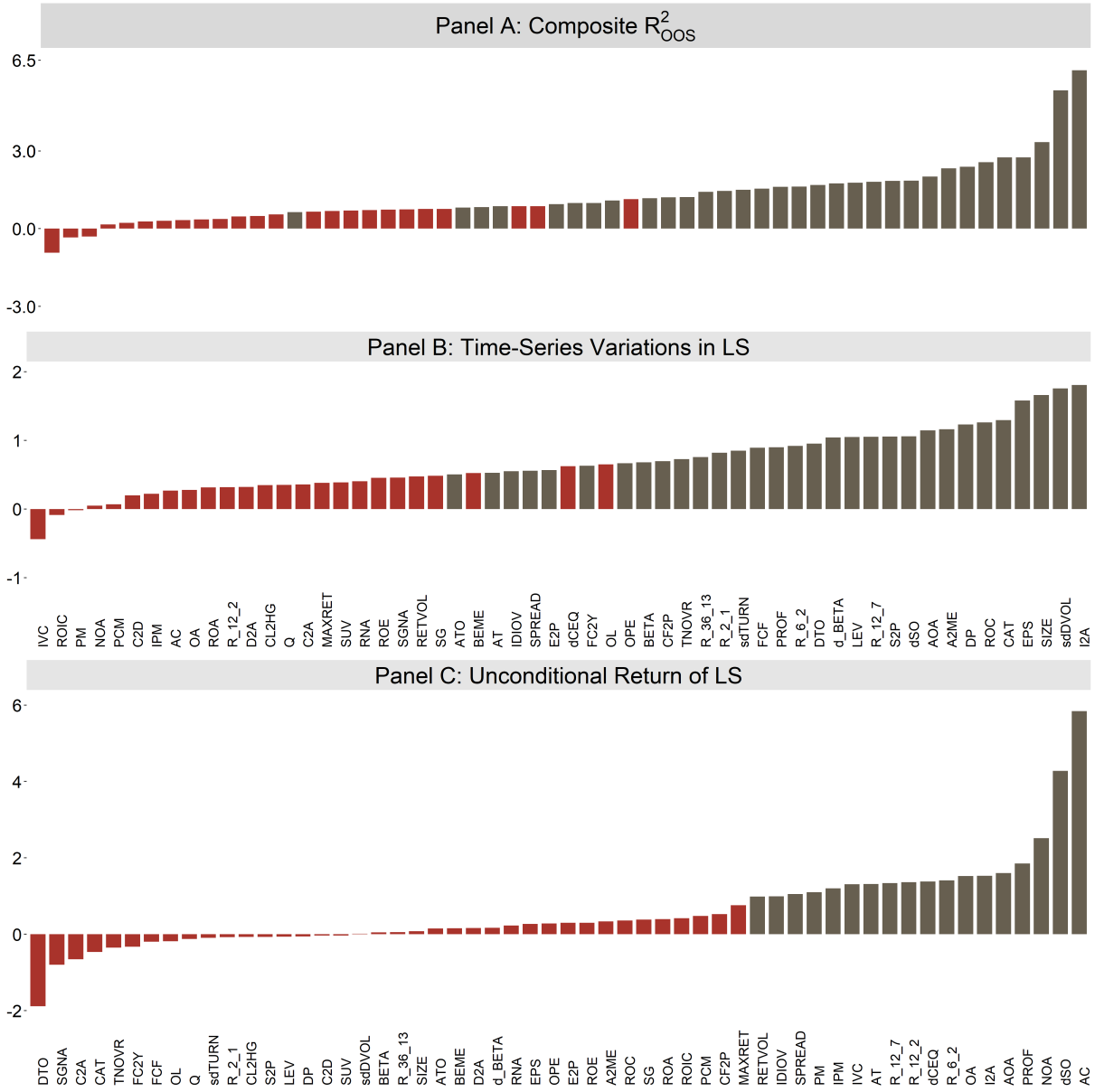

Figure 2: **Forecasting returns of long-short portfolios**
This figure reports the out-of-sample R^2 (R^2_{OOS}) of monthly re-balanced long-short portfolios formed from sorts on 56 characteristics listed in Table B.1 using forecasts from time $t-h$, where $h \in \{1, 2, 3, 13, 25, 37, 49, 61, 121\}$. Grey bars represent statistically significant R^2_{OOS} using the the Clark-West (2007) test at the 5% level. Panel A reports the composite R^2_{OOS}, Panel B reports the R^2_{OOS} contribution that comes from the ability of the neural network forecasts to better predict time-series variation in returns to long-short portfolios, and Panel C reports the R^2_{OOS} contribution that comes from the ability of the neural network forecasts to better predict the unconditional returns to long-short portfolios. The alternative model is the historical portfolio return computed from an expanding window mean with data from 1965. The sample period is from January 1995 to December 2018.

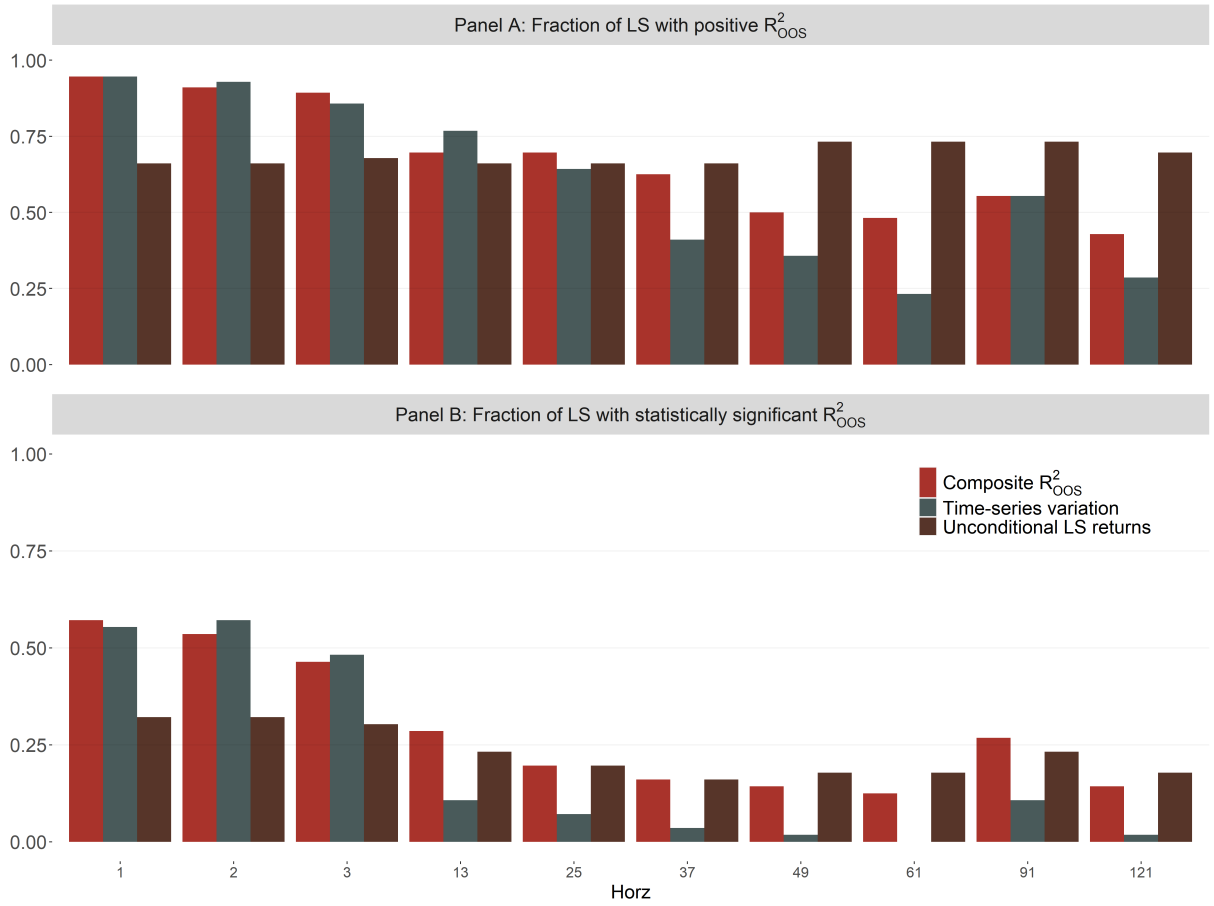

Figure 3: **Forecasting returns of long-short portfolios using forecasts for different horizons.**
This figure reports statistics for long-short portfolios formed from sorts on 56 characteristics listed in Table B.1. We report in panel A the fraction of 56 long-short portfolios where the neural network forecast has a positive R^2_{OOS} with respect to the historical portfolio return computed from an expanding window mean with data from 1965. Panel B reports the fraction of 56 long-short portfolios where the forecasts from the neural network model are statistically better than the forecasts from the zero-prediction model at the 5% level using the the Clark-West (2007) test. For each prediction $t-h$, we report fractions that pertain to results for the composite R^2_{OOS}, the contributions coming from better predicting time-series variations in return (grey), and improvements coming from predicting the unconditional portfolio return (brown). The sample period is from January 1995 to December 2018.

