Machine learning and return predictability across firms, time and portfolios☆

Fahiz Baba-Yara*
Nova SBE

September 25, 2020

Abstract

Using an information set of firm-level characteristics and aggregate-variables, we show that stock returns as far out as ten years are predictable using a neural network forecasting model. The strength of this predictability is highest in the short-run and falls with horizon. We disentangle the nature of these forecasts and show that most of the predictability we uncover comes from predicting the dominant factor in the pool of stocks; the equally weighted market return. We find that time-series variation in relative stock returns is only predictable in the short-run, consistent with the fact that longer-term discount rates do not vary much across firms. Finally, we show that a neural network model that closely adheres to economic theory generates forecasts that more robustly predict returns to the aggregate market and long-short characteristic sorted portfolios.

Keywords: Return Predictability, Long-run 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 Giorgio Ottonello, Fernando Anjos, Irem Demirci, Miguel Ferreira, Melissa Prado, 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 my own.

1 Introduction

In his presidential address, Cochrane (2011) laments that "the world is once again descending into chaos. Expected return strategies have emerged that do not correspond to market, value, and size betas." Since then, the zoo of return predictors has only gotten larger, currently numbering over 400 (Hou et al. (2020)). Using machine learning methods (ML), Gu et al. (2020a) show that one can reduce this high-dimensional vector of return predictors to a reasonably good proxy for expected stock return. Among the ML models they consider, neural networks consistently outperform in predicting returns across securities.

This paper shows that economic theory still has an integral role to play in this emerging field of financial machine learning. First, I show that a neural network model that adheres to economic theory outperforms a much simpler neural network model in predicting returns across multiple horizons. Whereas the simple neural network model can explain about 0.58% of the variation in next month's stock returns, the model that adheres more to economic theory can explain about 0.99%—a close to two-fold increase.

Second, I show that investors who employ the economically restricted neural network forecast enjoy large and robust economic gains. Specifically, a long-short portfolio that buys (sells) the 10% highest (lowest) expected return stocks, among the most liquid stocks, using these forecasts has an annualized average return of 15.82%, a Sharpe ratio of 0.78, and certainty equivalent of 11.73, using forecasts for the next month[1]. This result stands in contrast to Avramov et al. (2020), where the authors show that the simple neural network model extracts a non-trivial fraction of its perceived profitability from difficult to arbitrage stocks.

Third, I find that to produce forecasts that robustly generalize beyond the cross-section of stocks, restrictions implied by economic theory are crucial. The simple neural network model fails to robustly predict time-series variation in returns for the aggregate market and long-short portfolios. In contrast, the model grounded in economic theory produces forecasts that predict time-series variation in returns for 53 out of the 56 long-short portfolios I consider. Similarly, when aggregated to predict market returns, these

[1]This strategy is restricted to trading the most liquid stocks; the 500 largest market capitalized firms in the cross-section every. The sample period is January 1995 to December 2018.

forecasts predict time-series variation in market returns as far out as thirty-two months in the future.

A natural question one would ask at this point is,"Along which dimension does the economically restricted model help improve return forecasts?". To answer this question, I disentangle the nature of the observed predictability by decomposing stock forecasts into two components; explaining variations in a level factor (the equally-weighted market return) and a relative return factor (cross-sectional dispersion in returns across stocks)[2]. I find that the economically restricted model explains about 0.55% of the cross-sectional dispersion in returns across stocks whereas the simple model only explains 0.16%-a more than three-fold improvement.

Most papers in the financial machine learning literature study cross-section return predictability over the next month or at most the following year (cumulative return). Therefore, this paper is the first to study cross-section monthly stock return predictability as far as ten years into the future and documents new evidence on the cross-sectional properties of expected returns across horizons. I ask the question, "How predictable are future monthly stock returns conditional on an information set of firm characteristics and aggregate variables observed today?" The results are summarized as follows.