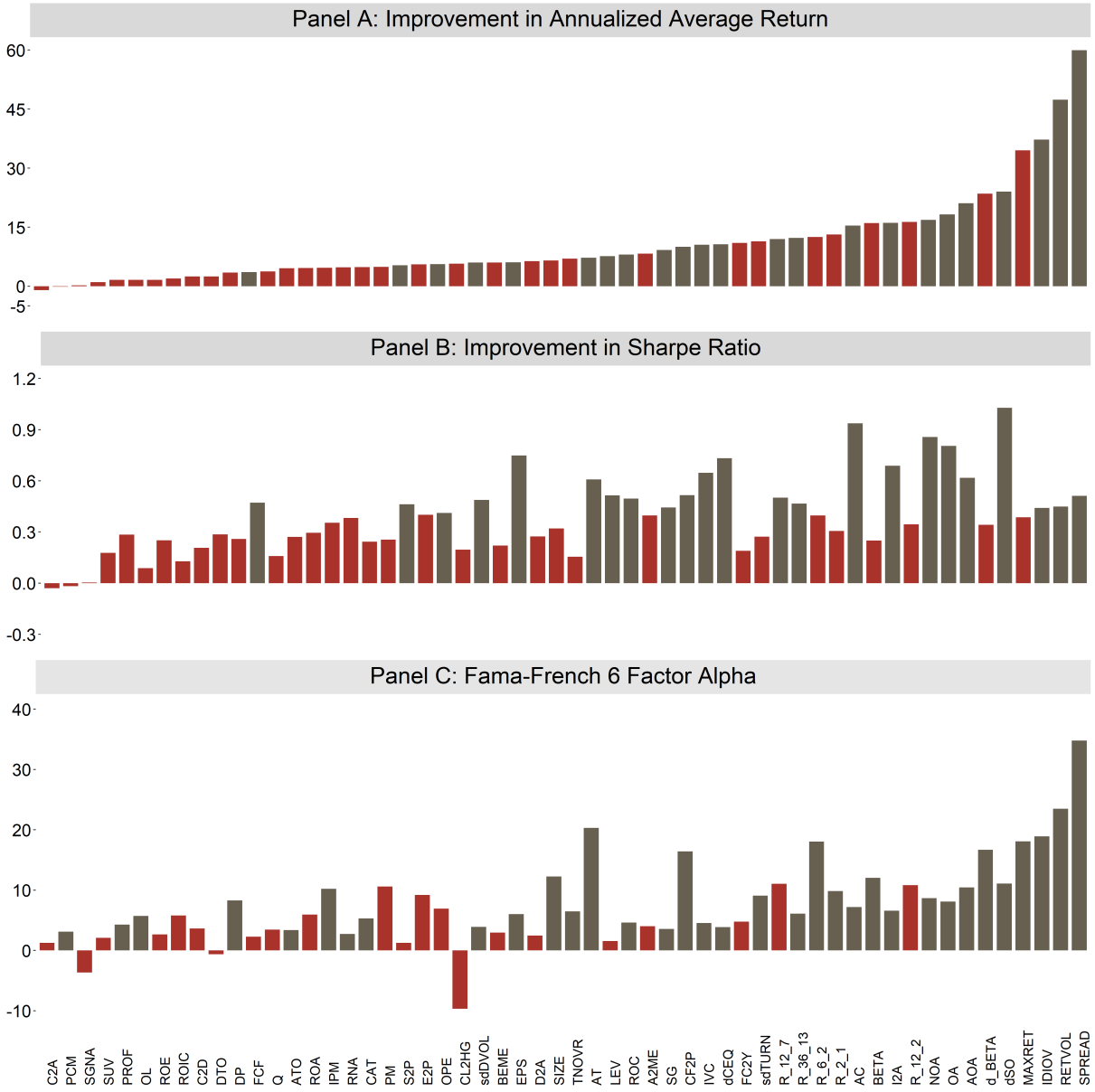

Figure 4: **Timing long-short characteristic sorted portfolios**
This figure reports the performance statistics of a portfolio that times each of the 56 long-short portfolios defined in Figure 2 and the risk-free asset. For Panels A and B, we report the performance statistics for the difference in returns between holding the timing portfolio and the buy and hold alternative. Panel A reports the annualized average return, Panel B reports the Sharpe ratio improvement, and panel C reports the Fama and French (2018) 6 factor model alpha (annualized) from regressing the returns of the timing portfolio on the pricing factors. Grey bars in Panels A and B represent long-short portfolios for which the improvement in returns is statistically significant at the 5% level. Grey bars in Panel C represent long-short portfolio returns with a statically significant alpha at the 5 % level with respect to the Fama-French 6 factor model. The sample period is from January 1995 to December 2018.

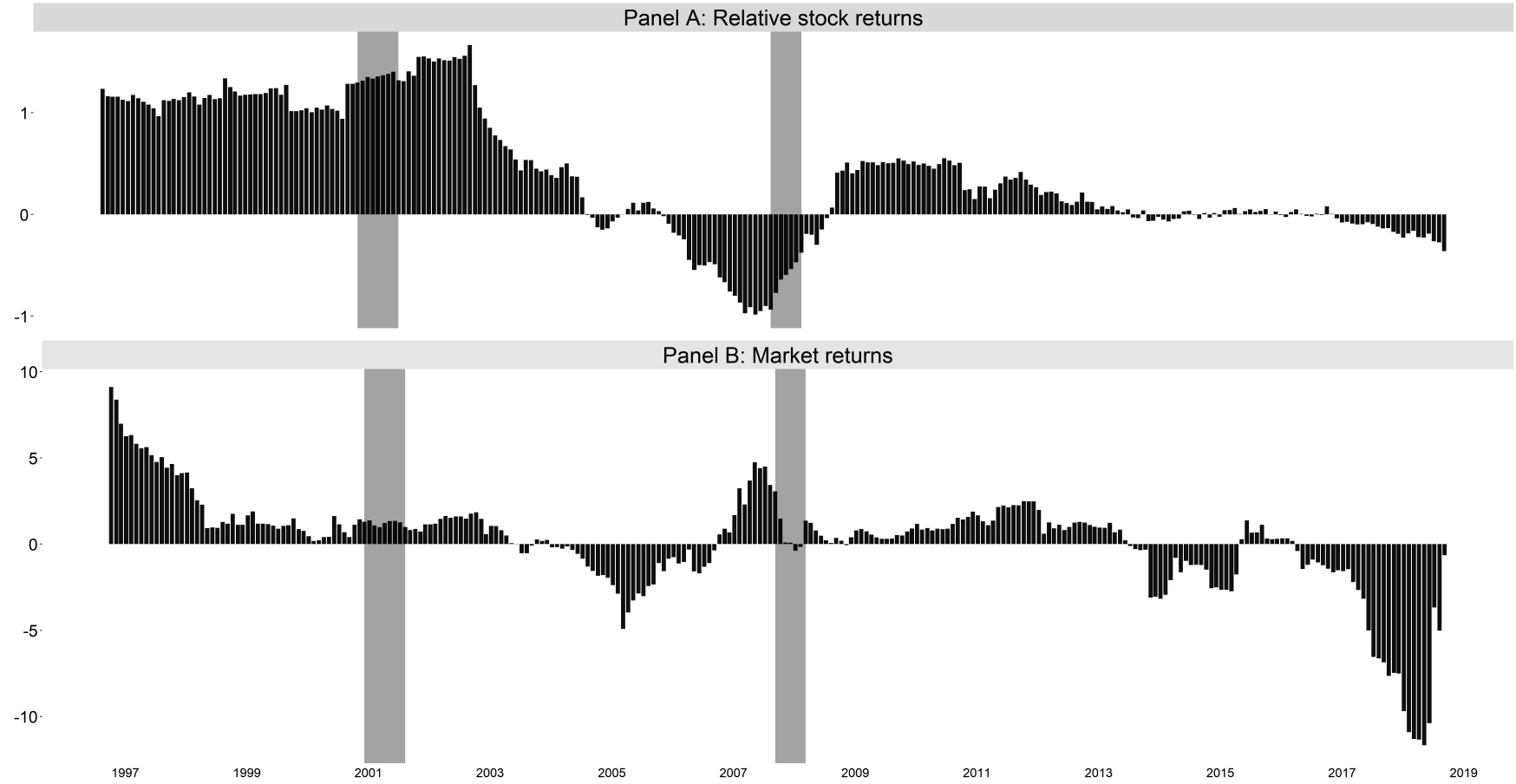

Figure 5: **Two-year rolling out-of-sample R^2**
This figure reports a two-year rolling out-of-sample R^2 (R^2_{OOS}) for demeaned monthly return forecasts from the economically restricted neural network model. The rolling estimates capture the changes in the neural network forecasts' ability to predict time-series variations in stock returns. In panel A, we report this estimate for the pool of stocks, and in panel B, we report this estimate for the value-weighted market. The alternative model is a zero prediction model and the period is $t+1$. The sample period is from January 1995 to December 2018.