First, stock returns are highly predictable in the short-run, and predictability wanes with horizon. The economically restricted neural network model can explain about 0.99% of the variations in next month's returns. This 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, I find that the nature of short-run stock return predictability is very different from long-run stock return predictability. This follows from the fact that individual stock return predictability is measured within the pool of stocks; hence it is instructive to consider the inherent factor structure that underpins stock returns. Once I do this, I find that a large fraction of observed stock return predictability comes from predicting well, the market, or level factor in the pool of stocks. When predicting next month's return, about 40% of the variation explained comes from explaining the market factor.

[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, $r_{i,t}^{RR} = r_{i,t} - (N_t^{-1})\sum_{i\epsilon t} r_{i,t}$.

When forecasts pertain to periods beyond the following year, over 95% of the variation explained comes from explaining variations in the market factor. This result implies that the forecasts' ability to explain cross-sectional variation in returns is short-lived. Return forecasts for the following month explain about 0.71% of time-series variation in relative stock returns, and this falls to about 0.08% when predicting returns thirteen months into the future.

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 expected return estimates is scant. I provide evidence that forecasts from the restricted neural network model line up well with true expected stock returns. I find that a one percent return forecast on average predicts a 0.97 percentage point increase in next month relative stock returns[3]. Similar to the conclusions drawn from the out-of-sample R^2 analysis, I find that this estimate wanes with horizon. In predicting the relative stock return realized one-year into the future, I find this estimate falls to 0.79 and further down to 0.24 when predicting returns two years into the future.

The question we ask in this paper is both important and interesting. First, expected returns are analogous to discount rates (cost of capital) conditional on firm/project characteristics. Presently, we do not have very good estimates of these discount rates. The proposed framework in this paper allows one to use project characteristics observed today as a basis for coming up with reasonable and consistent discount rate estimates for future cash-flows.

Second, although investors' horizon is very different, it is still a fact that most investors would like to harvest the highest spread in returns in the cross-section that their constraints will allow for. Therefore, it follows that if cross-sectional dispersion in returns is most predictable in the following month, $t+1$, then investors will have to trade-off how much of this spread they are willing to give up when they choose to re-balance less frequently. Specifically, I show that a long-short portfolio, trading the most liquid stocks in the cross-section, that re-balances monthly with forecasts for time $t+1$ has a Sharpe ratio of 0.78. In contrast, a strategy that re-balances with forecasts for time $t+13$ has a

[3]Relative stock returns is defined as stock returns minus the cross-sectional mean and so captures cross-sectional dispersion in returns

Sharpe ratio of 026.

Third, Martin and Nagel (2019) consider a world where agents have to condition on thousands of potentially relevant variables to forecast returns. A world not too dissimilar from our own. If agents are uncertain about how exactly cash-flows relate to these predictors, then a factor zoo will naturally emerge. Agents in this world need a learning-algorithm to help them reduce the uncertainty about coefficients over time. In this paper, I consider a neural network as an alternative learning algorithm to the Bayesian updating framework considered by the authors. Almost all machine-learning methods are equally efficient in a world with infinity data. However, in small samples, some algorithms are more efficient and less biased than others given the dataset one has and the problem to be solved[4]. And as Gu et al. (2020a) show, neural networks are one of the best approximators we have of expected returns given this particular history of returns.

This work builds on the emerging literature in economics and finance using machine learning methods to answer economic questions that are primarily prediction in nature. 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 a nonparametric method to study which subset of 62 characteristics they consider provide incremental information about the cross-section of expected returns. 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. Kelly et al. (2019) use dimension reduction methods to show that a latent factor model that employs time-varying loadings significantly improves upon the pricing power of static loading models. Haddad et al. (2020) extract the dominant principal components (PC) in the cross-section of characteristic sorted portfolios and show that the aggregated forecasts formed from these PCs can predict returns to about 36 of the 50 long-short portfolios they consider.

This work primarily extends the literature on stock return predictability. Lewellen (2015) studies expected returns across stocks as a linear function of stock level character-

[4]This phenomena is known as "the no know-free lunch" theorem in machine learning, see Wolpert (1996)

istics and finds evidence that the forecasts generated by the linear model explain some variation in returns. Gu, Kelly, and Xiu (2020b) further show that allowing for nonlinearities in this framework enhances the ability of the known predictors in the literature to predict better returns. Specifically, the authors show that firm-level characteristics can be combined with macroeconomic variables using different machine learning methods to predict returns better. I build on this framework by showing that a neural network architecture design that imposes much stricter economic restrictions further improves these forecasts. I also show that the standard set of cross-sectional predictors used in the literature are not only informative of stock return realizations for the next month but extends much further out into the future. Finally, I find that cross-sectional relative stock return predictability is short-lived such that after thirteen months, the forecasts from a machine learning model do no better than a zero forecast in discriminating between high and low expected return firms; a result consistent with longer-run discount rates coverge across firms (see Keloharju, Linnainmaa, and Nyberg (2019)).

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-sorted 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 the aggregate long-short portfolio forecasts from a neural network model can robustly predict time-series variation in 53 of the 56 long-short portfolios I consider. I also show that imposing economic restrictions on the corresponding machine learning model is important in producing forecasts that generalize beyond individual stocks. This generalization is what helps the aggregated forecasts predict variations in returns of these long-short portfolios.

2 Empirical Framework and Data

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

2.1 Factor Model

We assume that excess 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:*

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

where $r_{i,t+h}$ is the stock return of firm i at time $t+h$, $\beta_{i,t+h-1}$ is a $J \times 1$ dimensional vector of conditional factor loadings and $\varepsilon_{i,t+h}$ is an independent identically distributed normal random process, $\mathcal{N}(0, \sigma_{i,\varepsilon})$. We consider a subset of future time periods, $t+h$, where $h \in \{1, 2, 3, 13, 25, 37, 49, 61, 91, 121\}$.

From equation (1), expected returns for month $t+h$ follow the process:

$$\mathbb{E}_t[r_{i,t+h}] = \mathbb{E}_t[\beta'_{i,t+h-1}]\mathbb{E}_t[F_{t+h}] \tag{2}$$

where the $\mathbb{E}_t$ function is defined with respect to the information set, I_t.

Assumption 2. *The factor model admits the parametrization:*

$$\mathbb{E}_t[\beta'_{i,t+h-1}]\mathbb{E}_t[F_{t+h}] \approx g^*_h(z_{i,t}) \tag{3}$$

where $g^*_h(\cdot)$ is some real-valued deterministic function of P real variables, $z_{i,t}$. Specifically, $z_{i,t} = [I_{i,t} : I_t]$, where $I_{i,t}$ is firm specific and I_t is the same across firms. Following Gu et al. (2020b), we specify $I_{i,t}$ as a vector of firm level characteristics and I_t as a vector of aggregate variables.

It is important to point out that this parameterization imposes very minimal economic restrictions on approximating Eq. 2. All it seeks to do is approximate the sum product of conditional betas and conditional price of risk.

Despite its flexibility, the framework imposes a number of important constraints 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 we need to estimate and also increases the stability of the resulting estimates. This restriction is loose in that we re-estimate $g_h^*(\cdot)$ every two years which means that each subsequent 24 month set of stock forecasts comes from a slightly different approximation of $g_h^*(\cdot)$. Finally, the specification also assumes that the same information set is $[I_{i,t} : I_t]$ is relevant for making predictions beyond months $t+1$.