Figure 6: **Variable importance with Shapley values**
This figure reports the top four variables that increase return forecasts the most. A one-unit increase in a variable changes the return forecast by the corresponding Shapley value. A red (blue) bar represents a positive (negative) change in return forecasts. Shapley values are computed for each firm-month observation in the out-of-sample period and averaged over all observations. The sample period is from January 1995 to December 2018.

Table 1: **Predicting stock returns across horizons**
This table presents R^2_{OOS} estimates for firm-level return forecasts generated at time t for month $t+h$ where $h \in \{1, 2, 3, 13, 37, 61, 91, 121\}$. The R^2_{OOS} is defined with respect to a zero prediction benchmark. Panel A reports the R^2_{OOS} for a neural network forecasting model that adheres strictly to economic theory as defined in Equation (2) and a simpler neural network forecasting model with no economic restrictions as defined in Eq. (8). Panel B reports the decomposition of the R^2_{OOS} into three parts. The time-series variation dimension captures the forecasting model's ability to improve on forecasts from the benchmark model with respect to predicting time-series variation in relative stock returns. The unconditional relative stock return dimension captures the forecasting model's ability to improve upon on forecasts from the benchmark model with respect to predicting the unconditional relative stock return. Finally, the cross-sectional mean return dimension captures the forecasting model's ability to improve upon forecasts from the benchmark model with respect to predicting the level factor in stock returns; the equally-weighted market return. Bold fonts indicate horizons for which the forecasts from the neural network model and alternative model are statistically different at the 5% level or better using the Clark-West (2007) test. The sample period is from January 1995 to December 2018.

h	1	2	3	13	37	61	91	121
Panel A: Composite R^2_{OOS}								
Economically restricted model	**0.99**	**0.49**	**0.40**	**0.35**	**0.28**	**0.28**	**0.24**	**0.13**
Simple model	**0.58**	**0.41**	**0.40**	**0.37**	**0.35**	**0.34**	**0.27**	**0.18**
Panel B: Decomposed R^2_{OOS}								
	Economically restricted model							
Time-series variation	**0.71**	**0.16**	**0.09**	**0.08**	-0.06	-0.05	-0.07	-0.09
Unconditional rel. stock ret.	-0.16	-0.03	-0.02	-0.08	0.03	0.02	0.04	0.03
Cross-sectional mean return	**0.43**	**0.36**	**0.33**	**0.36**	**0.32**	**0.31**	**0.27**	**0.20**
	Simple model							
Time-series variation	**0.20**	**0.01**	**0.02**	**0.03**	-0.04	-0.04	-0.07	-0.07
Unconditional rel. stock ret.	-0.04	-0.01	-0.02	-0.05	-0.01	-0.00	0.00	0.00
Cross-sectional mean return	**0.41**	**0.41**	**0.41**	**0.40**	**0.40**	**0.39**	**0.33**	**0.24**

Table 2: **Historical average market return benchmark**
This table is similar to Table 1 but the benchmark model is the historical average equally-weighted market return computed from an expanding window using data from 1926. Panel A reports the R^2_{OOS} for a neural network forecasting model that adheres strictly to economic theory as defined in Equation (2), and panel B reports results for a simpler neural network forecasting model with no economic restrictions as defined in Eq. (8). Bold fonts indicate horizons for which the forecasts from the neural network model and alternative model are statistically different at the 5% level or better using the Clark-West (2007) test. The sample period is from January 1995 to December 2018.

h	1	2	3	13	37	61	91	121
Panel A: Economically restricted model								
Composite R^2_{OOS}	**0.64**	**0.14**	**0.04**	0.00	**0.05**	**0.16**	**0.37**	**0.62**
Time-series variation	**0.71**	**0.16**	**0.09**	**0.08**	-0.06	-0.05	-0.07	-0.09
Unconditional rel. stock ret.	-0.16	-0.03	-0.02	-0.08	0.03	0.02	0.04	0.03
Cross-sectional mean return	**0.08**	**0.01**	-0.02	**0.01**	**0.08**	**0.19**	**0.40**	**0.69**
Panel B: Simple model								
Composite R^2_{OOS}	**0.20**	**0.03**	**0.01**	-0.00	**0.04**	**0.17**	**0.32**	**0.55**
Time-series variation	**0.20**	**0.01**	**0.02**	**0.03**	-0.04	-0.04	-0.07	-0.07
Unconditional rel. stock ret.	-0.04	-0.01	-0.02	-0.05	-0.01	-0.00	0.00	0.00
Cross-sectional mean return	**0.03**	**0.03**	**0.02**	**0.02**	**0.09**	**0.21**	**0.39**	**0.61**

Table 3: **Predicting market returns across horizons**
This table presents out-of-sample (R^2_{OOS}) estimates for market forecasts for different months in the future $t+h$, $h \in \{1, 2, 3, 13, 37, 61, 91, 121\}$, conditional on information observed at time t. Panel A reports results for a neural network forecasting model that adheres strictly to economic theory as defined in Equation (2). Panel B reports results for a simpler neural network forecasting model with no economic restrictions as defined in Eq. (8). The market forecasts are defined as the value-weighted cross-sectional average stock forecasts and the market return as the value-weighted cross-sectional average stock return. I report results for two benchmark models, a zero-prediction model that predicts a return of zero for all horizons and the historical equity market benchmark that forecasts market return as the average market return using data from July 1926 upto time t. I decompose the R^2_{OOS} into contributions coming from the neural network forecasts ability to explain better time-series variation (Time-series variation) and the unconditional market return (Uncond. market ret.). Bold fonts indicate horizons for which the forecasts from the neural network model and alternative model are statistically different at the 5% level or better using the Clark-West (2007) test. The sample period is from January 1995 to December 2018.

h	1	2	3	13	37	61	91	121
Panel A: Economically restricted model								
	Zero prediction benchmark							
Composite R^2_{OOS}	**5.35**	**5.30**	**5.17**	**4.90**	**4.42**	**3.05**	**3.48**	**2.77**
Time-series variation	**1.92**	**1.82**	**1.69**	**1.51**	**1.10**	-0.03	0.19	-0.42
Uncond. market ret.	**3.42**	**3.49**	**3.49**	**3.39**	**3.31**	**3.08**	**3.30**	**3.19**
	Historical average market benchmark							
Composite R^2_{OOS}	**2.05**	**2.03**	**1.89**	**1.64**	1.09	-0.40	0.13	-0.63
Time-series variation	**2.06**	**1.97**	**1.84**	**1.68**	**1.21**	-0.04	0.26	-0.39
Uncond. market ret.	-0.01	0.06	0.05	-0.04	-0.12	-0.36	-0.13	-0.24
Panel B: Simple model								
	Zero prediction benchmark							
Composite R^2_{OOS}	**1.43**	**2.53**	**2.37**	**3.08**	**3.65**	**4.14**	**3.47**	**3.72**
Time-series variation	0.21	0.48	0.26	0.54	0.60	0.70	0.01	0.03
Uncond. market ret.	**1.23**	**2.06**	**2.10**	**2.54**	**3.05**	**3.44**	**3.46**	**3.69**
	Historical average market benchmark							
Composite R^2_{OOS}	-2.29	-1.13	-1.31	-0.52	0.04	0.47	-0.16	0.07
Time-series variation	0.29	0.59	0.36	0.70	0.73	0.75	0.10	0.09
Uncond. market ret.	-2.59	-1.72	-1.67	-1.22	-0.69	-0.28	-0.25	-0.02

Table 4: **Predicting long-short characteristic sorted portfolio returns (Simple model)** This table reports the out-of-sample R^2 (R^2_{OOS}) estimates for monthly rebalanced long-short portfolios formed from sorts on book-to-market (BEME), investments (INV), size (SIZE), mom (Momentum), and profitability (PROF). The results are for the simple neural network forecasting model, as defined in Equation (8). The benchmark model is the zero prediction model. Bold fonts indicate horizons for which the forecasts from the simple neural network model and alternative model are statistically different at the 5% level or better using the Clark-West (2007) test. The sample period is from January 1995 to December 2018.

h	1	2	3	13
		BEME		
Composite R^2_{OOS}	-1.94	-1.53	-2.10	-0.78
Time-series variation	-2.64	-1.33	-1.48	-0.32
Unconditional LS Ret.	0.70	-0.20	-0.62	-0.46
		INV		
Composite R^2_{OOS}	-0.78	**2.91**	**3.60**	0.41
Time-series variation	0.15	0.89	0.97	0.63
Unconditional LS Ret.	-0.92	**2.02**	**2.62**	-0.22
		SIZE		
Composite R^2_{OOS}	0.44	0.61	0.48	-0.15
Time-series variation	0.45	0.88	0.89	0.11
Unconditional LS Ret.	-0.01	-0.27	-0.41	-0.26
		MOM		
Composite R^2_{OOS}	-1.47	-0.98	-2.60	-2.29
Time-series variation	-1.24	-0.53	-1.31	-0.96
Unconditional LS Ret.	-0.23	-0.44	-1.29	-1.34
		PROF		
Composite R^2_{OOS}	-0.56	0.14	0.28	1.23
Time-series variation	-0.74	-0.06	0.28	1.00
Unconditional LS Ret.	0.18	0.20	0.00	0.23