2.1.1 Approximating $g_h^*(\cdot)$

Our learning algorithm of choice is a neural network because it has been shown to be the best approximator of $g_1^*(\cdot)$ with respect to our information set (see Gu et al. (2020b)). Specifically, we approximate Eq. (3) using a three layer deep neural network which we define as:

$$Y_1 = \psi(z_{i,t}W_0 + b_0) \tag{4}$$

$$Y_2 = \psi(Y_1W_1 + b_1) \tag{5}$$

$$Y_3 = \psi(Y_2W_2 + b_2) \tag{6}$$

$$\hat{r}_{i,t+1} = Y_3W_3 + b_3 \tag{7}$$

where $W_0 \in \mathbb{R}^{64\times32}$, $W_1 \in \mathbb{R}^{32\times16}$, $W_2 \in \mathbb{R}^{16\times8}$, $W_3 \in \mathbb{R}^{8\times1}$, $b_0 \in \mathbb{R}^{1\times32}$, $b_1 \in \mathbb{R}^{1\times16}$, $b_2 \in \mathbb{R}^{1\times8}$ and $b_3 \in \mathbb{R}$ are unknown parameters, θ, to be estimated. ψ is a non-linear function applied element-wise after linearly transforming an input vector, either $z_{i,t}$ or Y_k. The activation function(ψ) we use is the rectified linear unit (RELU), $\psi(\cdot) = max(y, 0)$.

2.1.2 Loss Function

We estimate the neural network model by minimizing the mean squared error loss function with an l_1 penalty:

$$\mathcal{L}(\theta) = (N_t T)^{-1} \sum_{i=1}^{N_t} \sum_{t=1}^{T} \left(R_{t+1} - \hat{R}_{t+1}\right)^2 + \lambda_1 \left|\left|\theta\right|\right|_1 \tag{8}$$

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. We minimize the empirical loss function over a pool of firm-month observations. We choose hyper-parameters such as λ_1 via a validation set.

2.1.3 Estimation

We use the AdaBound learning algorithm from Luo et al. (2019) to estimate the unknown parameters (θ)[5].

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

2.1.4 Sample Splitting

The dataset starts from January 1965 and ends in December 2018. We 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, we estimate the model using 15 years of data that starts from January 1975 and ends in December 1989. We choose hyper-parameters by comparing estimated model performance

[5]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 decent optimizer. In addition, AdaBound has theoretical convergence guarantees which other optimizers lack.

over a validation dataset starting from January 1990 to December 1994. We use the optimal model to make one-month ahead return predictions from January 1995 to December 1996. Figure 1 illustrates this exercise. We 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, we estimate the model using 15 years of data that starts from December 1974 and ends in November 1989. We choose optimal hyper-parameters by comparing estimated model performance over a validation dataset starting from December 1989 to November 1994. We use the optimal model to make two-month ahead predictions starting from from December 1994 to November 1996. This ensures that the returns are realized from January 1995 to December 1996, thereby aligning return realization dates across prediction periods, h. Similar to $t+1$, we move the training, validation and test set forward by two years and repeat the process.

We always predict returns from January 1995 to December 2018 by shifting the conditioning information further into the past. This allows us to maintain the same training, validation and testing data size, in months, across horizons. Although this also allows us to compare forecasts from different horizons for some month s, the set of firms change as we condition on firms present during the forecasting date. We choose to estimate monthly forecasts because this allows us to bring the entire asset pricing econometric toolset to the problem and side step the econometric issues inherent in compounded returns.

2.2 Data

We obtain monthly market data for US stocks traded on AMEX, NASDAQ and NYSE stock exchanges from CRSP. We match market data with annual and quarterly fundamental data from COMPUSTAT. We build a set of 56 firm-level characteristics from this panel.[6] The characteristics are computed as per their definitions in Freyberger et al. (2020) and Green et al. (2017). We obtain one-month risk free rate from Kenneth French's website. To avoid forward-looking bias, we follow the standard practice in the literature and delay monthly, quarterly and annual characteristics, by a month, 4 months and 6

[6]The details of the characteristics are provided in Table A.

months respectively (e.g. Green et al., 2017; Gu et al., 2020b). To be included in the sample for some month t, a firm must have 30 non-missing characteristic observations. We rank-normalize the characteristics to the interval [-1,1] and replace missing values with zero.

The aggregate variables we use are 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[7].

3 Neural network forecasts in the cross-section of stocks across horizons

Chen et al. (2019) and Gu et al. (2020a) show that neural network forecasts are one of the best predictors of realized returns in the cross-section of stock. In this section, we seek to extend our understanding of exactly which dimension of returns these forecasts are predicting, specifically, we ask, "Are the forecasts predicting correctly cross-sectional variation in returns in addition to predicting correctly the level of returns?" We do this across horizons which allows us to contrast short run expected return dynamics with longer run dynamics.

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

$$R^2_{OOS} = 1 - \frac{\sum_{(t) \in oss}(R_t - \tilde{R}_{t,1})^2}{\sum_{(t) \in oss}(R_t - \tilde{R}_{t,2})^2} \tag{9}$$

R_t is the time t vector of realized stock returns. 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 of model 2.

We formally test the null hypothesis that forecasts from model 1 are no different from

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

forecasts from model 2 in explaining variations in stock returns using the Clark-West (2007) test with Newey-West (1987) adjusted standard errors.

3.1 Can neural networks predict stock returns across horizons?

To answer this question, we define forecasts for model 1 as forecasts from the neural network model. We compare the neural network forecasts to two standard alternatives in the literature that generate monthly firm return forecasts as a function of firm historical returns. The first alternative model forecasts 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 model. The second alternative 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 model.

[Insert Table 1 about here]

The first two rows of Table 1 report results for these alternative models described above. The R^2_{OOS}s are positive for all future periods and statistically significant. The estimates are larger for predictions made for periods closer to time t, the date of information revelation. However, these positive R^2_{OOS} are inflated because firm average returns are very noisy estimates of expected returns.

To see this we consider a tougher benchmark currently used in the literature. This benchmark forecasts a zero return for all firms across time; the zero-prediction model. The neural network model improves upon the zero prediction forecasts as the R^2_{OOS} is positive across all horizons. All the improvements are also statistically significant at the 5% significance level. However the reported R^2_{OOS} in row three are much lower compared to rows one and two which shows that indeed the first two alternative benchmarks inflate the R^2_{OOS}.

From the results above, we can conclude that the neural network forecasts are better predictors of stock returns compared to all three alternatives. This out-performance is much stronger in the short-run and the strength decreases with horizon.

3.2 Disentangling the composite R^2_{OOS}

The R^2_{OOS} tells us how much model 1 improves upon the forecasts of model 2. It fails to tell us along which dimension the observed improvement is coming from. To shed light on this, we assume a two factor structure holds for these forecasts. We 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. This allows us to decompose the return forecasts from some model m for a firm i at some time t into two parts:

$$r_{m,i,t} = (N_t^{-1}) \sum_{i \epsilon t} r_{1,i,t} + r^{RR}_{m,i,t} \tag{10}$$

where $(N_t^{-1}) \sum_{i \epsilon t} r_{m,i,t}$ captures the cross-sectional mean forecasts of model m and $r^{RR}_{m,i,t}$ captures the cross-sectional variation in forecasts across firms.

This decomposition follows from a lot of asset pricing models in the literature. To see this, consider a world were a version of the CAPM holds and return forecasts follow the process; $r_{m,i,t} = \beta_m * r_{1,MKT,t}$. Decomposing return forecasts in this world according to Eq. (10) results in:

$$r_{m,i,t} = r_{1,MKT,t} + (\beta_m * r_{1,MKT,t} - r_{1,MKT,t}) \tag{11}$$

where $r_{1,MKT,t}$ captures the level of the market factor in the cross-section similar to $(N_t^{-1}) \sum_{i \epsilon t} r_{1,i,t}$ and $(\beta_m * r_{1,MKT,t} - r_{1,MKT,t})$ is now a market-neutral factor that strictly explains the cross-sectional variation across stocks similar to $r^{RR}_{m,i,t}$. It is straight forward to extend this to other multi-factor models in the literature as the extra factors are always defined to be market-neutral and so are all subsumed by $r^{RR}_{m,i,t}$.