Table 5: **Predicting long-short portfolio returns (Economically restricted model)** This table reports the out-of-sample R^2 (R^2_{OOS}) estimates for monthly rebalanced long-short portfolios formed from sorts on book-to-market (BEME), investments (INV), size (SIZE), mom (Momentum), and profitability (PROF). The results are for the neural network forecasting model that adheres strictly to economic theory as defined in Equation (2). The benchmark model is the historical average return of the long-short portfolio computed from an expanding window using data from 1964. Bold fonts indicate horizons for which the forecasts from the neural network model and alternative model are statistically different at the 5% level or better using the Clark-West (2007) test. The sample period is from January 1995 to December 2018.

h	1	2	3	13	37	61	91	121
				BEME				
Composite R^2_{OOS}	0.68	0.80	0.73	0.81	-0.30	-0.17	-0.34	-0.23
Time-series variation	0.52	0.62	0.51	-0.03	-0.30	-0.13	-0.20	0.01
Unconditional LS Ret.	0.15	0.18	0.22	0.84	-0.00	-0.04	-0.14	-0.24
				SIZE				
Composite R^2_{OOS}	**1.74**	**1.48**	**1.32**	**0.71**	0.14	0.01	-0.45	-0.56
Time-series variation	**1.66**	**1.43**	**1.28**	**0.78**	0.10	-0.06	-0.41	-0.44
Unconditional LS Ret.	0.08	0.05	0.04	-0.08	0.04	0.07	-0.04	-0.12
				PROF				
Composite R^2_{OOS}	**0.95**	**0.97**	0.75	0.55	-0.06	0.07	0.24	-0.12
Time-series variation	**0.67**	**0.72**	**0.51**	0.28	-0.03	-0.05	0.15	-0.14
Unconditional LS Ret.	0.28	0.24	0.25	0.27	-0.03	0.12	0.09	0.01
				INV				
Composite R^2_{OOS}	**3.34**	**3.49**	**3.23**	**1.93**	0.30	0.17	0.00	0.36
Time-series variation	**1.81**	**1.95**	**1.67**	0.56	0.07	-0.13	-0.09	-0.01
Unconditional LS Ret.	**1.53**	**1.55**	**1.56**	**1.37**	**0.23**	0.30	0.10	0.37
				MOM				
Composite R^2_{OOS}	**1.34**	**0.84**	0.43	**0.41**	-0.21	-0.10	-0.25	-0.50
Time-series variation	0.22	0.18	0.22	0.28	0.02	-0.02	-0.04	-0.14
Unconditional LS Ret.	**1.12**	**0.66**	0.20	**0.14**	-0.24	-0.08	-0.21	-0.37

Table 6: **Time-series regression of returns on forecasts**

This table reports the results from a time-series regression of returns on the economically restricted model's demeaned forecasts. Panel A reports results from regressing stock returns on demeaned relative stock forecasts. Panel B reports results from regressing the value-weighted market return on demeaned value-weighted market forecasts. Panels C and D report results from regressing long-short portfolio returns on demeaned long-short portfolio forecasts. β_0 and β_1 are the intercept and slope coefficients from these regressions. For each of the estimates, we report the 95% confidence interval constructed from Newey-West 1987 corrected standard errors. The sample period is from January 1995 to December 2018.

h	1	2	3	13	25	37	49	61	91	121
					Panel A: Pool of stocks					
β_0	1.08	1.08	1.08	1.02	0.93	0.83	0.78	0.72	0.59	0.47
(β_0^l, β_0^u)	(1.05,1.11)	(1.05,1.11)	(1.05,1.11)	(1.00,1.05)	(0.90,0.96)	(0.81,0.86)	(0.76,0.80)	(0.70,0.74)	(0.57,0.60)	(0.46,0.49)
β_1	0.97	0.80	0.68	0.79	0.24	-0.02	0.03	0.11	-0.12	-0.04
(β_1^l, β_1^u)	(0.94,1.00)	(0.76,0.84)	(0.64,0.72)	(0.74,0.85)	(0.19,0.30)	(-0.09,0.05)	(-0.04,0.09)	(0.04,0.17)	(-0.19,-0.05)	(-0.10,0.02)
R^2	0.71	0.17	0.11	0.09	0.01	0.00	0.00	0.00	0.00	0.00
$NOS.$	1507557	1508884	1510083	1519230	1525769	1529073	1532204	1536016	1543943	1535857
					Panel B: Value-weighted market					
β_0	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79
(β_0^l, β_0^u)	(0.28,1.30)	(0.28,1.30)	(0.28,1.30)	(0.28,1.30)	(0.28,1.30)	(0.28,1.30)	(0.28,1.30)	(0.27,1.30)	(0.27,1.30)	(0.27,1.30)
β_1	1.78	1.58	1.34	1.69	1.45	1.63	1.66	0.42	0.44	-0.22
(β_1^l, β_1^u)	(0.44,3.12)	(0.24,2.92)	(0.09,2.60)	(0.23,3.16)	(0.11,2.79)	(0.01,3.24)	(-0.24,3.57)	(-1.61,2.45)	(-1.14,2.02)	(-2.17,1.73)
R^2	2.18	1.79	1.53	1.53	1.49	1.05	0.82	0.05	0.08	0.02
$NOS.$	282	282	282	282	282	282	282	282	282	282
					Panel C: Pool of long-short portfolios					
β_0	0.31	0.13	0.28	0.31	0.25	0.29	0.23	0.23	0.27	0.26
(β_0^l, β_0^u)	(-0.04,0.65)	(-0.16,0.42)	(-0.05,0.61)	(-0.03,0.65)	(-0.08,0.57)	(-0.05,0.63)	(-0.09,0.56)	(-0.08,0.55)	(-0.05,0.58)	(-0.05,0.56)
β_1	1.96	-2.93	2.29	2.20	2.12	2.07	0.35	-0.73	-0.98	-0.99
(β_1^l, β_1^u)	(0.88,3.04)	(-6.34,0.48)	(0.67,3.90)	(1.19,3.21)	(-0.34,4.58)	(0.98,3.16)	(-2.87,3.56)	(-4.67,3.22)	(-4.67,2.71)	(-4.32,2.34)
R^2	1.74	0.22	0.66	1.81	0.23	1.48	0.00	0.01	0.02	0.03
$NOS.$	282	282	282	282	282	282	282	282	282	282

Continued

h	1	2	3	13	25	37	49	61	91	121
					Panel D: Individual long-short portfolios					
					BEME					
β_1	1.39	-0.21	0.57	2.73	-1.69	2.45	-8.14	-9.46	-3.97	-10.64
(β_1^l, β_1^u)	(-1.33,4.11)	(-13.18,12.77)	(-4.86,6.01)	(-1.12,6.58)	(-10.29,6.90)	(-1.69,6.60)	(-18.32,2.05)	(-18.59,-0.34)	(-11.92,3.98)	(-22.79,1.51)
R^2	0.56	0.00	0.02	1.04	0.09	0.77	1.16	1.62	0.27	0.95
					SIZE					
β_1	4.99	-6.17	6.55	4.36	4.20	3.80	3.67	3.49	-0.93	-6.85
(β_1^l, β_1^u)	(1.98,8.00)	(-12.56,0.22)	(1.20,11.91)	(1.55,7.17)	(-1.67,10.07)	(1.11,6.49)	(-5.70,13.04)	(-7.75,14.74)	(-9.19,7.32)	(-13.13,-0.56)
R^2	4.60	1.23	2.83	3.51	0.78	2.80	0.22	0.10	0.01	1.29
					PROF					
β_1	4.72	-4.49	3.44	5.40	2.95	4.65	0.38	2.44	-1.60	5.85
(β_1^l, β_1^u)	(0.46,8.98)	(-15.44,6.45)	(-1.24,8.12)	(1.42,9.38)	(-2.56,8.47)	(0.11,9.18)	(-6.85,7.60)	(-8.36,13.24)	(-10.77,7.57)	(-3.72,15.42)
R^2	1.75	0.27	0.59	2.13	0.26	1.32	0.00	0.06	0.03	0.51
					INV					
β_1	4.03	0.31	3.01	5.27	2.94	4.66	1.52	-2.55	-2.52	-1.26
(β_1^l, β_1^u)	(1.50,6.57)	(-8.96,9.58)	(-0.75,6.76)	(2.38,8.16)	(-2.07,7.95)	(1.66,7.65)	(-3.95,7.00)	(-9.43,4.34)	(-8.61,3.57)	(-8.27,5.75)
R^2	4.25	0.00	1.01	5.85	0.46	4.53	0.08	0.18	0.13	0.04
					MOM					
β_1	1.41	-9.04	4.25	2.40	4.92	1.85	3.88	3.96	-1.31	-1.58
(β_1^l, β_1^u)	(-1.88,4.70)	(-25.93,7.85)	(-2.13,10.62)	(-1.74,6.54)	(-4.43,14.27)	(-2.56,6.25)	(-10.24,18.01)	(-12.25,20.17)	(-18.03,15.41)	(-17.78,14.62)
R^2	0.34	0.39	0.67	0.50	0.38	0.28	0.10	0.09	0.01	0.02
$NOS.$	282	282	282	282	282	282	282	282	282	282