We opt for this simple intuitive specification because it allows us to straightforwardly answer the question, "What fraction of the R^2_{OOS} comes from getting correctly the level factor (equally-weighted market return) and what fraction comes from predicting correctly return variations in the cross-section?"

To see why this is an important point, consider a world where there are two firms, A and B. The realized return for firm A is 20% and that of firm B is 10%. We have a forecasting model that forecasts a return of 8% for Firm A and 12% for firm B. Estimating

the R^2_{OOS} against a zero prediction model, we get 70.4%. However when we decompose this R^2_{OOS} as above, we find all the positive R^2_{OOS}, as much as 80%, comes from predicting a cross-sectional mean return of 10% which is much closer to the truth 15%. The alternative model, zero prediction, is worse as it is 15 percentage points from the truth. The composite R^2_{OOS} is hiding the fact that the forecasting model does a very poor job predicting cross-sectional variation in returns. It predicts a -2% relative return for Firm A and a 2% relative return for Firm B against truth values of 5% and -5% respectively. The contribution of the relative return dimension ($r^{RR}_{m,i,t}$) to the composite R^2_{OOS} is -9.6%.

The example above is a one-period example and so hides the fact that relative stock returns are timing varying. To capture this fact, we further decompose the relative forecasts ($r^{RR}_{m,i,t}$) into an unconditional component ($\mu^{RR}_{1,i}$) and conditional component ($\tilde{r}^{RR}_{1,i,t}$).

$$r^{RR}_{m,i,t} = \mu^{RR}_{m,i} + \tilde{r}^{RR}_{m,i,t} \tag{12}$$

where $\tilde{r}^{RR}_{1,i,t}$ is mean zero and captures the residual time-series forecasts of model m and $\mu^{RR}_{1,i}$ captures the residual unconditional stock forecast[8]. Intuitively, $(N_t^{-1}) \sum_{i \in t} r_{1,i,t}$ captures the idea that there is a single dominant factor in the cross-section of stocks. If a forecasting model sufficiently matches this single factor, it will outperform a zero-prediction model. $\tilde{r}^{RR}_{1,i,t}$ and $\mu^{RR}_{2,i}$ capture the fact that even after accounting for the market, there still exists residual variations in stock returns that are predictable. This predictability has been widely documented in the literature in the form of characteristics that explain the cross-section of stock returns (see Hou et al. (2020)).

[Insert Table 2 about here]

In Panel A of Table 2, we present the results for the decomposition of the R^2_{OOS} where the benchmark model is the zero prediction model. From row 4 of this panel, it is evident that a majority of the positive R^2_{OOS} comes from explaining variations in the dominant factor in the cross-section of stocks. For instance, about 70% of the composite R^2_{OOS} for month $t+1$ comes from predicting better the cross-sectional mean return. The remaining 30% comes from predicting better time-series variation in relative stock returns than a zero-prediction.

[8] See section C in the Appendix for more details on the decomposition.

For the other short-run forecasts ($t+h$, where $h \leq 31$), about 100% of the positive composite R^2_{OOS} comes from better explaining variations in the cross-sectional mean return and another 5% comes from explaining time-series variations in relative stock returns. Forecasts for the unconditional relative stock return contribute negatively to the composite R^2_{OOS} at these horizons. The positive R^2_{OOS} for long horizon forecasts all come from the forecasts' ability to predict variations in the cross-sectional mean.

The results in this section highlight the fact that intermediate and long-run forecasts are very different from short-run forecasts. Whereas short-run forecasts strongly discriminate between high and low expected return stocks (statistically significant relative forecasts), intermediate and long-run forecasts all come from matching the equally-weighted market return.

If these forecasts are to be used as discount rates for projects with ex-ante known characteristics, the results imply we can get project specific discount rates only up to about a year out. For project cash-flows that accrue over much longer periods, the consistent discount rate to use is the market return. This is because conditional on this information set, we cannot predict variations in expected returns in the cross-section for the long-term.

3.2.1 Is the zero-prediction model inflating the R^2_{OOS}?

The first question raised by the decomposition above is whether or not the zero-prediction model is the wrong benchmark to compare against when forecasting returns in the pool of individual stocks. This question is informed by the fact that most of the positive R^2_{OOS} comes from matching the cross-sectional mean return. This suggests an alternative benchmark that does reasonably well along this dimension will be a much tougher benchmark for the neural network to beat. From Goyal and Welch (2008), we know one such benchmark is the the average equally-weighted market return. We define this benchmark's $t+h$ stock return forecast as the time t average market return computed using data from 1926.

The results are reported in Panel B of Table 2. In the short-run, we find a more than 65% reduction in the composite R^2_{OOS} compared to the zero-prediction model. This result shows that a majority of the perceived improvement in forecasting short-run stock

returns are as a result of using a weak benchmark. In the long-run, we see a general increase in the composite R^2_{OOS} compared to the zero-prediction model (row one of panel A against row one of panel B), which means that the zero-prediction model is a more challenging benchmark in the long-run.

To tease out why the historical average market return is a more challenging benchmark in the short-run but a weaker benchmark in the long-run, we drop all firms that fall out of the sample over time and report the results in Panel C of Table 2. We now find a consistent reduction in the composite R^2_{OOS} across horizons compared to those reported for the zero-prediction model. The difference between Panels B and C come from the fact that about 50% of firms alive today fall out of the sample in the long-run. This means that conditional on firms alive at time t, about half of firms have an implicit return of zero in the long-run, which draws the cross-sectional mean-return in the long run closer to zero.

3.2.2 Can the forecasts predict market returns across horizons?

The second question raised by the decomposition in Panel A of Table 2 is whether or not the aggregated stock level forecasts can predict market returns across horizons[9]. To answer this question, we define the market forecast as the value-weighted monthly stock forecasts for period $t + h$ and define the market return as the value-weighted monthly stock returns of firms in the CRSP file at time $t+h$.To capture the pure effect of different forecasts, we always use market-caps from time t but allow the forecasts to vary across horizons. We 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]

Table 3 presents the result. The composite R^2_{OOS}, against the zero-prediction model, show that neural network forecasts can predict market returns across horizons. The decomposition of this R^2_{OOS} shows that about 90% of the reported improvements against a zero-prediction is coming from predicting the unconditional market return, with the

[9]It is worth pointing out that the results in Panels B and C of Table 2 do not rule out this hypothesis because of the R^2_{OOS} in row 4 of Panels B and C, although small, are still mostly statistically significant.

other 10% coming from predicting time-series variations in market returns. As we show in the case of individual stock returns, using the wrong benchmark model can lead to inflated composite R^2_{OOS}. We, therefore, replace the zero prediction benchmark with an alternative model where the market forecast is defined as the historical average market return.

Panel B of Table 3 presents the results. The first thing we observe is that the composite R^2_{OOS} across horizons is now negative. This means that the average historical market return is a better predictor of market return compared to the neural network forecasts. However when we focus on the decomposed R^2_{OOS} it becomes clear that the historical market return only out-performs the neural network model because it better predicts the unconditional market return.

3.2.3 Can neural networks predict returns to long-short portfolios?

The positive and statistically significant R^2_{OOS} in row 2 of Panel A in Table 2 suggest that the neural network forecasts may 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 which translates into returns of long-short portfolios. The answer to this question does not follow directly from the results in Table 2 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.

To answer the question, we sort stocks on the five characteristics in the Fama and French (2018) factor model; book-to-market, investment, size, operating profit and momentum[10]. For characteristics that condition on balance sheet or income statement variables, we update characteristics at the end of June of year s using the characteristic observations from the fiscal year-end $s-1$. For characteristics that condition on only market variables from CRSP, we update the characteristic observation at the end of each month and re-balance accordingly. We form decile portfolios from the sorts and value-weight to reduce the effect of small stocks. The return (forecast) to the long-short

[10]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.

portfolio is the value-weighted return (forecast) of portfolio ten minus the value-weighted return (forecast) to portfolio one.