Table 7: **Timing Strategies**
This table reports performance statistics for portfolios formed from sorts on neural network forecasts. I report the annualized average return (Avg. ret), annualized Sharpe ratio (Sharpe), the annualized certainty equivalent (Utility) in percentages for an agent with a mean-variance utility function, and a risk aversion parameter of 2, the CAPM and Fama-French 6 factor model alphas. The two timing strategies invests: $w_{t,h} = \frac{\tilde{r}_{t,h} - r^f_{t+1}}{\gamma \sigma^2(\tilde{r}_{1:t-1,h})}$ in a risky security and $1 - w_{t,h}$ in the risk-free asset. I fix $\sigma(\tilde{r}_{1:t-1,h})$ at $\frac{0.15}{\sqrt{12}}$ to make the results comparable across forecasts. Panel A reports results for a strategy where the expected return on the value-weighted market ($\tilde{r}_{t,h}$) is computed as a function of different forecasts ($h \in \{1, 2, 3, 13, 25, 37, 49, 61, 121\}$) from the neural network model. We restrict the cross-section to the 500 largest stocks. The buy and hold strategy buys all 500 stocks in this restricted cross-section for each month t. Panel B reports results for a strategy where the expected return on an equally-weighted portfolio of the five long-short portfolios (value-weighted) formed from the characteristics in the Fama and French (2018) model. The buy and hold strategy buys all five portfolios in each month t whereas the timing strategies lever up and down this portfolio based on expected next month returns. The sample period is 1995 to 2018.

	Utility	Avg. ret	Sharpe	α_{capm}	α_{FF6}
Panel A: Market					
Buy & hold	8.18	**10.39**	0.70	0.23	0.04
Rotation (h)					
1	12.86	**17.21**	0.83	**8.41**	4.14
2	12.47	**15.79**	0.87	**8.53**	3.49
3	12.76	**15.70**	0.92	**8.95**	3.40
13	10.54	**12.33**	0.92	**7.31**	**4.64**
25	10.60	**12.61**	0.89	**7.43**	**4.77**
37	9.19	**10.39**	0.95	**5.76**	**4.53**
49	7.75	**8.83**	0.85	**4.50**	**3.97**
61	5.31	**6.66**	0.57	3.01	2.03
91	7.04	**8.15**	0.77	**3.42**	**3.37**
121	4.77	**6.86**	0.47	2.97	4.12
Panel B: Equally-weighted portfolio of long-short portfolios					
Buy & hold	3.02	**3.88**	0.42	**5.73**	1.70
Rotation (h)					
1	5.04	**6.72**	0.52	**9.05**	**5.28**
2	4.78	**6.22**	0.52	**8.05**	**4.48**
3	4.40	**5.78**	0.49	**7.57**	**3.90**
13	3.48	**3.82**	0.65	**4.79**	**2.76**
25	2.78	**2.92**	0.79	**3.49**	**2.39**
37	2.33	**2.40**	0.95	**2.78**	**1.94**
49	2.39	**2.44**	1.12	**2.79**	**1.91**
61	2.39	**2.43**	1.25	**2.68**	**2.07**
91	2.14	**2.17**	1.13	**2.44**	**1.65**
121	1.62	**1.64**	1.13	**1.78**	**1.25**

Table 8: **Rotation strategies**
This table reports performance statistics, as in Table 8. The strategy in panel A goes long (short) the value-weighted portfolio of the top (bottom) 10% of stocks with the highest (lowest) forecasted return for each month t. I restrict the cross-section of stocks to the 500 largest market-cap firms at each time t. The buy and hold strategy buys all 500 stocks in this restricted cross-section for each month t. In Panel B, the strategy buys (sells) an equally-weighted portfolio of the two characteristic-sorted long-short portfolios with an expected return above (below) the median forecast for each month t. The buy and hold strategy buys all long-short portfolios in each month t. I restrict the characteristic-sorted long-short portfolios to the five Fama and French (2018) characteristics. To highlight the importance of timely conditioning information, we consider variants of the strategy where the expected month t stock return, $\mathbb{E}_{t-h}[R_t]$, comes from different horizons, $h \in \{1, 2, 3, 13, 25, 37, 49, 61, 121\}$. The sample period is 1995 to 2018.

	Utility	Avg. ret	Sharpe	α_{capm}	α_{FF6}
			Panel A: Stocks		
Buy & hold	8.18	**10.39**	0.70	0.23	0.04
Rotation (h)					
1	11.73	**15.82**	0.78	**21.19**	**11.05**
2	8.74	**13.89**	0.61	**20.01**	**8.37**
3	3.98	**9.15**	0.40	**15.27**	3.46
13	0.26	6.26	0.26	**11.22**	0.07
25	-0.76	3.69	0.18	7.75	0.33
37	-1.60	0.45	0.03	2.02	-0.04
49	-1.43	0.62	0.04	2.08	1.80
61	0.85	2.74	0.20	4.99	2.46
91	-0.57	1.50	0.10	2.89	1.63
121	-0.14	1.43	0.11	0.96	0.17
			Panel B: Long-short portfolios		
Buy & hold	3.02	**3.88**	0.42	**5.73**	1.70
Rotation (h)					
1	10.45	**17.06**	0.66	**24.51**	**9.94**
2	12.77	**19.41**	0.75	**26.36**	**12.47**
3	9.80	**16.65**	0.64	**23.88**	**10.15**
13	3.01	8.96	0.37	**14.27**	4.20
25	2.04	7.62	0.32	**11.90**	5.79
37	-5.75	-1.06	-0.05	0.25	-1.46
49	-6.33	-1.27	-0.06	0.65	-0.91
61	-6.85	-2.50	-0.12	-0.62	-3.48
91	-8.01	-3.80	-0.19	-2.48	-4.38
121	-7.00	-3.22	-0.17	-2.40	-2.96

Internet Appendix

A Robustness Checks

In this section, I consider alternative benchmarks and additional checks of the robustness of the results in the paper.

A.1 Alternative benchmarks

To highlight the general performance of the neural network forecasts, I consider two additional benchmarks from the literature. These benchmark models forecast monthly firm returns as a function of historical individual firm returns. The first model predicts the time $t + h$ return of firm i as the average firm i return computed over a five-year rolling window; the five-year rolling window firm average return benchmark. The second model predicts the time $t + h$ return of firm i as the average firm i return from the start of the sample up to time t; the expanding window firm average return benchmark.

[Insert Table IA.1 about here]

Panels A and B of Table IA.1 report results for the alternative benchmarks described above. The R^2_{OOS} is positive for all future periods and statistically significant for both models. Prediction accuracy is higher for short-run periods and falls with the horizon. The positive R^2_{OOS} of the neural network forecasts against these benchmarks are much larger than the estimates against the zero prediction benchmark because firm average returns are very noisy estimates of expected firm returns.

Comparing the R^2_{OOS} estimates of the simple model to the economically restricted model across horizons and benchmarks, it is evident that economic restrictions generally improve the forecasts. Most of this improvement is concentrated in the short-run. In predicting next month's return, the economically restricted model has an R^2_{OOS} of about 2.18% against the five-year rolling window benchmark, whereas the simple model has an R^2_{OOS} of about 1.76%. In predicting returns ten-years into the future, the economically restricted model has an R^2_{OOS} of 0.63% against the expanding window firm average return benchmark, and the simple model has an R^2_{OOS} of about 0.64%.

Table IA.1: **Alternative benchmarks**
This table presents out-of-sample R^2 (R^2_{OOS}) estimates for firm-level return forecasts generated by neural network models at time t for month $t+h$ where $h \in \{1, 2, 3, 13, 25, 37, 49, 61, 121\}$. Panel A reports the R^2_{OOS} for a neural network forecasting model that adheres strictly to economic theory as defined in Equation (2), and panel B reports results for a simpler neural network forecasting model with no economic restrictions as defined in Eq. (8). I present the R^2_{OOS} for the neural network model against two alternative models; a five-year rolling window firm average return model, and an expanding window firm average return model. Bold fonts indicate horizons for which the forecasts from the neural network model and alternative model are statistically different at the 5% level or better using the Clark-West (2007) test. The sample period is from January 1995 to December 2018.

h	1	2	3	13	37	61	91	121
Panel A: Economically restricted model								
Five-Year Rolling Window	**2.18**	**1.78**	**1.71**	**1.75**	**1.34**	**1.54**	**1.54**	**1.58**
Expanding Window	**1.26**	**0.81**	**0.72**	**0.74**	**0.58**	**0.59**	**0.62**	**0.63**
Panel B: Simple model								
Five-Year Rolling Window	**1.76**	**1.64**	**1.62**	**1.58**	**1.28**	**1.51**	**1.42**	**1.38**
Expanding Window	**0.91**	**0.75**	**0.72**	**0.70**	**0.59**	**0.68**	**0.65**	**0.64**

B Variable Construction

Table B.1: **Characteristics**
This table lists the characteristics used in this paper. For each characteristic, we present the associated acronym, the original source and the definition of the characteristic.

Acronym	Author(s)	Definition
A2ME	Bhandari (1988)	Total assets (at) over market capitalization (prc x shrout)
AC	Sloan (1996)	Change in operating working capital per split adjusted share from fiscal year $t-2$ to $t-1$ to book equity, (BEME), per share. Operating working capital per split-adjusted share is defined as current assets (ACT) minus cash and short-term investments (che) minus current liabilites (lct) minus debt in current liabilities (dlc) minus income taxes payable (txp).
AOA	Bandyopadhyay et al. (2010)	Absolute value of OA
AT	Gandhi and Lustig (2015)	Total assets (at)
ATO	Soliman (2008)	Net sales (sales) over lagged net operating assets. Net operating assets is the difference between operating assets and operating liabilities. Operating Assets is total assets (at) minus cash and short-term investments (che) minus investments and other advances (ivao). Operating Liabilities is total assets (at) minus debt in current liabilities (dlc) minus long-debt debt (dltt) minus minority interest (mib) minus preferred stock (pstk) minus common equity (ceq).
BEME (BM)	Davis et al. (2000)	Book equity to market equity. Book equity is shareholders' equity (seq), (if missing, common equity (ceq) plus preferred stock (pstk), if missing, total assets (at) minus total liabilities (lt)), plus deferred taxes and investment tax credit (txditc) minus preferred stock (pstrkrv), (if missing, liquidation value, (pstkl), if missing par value (pstk)). Market value of equity is shares outstanding (shrout) times price (prc).
BETA	Frazzini and Pedersen (2014)	The product of the correlation between stock excess returns and market excess returns and the ratio of volatilities. Ratio of volatilities is the volatility of stock excess returns to the volatility of market excess returns. Volatility is computed from the standard deviations of daily log excess returns over a one-year horizon requiring at least 120 observations. Correlations is computed using overlapping three-day log excess returns over a five-year period requiring at least 750 non-missing observations.
$BETA_d$	Lewellen and Nagel (2006)	The sum of the regression coefficients of daily excess returns on market excess returns and the lag of market excess returns.
C2A	Palazzo (2012)	Cash and short-term investments (che) to total assets (at).
C2D	Ou and Penman (1989)	Cashflow to debt. Cashflow is the sum of income and extraordinary items (ib) and depreciation and amortization (dp). And debt is to total liabilities (lt).

Continued

Acronym	Author(s)	Definition
CAT	Haugen and Baker (1996)	Sales (sale) to lagged total assets (at).
CF2P	Desai et al. (2004)	Cashflow to book value of equity is the ratio of net income (ni), depreciation and amortization (dp) less change in working capital (wcapch) and capital expenditure (capx) over the book-value of equity (BEME).
CL2HG	George and Hwang (2004)	ratio of last month closing price to the max closing price over the last 52 weeks.
D2A	Gorodnichenko and Weber (2016)	Depreciation and amortization (dp) to total assets (at).
D2P	Litzenberger and Ramaswamy (1979)	Debt to price. Debt is long-term debt (dltt) plus debt in current liabilities (dlc). Market capitalization is the product of shares outstanding (shrout) and price (prc).
dCEQ	Richardson et al. (2005)	Annual % change in book value of equity (ceq).
dGS	Abarbanell and Bushee (1997)	% change in gross margin minus % change in sales (sale). Gross margin is the difference in sales (sale) and cost of goods sold (cogs).
dPIA	Lyandres et al. (2008)	Change in property, plants and equipment (ppegt) and inventory (invt) over lagged total assets (at).
dSO	Fama and French (2008)	Log change in the product of shares outstanding (csho) and the adjustment factor (ajex).
dSOUT	Pontiff and Woodgate (2008)	Annual % change in shares outstanding (shrout).
DP	Litzenberger and Ramaswamy (1979)	Sum of monthly dividend over the last 12 months to last month's price (prc).
DTO	Garfinkel (2009)	Daily volume (vol) to shares outstanding (shrout) minus the daily market turnover and detrended by the 180 trading day median. To address the double counting of volume for NASDAQ securities, we follow Anderson and Dyl (2005) and scale down the volume of NASDAQ securities by 50% before and by 38% after 1997.
E2P	Basu (1983)	Income before extraordinary items (ib) to market capitalization (prc x shrout).
EPS	Basu (1977)	Income before extraordinary items (ib) to shares outstanding (shrout).
FC2Y	D'Acunto et al. (2018)	Ratio of selling, general and administrative expenses (xsgs), research and development expenses (xrd) and advertising expenses (xad) to net sales.
FCF	Hou et al. (2011)	Ratio of net income (ni), depreciation and amortization (dp) less change in working capital (wcapch) and capital expenditure (capx) over book value of equity as defined in BEME.
I2A (INV)	Cooper et al. (2008)	Annual % change in total assets (at).
IDIOV	Ang et al. (2006)	Standard deviation of the residuals from a regression of excess returns on the Fama and French (1993) three-factor model.
IPM		Pre-tax income (pi) over sales (sale).
IVC	Thomas and Zhang (2002)	Annual change in inventories (invt) in the last two fiscal years over the average total assets (at) over the last two fiscal years.
LEV	Lewellen (2015)	long-term debt (dltt) plus current liabilities (dlc) over the sum of long term debt (dltt), debt in current liabilities (dlc) and stockholders equity (seq).

Continued

Acronym	Author(s)	Definition
MAXRET	Bali et al. (2011)	Last months stock price (prc) over previous 52 week max price.
NOA	Hirshleifer et al. (2004)	Operating assets minus operating liabilities to lagged total assets (at). Operating assets is total assets (at) minus cash and short term investments (che) minus investment and other advances (ivao). Operating liabilities is total assets (at) minus debt in current liabilities (dlc) minus long-term debt (dltt) minus minority interest (mib) minus preferred stock (pstk) minus common equity (ceq).
OA	Sloan (1996)	Changes in non-cash working capital minus depreciation (dp), all scaled by lagged total assets (at). Changes in non-cash working capital is difference in current assets (act) minus difference in cash and short-term investments (che) minus difference in current liabilities (lct) minus difference in debt in current liabilities (dlc) minus difference in taxes payable (txp).
OL	Novy-Marx (2011)	Sum of cost of goods sold (cogs) and selling, general and administrative expense (xsga) over total assets (at).
PCM	Gorodnichenko and Weber (2016)	Net sales (sale) minus cost of goods sold (cogs) all scaled by net sales (sale).
PM	Soliman (2008)	Operating Income after depreciation (oiadp) to sales (sale).
PROF	Ball et al. (2015)	Gross profitability (gp) over book equity as defined in $BEME$.
Q		Total assets (at) plus market value of equity (shrout x prc) minus cash and short-term investments (ceq) minus deferred taxes (txdb) scaled by total assets (at).
$R_{12,2}$	Fama and French (1996)	Cumulative return from 12 months to 2 months ago.
$R_{12,7}$	Novy-Marx (2012)	Cumulative return from 12 months to 7 months ago.
$R_{2,1}$	Jegadeesh (1990)	Lagged one month return.
$R_{36,13}$	De Bondt and Thaler (1985)	Cumulative return from 36 months to 13 months ago.
$R_{6,2}$	Jegadeesh and Titman (1993)	Cumulative return from 6 months to 2 months ago.
RETVOL	Ang et al. (2006)	Standard deviation of residuals from a regression of excess returns on a constant using one month of daily data. We require there to be at least 15 non-missing observations.
RNA	Soliman (2008)	Operating income after depreciation (oiadp) scaled by lagged net operating assets. Net operating assets is operating assets minus operating liabilities. Operating assets is total assets (at) minus cash and short term investments (che) minus investment and other advances (ivao). Operating liabilities is total assets (at) minus debt in current liabilities (dlc) minus long-term debt (dltt) minus minority interest (mib) minus preferred stock (pstk) minus common equity (ceq).
ROA	Balakrishnan et al. (2010)	Income before extraordinary items (ib) to lagged total assets (at).

Continued

Acronym	Author(s)	Definition
ROC	Chandrashekar and Rao (2009)	Market value of equity (shrout x prc) plus long-term debt (dltt) minus total assets (at) all over cash and short-term investments (che).
ROE	Haugen and Baker (1996)	Income before extraordinary items (ib) to lagged book-value of equity.
ROIC	Brown and Rowe (2007)	Earnings before interest and taxes (ebit) less non-operating income (nopi) to the sum of common equity (ceq), total liabilities (lt), and cash and short-term investments (che).
S2P	Lewellen (2015)	Net sales (sale) to market capitalization (shrout x prc).
sdDVOL	Chordia et al. (2001)	Standard deviation of residuals from a regression of daily volume (vol) on a constant. Use one month of daily data requiring at-least 15 non-missing observations.
sdTURN	Chordia et al. (2001)	Standard deviation of residuals from a regression of daily turnover on a constant. Turnover is volume (vol) times shares outstanding (shrout). Use one month of daily data requiring at-least 15 non-missing observations.
SG	Lakonishok et al. (1994)	% growth rate in sales (sale).
SGNA		Selling, general and administrative expenses (XSGA) to net sales (sale).
SIZE	Fama and French (1992)	Price (prc) times shares outstanding (shrout) .
SPREAD	Chung and Zhang (2014)	Average daily bid-ask spread in the previous month.
SUV	Garfinkel (2009)	Difference between actual volume and predicted volume. Predicted volume is from a regression of previous month's daily volume on a constant and the absolute values of positive and negative previous month's returns. Unexplained volume is standardized by the standard deviation of the residuals from the regression.
TNOVR	Datar et al. (1998)	Volume (vol) over shares outstanding (shrout).

C Details of Algorithms

Denote the penalized loss function of the neural network model as $\mathcal{L}(\theta;\cdot)$. The standard method for finding the optimal parameters (θ^*) that minimizes $\mathcal{L}(\theta;\cdot)$ is stochastic gradient descent (SGD), Goodfellow et al. (2016). Minimizing this loss function with SGD is slow and inefficient because it is a first-order optimization procedure. In this study we use a recent variant of SGD called AdaBound, Luo et al. (2019) which uses second-order information and has theoretical convergence guarantees.

We initialize θ by sampling θ_0 from $\mathcal{N}(0, n_h^{-1})$ where n_h is the size of the input vector of layer h. A single training step t, consists of a randomly sampling 10000 firm-month observations from the training set and running Algorithm 1.

Algorithm 1: AdaBound Variant of Stochastic Gradient Descent

Initialization : $\theta_0 \sim \mathcal{N}(0, n_h^{-1})$. $\alpha = 10^{-1}$. $m_0 = 0$. $v_o = 0$;
while θ_t *not coverge* **do**
  $t \leftarrow t + 1$;
  $g_t = \nabla_\theta \mathcal{L}_t(\theta_{t-1};\cdot)$;
  $m_t = \beta_1 m_{t-1} + (1 - \beta_{1,t}) g_t$;
  $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$;
  $V_t = diag(v_t)$;
  $\hat{\eta}_t = Clip(\alpha/\sqrt{V_t}, \eta_l(t), \eta_u(t))$;
  $\eta_t = \hat{\eta}_t/\sqrt{t}$;
  $\theta_t = \prod_{diag(\eta_t^{-1})}(\theta_{t-1} - \eta_t \odot m_t)$;
end
Result: Final parameter estimate $\theta_{\tilde{t}}$;

where $Clip(.)$ is a clipping function that bounds the learning rate (α) to the interval $[\eta_l, \eta_u]$.

Algorithm 2: Batch Normalization

Input : Values of x for each activation over a single batch $B = \{x_1, x_2, x_3, .., x_N\}$;
$\mu_B \leftarrow \frac{1}{N}\sum_{i=1}^{N} x_i$;
$\sigma_B^2 \leftarrow \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_B)^2$;
$\hat{x_i} \leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}}$;
$y_i \leftarrow \gamma x_i + \beta = BN_{\gamma,\beta(x_i)}$;
Result: $y_i = BN_{\gamma,\beta(x_i):i=1,2,3,...,N}$;

Table C.1: **Hyper-parameters**
This table reports all hyper-parameters in the paper and the range in which they were chosen via the validation set.

Name	Range
l_1 Penalty (λ_1)	$(10^{-1}, 10^{-6})$
Learning rate	$(10^{-2}, 10^{-4})$
Batch size	10000
Epochs	100
Adabound parameters	Default (amsbound=True)
Patience	5
Ensemble	100

D Decomposing the out-of-sample R^2

We start with the definition of the out-of-sample R^2 (R^2_{OOS});

$$R^2_{OOS} = 1 - \frac{\sum_{(t) \in oss}(R_t - R_{t,1})^2}{\sum_{(t) \in oss}(R_t - R_{t,2})^2} \tag{D.1}$$

We define the relative stock return forecast of some firm i at time t, as the time t return forecast for firm i minus the time t cross-sectional mean forecast.

Specifically, we re-define stock return forecast as follows:

$$r_{1,i,t} = \mu^{RR}_{1,i} + (N_t^{-1}) \sum_{i \in t} r_{1,i,t} + \tilde{r}^{RR}_{1,i,t} \tag{D.2}$$

$$r_{2,i,t} = \mu^{RR}_{2,i} + (N_t^{-1}) \sum_{i \in t} r_{1,i,t} + \tilde{r}^{RR}_{2,i,t} \tag{D.3}$$

This definition allows model one to improve upon forecast from model two along three possible dimensions. First, model one can improve the forecasts of model two with respect to predicting time-series variation in relative stock returns; $\tilde{r}^{RR}_{1,i,t}$ vs. $\tilde{r}^{RR}_{2,i,t}$. Second, the improvement maybe in predicting better the unconditional relative stock returns; $\mu^{RR}_{1,i}$ vs. $\mu^{RR}_{2,i}$. Finally, the improvement maybe in predicting the cross-sectional mean better; $(N_t^{-1}) \sum_{i \in t} r_{1,i,t}$ vs. $(N_t^{-1}) \sum_{i \in t} r_{1,i,t}$. The decomposition therefore allows us to directly pin-point along which dimension one forecaster is doing better than another.

We claim that:

$$R^2_{OOS} = R^2_{TS} + R^2_{UN} + R^2_{CS} \tag{D.4}$$

where R^2_{TS}; captures the ability of model 1 to improve on forecasts from model 2 with respect to predicting time series variation in relative stock returns, R^2_{UN}; captures the ability of model 1 to improve on forecasts from model 2 with respect to predicting the unconditional relative stock returns and R_{CS}; cross-sectional mean dimension (time series plus unconditional cross-sectional market return).

We begin by expanding R^2_{TS}:

$$R^2_{TS} = 1 - \frac{\sum r_{i,t} - \mu_{2,i,t,RR} - \tilde{r}_{1,i,t,RR} - \sum_t r_{2,i,t}{}^2}{\sum (r_{i,t} - r_{2,i,t})^2} \tag{D.5}$$

$$A_1 = r_{i,t}; \quad B_2 = \mu_{2,i,t,RR}; \quad C_1 = \tilde{r}_{1,i,t,RR}; \quad D_2 = \sum r_{2,i,t} \tag{D.6}$$

$$R^2_{TS} = 1 - \frac{\sum (A_1 - B_2 - C_1 - D_2)^2}{\sum (r_{i,t} - r_{2,i,t})^2} \tag{D.7}$$

$$R^2_{TS} = 1 - \frac{\sum (A_1^2 - 2A_1B_2 - 2A_1C_1 - 2A_1D_2 + B_2^2 + 2B_2C_1 + 2B_2D_2 + C_1^2 + 2C_1D_2 + D_2^2)}{\sum (r_{i,t} - r_{2,i,t})^2} \tag{D.8}$$

Expanding R^2_{UN}:

$$R^2_{UN} = 1 - \frac{\sum r_{i,t} - \mu_{1,i,t,RR} - \tilde{r}_{2,i,t,RR} - \sum_t r_{2,i,t}{}^2}{\sum (r_{i,t} - r_{2,i,t})^2} \tag{D.9}$$

$$A_1 = r_{i,t}; \quad B_1 = \mu_{1,i,t,RR}; \quad C_2 = \tilde{r}_{2,i,t,RR}; \quad D_2 = \sum r_{2,i,t} \tag{D.10}$$

$$R^2_{UN} = 1 - \frac{\sum (A_1 - B_1 - C_2 - D_2)^2}{\sum (r_{i,t} - r_{2,i,t})^2} \tag{D.11}$$

$$R^2_{UN} = 1 - \frac{\sum (A_1^2 - 2A_1B_1 - 2A_1C_2 - 2A_1D_2 + B_1^2 + 2B_2C_2 + 2B_1D_2 + C_2^2 + 2C_2D_2 + D_2^2)}{\sum (r_{i,t} - r_{2,i,t})^2} \tag{D.12}$$

Expanding R^2_{CS}:

$$R^2_{CS} = 1 - \frac{\sum (r_{i,t} - \mu_{2,i,t,RR} - \tilde{r}_{2,i,t,RR} - \sum_t r_{1,i,t})^2}{\sum (r_{i,t} - r_{2,i,t})^2} \tag{D.13}$$

$$A_1 = r_{i,t}; \quad B_2 = \mu_{2,i,t,RR}; \quad C_2 = \tilde{r}_{2,i,t,RR}; \quad D_1 = \sum r_{1,i,t} \tag{D.14}$$

$$R^2_{CS} = 1 - \frac{\sum (A_1 - B_2 - C_2 - D_1)^2}{\sum (r_{i,t} - r_{2,i,t})^2} \tag{D.15}$$

$$R^2_{CS} = 1 - \frac{\sum (A_1^2 - 2A_1B_2 - 2A_1C_2 - 2A_1D_1 + B_2^2 + 2B_2C_2 + 2B_2D_1 + C_2^2 + 2C_2D_1 + D_1^2)}{\sum (r_{i,t} - r_{2,i,t})^2} \tag{D.16}$$

From the expansions above we have:

$$R^2_{TS} + R^2_{UN} + R^2_{CS} = ...$$
$$3 - \frac{\sum(A_1^2 - 2A_1B_2 - 2A_1C_1 - 2A_1D_2 + B_2^2 + 2B_2C_1 + 2B_2D_2 + C_1^2 + 2C_1D_2 + D_2^2)}{\sum\left(r_{i,t} - r_{2,i,t}\right)^2}$$
$$- \frac{\sum\left(A_1^2 - 2\ A_1B_1 - \ 2A_1C_2\ -\ 2A_1D_2\ +\ B_1^2\ +\ 2\ B_1C_2 + \ 2B_1D_2\ +\ C_2^2\ +\ 2C_2D_2\ +\ D_2^2\right)}{\sum\left(r_{i,t} - r_{2,i,t}\right)^2}$$
$$- \frac{\sum\left(A_1^2 - 2\ A_1B_2 - \ 2A_1C_2\ -\ 2A_1D_1\ +\ B_2^2\ +\ 2\ B_2C_2 + \ 2B_2D_1\ +\ C_2^2\ +\ 2C_2D_1\ +\ D_1^2\right)}{\sum\left(r_{i,t} - r_{2,i,t}\right)^2}$$

Cross-products with C, where subscripts are different fall out because we define the time varying relative return forecasts to have mean zero. We further assume; $\tilde{r}_{2,i,t,RR} \perp \sum_t r_{1,i,t}$ and $\tilde{r}_{1,i,t,RR} \perp \sum_t r_{2,i,t}$.

$$R^2_{TS} + R^2_{UN} + R^2_{CS} = ...$$
$$3 - \frac{\sum\left(A_1^2 - \ 2A_1C_1 + \ C_1^2 - 2\ A_1B_1 + \ B_1^2 - \ 2A_1D_1 + \ D_1^2\right)}{\sum\left(r_{i,t} - r_{2,i,t}\right)^2}$$
$$- \frac{\sum\left(A_1^2 + \ 2C_2D_2 + \ D_2^2 + \ B_2^2 + \ 2\ B_2C_2 + \ C_2^2 - \ 2A_1C_2 - 2\ A_1B_2 - \ 2A_1D_2\right)}{\sum\left(r_{i,t} - r_{2,i,t}\right)^2}$$
$$- \frac{\sum\left(A_1^2 + \ B_2^2 + \ 2B_2D_2 + \ D_2^2 + \ C_2^2 - \ 2A_1C_2 - \ 2A_1D_2 - 2\ A_1B_2\right)}{\sum\left(r_{i,t} - r_{2,i,t}\right)^2}$$

Add $(2B_1C_1 - 2B_1C_1)$, $(2B_1D_1 - 2B_1D_1)$, and $(2C_1D_1 - 2C_1D_1)$ to the first fraction, $(2A_1D_2 - 2A_1D_2)$ to the second fraction and $(2B_2C_2 - 2B_2C_2)$ and $(2C_2D_2 - 2C_2D_2)$ to the last fraction.

$$R^2_{TS} + R^2_{UN} + R^2_{CS} = ...$$
$$3 - \frac{\sum(A_1^2 - 2A_1C_1 + C_1^2 - 2A_1B_1 + B_1^2 - 2A_1D_1 + D_1^2 + (2B_1C_1 - 2B_1C_1) + (2B_1D_1 - 2B_1D_1) + (2C_1D_1 - 2C_1D_1))}{\sum\left(r_{i,t} - r_{2,i,t}\right)^2}$$
$$- \frac{\sum(A_1^2 + 2C_2D_2 + D_2^2 + B_2^2 + 2B_2C_2 + C_2^2 - 2A_1C_2 - 2A_1B_2 - 2A_1D_2) + (2B_2D_2 - 2B_2D_2)}{\sum\left(r_{i,t} - r_{2,i,t}\right)^2}$$
$$- \frac{\sum(A_1^2 + B_2^2 + 2B_2D_2 + D_2^2 + C_2^2 - 2A_1C_2 - 2A_1D_2 - 2A_1B_2) + (2B_2C_2 - 2B_2C_2) + (2C_2D_2 - 2C_2D_2)}{\sum\left(r_{i,t} - r_{2,i,t}\right)^2}$$

$$R^2_{TS} + R^2_{UN} + R^2_{CS} = ...$$
$$3 - \frac{\sum(A_1^2 - 2A_1C_1 + C_1^2 - 2A_1B_1 + B_1^2 - 2A_1D_1 + D_1^2 + 2B_1C_1 + 2B_1D_1 + 2C_1D_1)}{\sum(r_{i,t} - r_{2,i,t})^2}$$
$$- \frac{\sum(A_1^2 + 2C_2D_2 + D_2^2 + B_2^2 + 2B_2C_2 + C_2^2 - 2A_1C_2 - 2A_1B_2 - 2A_1D_2 + 2B_2D_2)}{\sum(r_{i,t} - r_{2,i,t})^2}$$
$$- \frac{\sum(A_1^2 + B_2^2 + 2B_2D_2 + D_2^2 + C_2^2 - 2A_1C_2 - 2A_1D_2 - 2A_1B_2 + 2B_2C_2 + 2C_2D_2)}{\sum(r_{i,t} - r_{2,i,t})^2}$$
$$- \frac{\sum(2B_1C_1 + 2B_1D_1 + 2C_1D_1 + 2B_2D_2 + 2B_2D_2 + 2B_2C_2 + 2C_2D_2)}{\sum(r_{i,t} - r_{2,i,t})^2}$$

All cross-terms with C fall out:

$$R^2_{TS} + R^2_{UN} + R^2_{CS} = ...$$
$$3 - \frac{\sum(A_1 - B_1 - C_1 - D_1)^2}{\sum(r_{i,t} - r_{2,i,t})^2} - \frac{\sum(A_1 - B_2 - C_2 - D_2)^2}{\sum(r_{i,t} - r_{2,i,t})^2} - \frac{\sum(A_1 - B_2 - C_2 - D_2)^2}{\sum(r_{i,t} - r_{2,i,t})^2} - \frac{\sum(2B_1D_1 + 4B_2D_2)}{\sum(r_{i,t} - r_{2,i,t})^2} \tag{D.17}$$

$$R^2_{TS} + R^2_{UN} + R^2_{CS} = ...$$
$$3 - \frac{\sum(A_1 - B_1 - C_1 - D_1)^2}{\sum(r_{i,t} - r_{2,i,t})^2} - \frac{\sum(A_1 - B_2 - C_2 - D_2)^2}{\sum(r_{i,t} - r_{2,i,t})^2} - \frac{\sum(A_1 - B_2 - C_2 - D_2)^2}{\sum(r_{i,t} - r_{2,i,t})^2} \tag{D.18}$$

Where $r_{2,i,t} = B_2 + C_2 + D_2$ and $r_{1,i,t} = B_1 + C_1 + D_1$.

$$R^2_{TS} + R^2_{UN} + R^2_{CS} = 3 - \frac{\sum(A_1 - B_1 - C_1 - D_1)^2}{\sum(r_{i,t} - r_{2,i,t})^2} - 1 - 1$$

Putting it all together we have:

$$R^2 = R^2_{TS} + R^2_{UN} + R^2_{CS} = 1 - \frac{\sum(r_{i,t} - r_{1,i,t})^2}{\sum(r_{i,t} - r_{2,i,t})^2}$$

where $\tilde{r}_{t,i,1}$ and $\tilde{r}_{t,i,2}$ have mean zero. It then follows that re-writing $R_{t,1}$ as $(\mu_{i,1} - \mu_{i,1}) + \tilde{r}_{t,i,1} + \mu_{i,2}$, and comparing this to forecasts from model 2 allows us to focus on the time-series variations in the forecasts. This is because, this specific re-write forces model 1 and model 2 to be equal in their ability to explain the unconditional stock return (all equal to mu_2), and the difference in forecasting ability now comes from $\tilde{r}_{t,i,1}$ versus $\tilde{r}_{t,i,2}$. Re-writing model 1 forecasts as $\mu_{i,1} + (\tilde{r}_{t,i,1} - \tilde{r}_{t,i,1}) + \tilde{r}_{t,i,2}$ forces the two models to match in

explaining the time-series variation in returns ($\tilde{r}_{t,i,2}$), and differ in their ability to explain the unconditional stock return ($\mu_{i,2}$ vs $\mu_{i,1}$) over the sample.