Similar to analyzing return predictability in stocks, we use the decomposed out-of-sample (R^2_{OOS}) to investigate predictability of these long-short characteristic portfolios. We define the alternative model as the zero-prediction model.

[Insert Table 4 about here]

Table 4 reports the decomposed R^2_{OOS} for these five portfolios. The general conclusion from the results is that the neural network forecasts fail to predict returns to long-short portfolios.

4 Regularization using Economic Theory

Regularization is the process of adding information to a machine learning problem to reduce overfitting and improve generalization (Bülhmann and van der Geer (2011)). This is usually achieved by introducing an additional term in the loss function we are minimizing so as to constrain the space of permissible parameter values, Θ. For example, is Eq. (8) where we add an L_1 penalty to our mean-squared loss function.

Choosing a loss function introduces an inductive bias that pushes the resulting approximating function towards our preferred function class. Adding a penalty term can also be viewed as choosing between simple and complex functions (Bülhmann and van der Geer (2011)). For instance, the choice of an L1 norm will result in a simpler model with fewer covariates out of the available covariates. In contrast, an L2 norm results in a more complex model that prefers to include all available covariates.

We do not need to explicitly declare regularization terms in our loss function to prevent overfitting. We can implicitly introduce regularization through our choice of a learning algorithm and our approximating function class choice. An example of such a decision is the use of convolutions over affine functions as neural network primitives (constitutional neural networks) in solving vision problems.

Following this discussion, we propose to regularize the of learning Eq. (2) by adhering as close to economic theory as possible. In the following subsections, we detail how we

allow economic theory to shape the nature of the approximating function.

4.1 Economic restrictions

Guided economic theory, we introduce a few additional assumptions to pin down the structural nature of Equation (2).

Assumption 3. *Expected returns are linear in conditional betas and conditional price of risk:*

$$\mathbb{E}_t[\beta'_{i,t+h-1}]\mathbb{E}_t[F_{t+h}] \approx b^*_h(\cdot)' * f^*_h(\cdot) \tag{13}$$

where $b^*_h(\cdot)$ is a function that approximates the time $t+h$ expected conditional betas of firm i and $f^*_h(\cdot)$ is a function that approximates the time $t+h$ expected conditional price of risk. The crucial assumption here is that expected firm returns is the sum of the products of conditional risk loadings (betas) and 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.

We can impose this linearity assumption only because we model the conditional price of risk and conditional beta functions separately. This separation also allows us to treat conditioning information more in line with findings in the literature. Specifically, we treat characteristic realizations as being informative of risk loadings as in Kelly et al. (2019). And we treat the conditional price of risk as arising from linear combinations of trade-able portfolios formed from sorts on characteristics as in Fama and French (1996).

4.1.1 The conditional price of risk function

The conditional price of risk function, $f^*_h(\cdot)$, is initialized with a $(P+2)$-dimensional vector of portfolio average returns, $\bar{r}_{p,t}$, when predicting returns for time $t+h$. This vector comprises an expanding window average return of long-short portfolios formed from sorts on the P firm-characteristics introduced in the previous chapter. We concatenate this vector with the expanding window average return of the equally-weighted market portfolio

and risk-free rate[11]. Specifically, the conditional price of risk function is defined as:

$$\mathbb{E}_t[F_{t+h}]' = \bar{r}_{p,t} W_0 + b_0 \tag{14}$$

where $W_0 \in \mathbb{R}^{58\times 3}$ and $b_0 \in \mathbb{R}^{1\times 3}$ are unknown parameters to be estimated[12]. This parameterization allows for the pricing function to be dense in portfolio and security returns (58 average returns) and simultaneously remain sparse in pricing factor returns (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, we set the number of pricing factors to 3[13]. It is worth highlighting the fact that the small number of factors we impose does not restrict the resulting approximator to the same functional space as a three principal component (PC) model. This is because loadings in Eq.(13) are time-varying as opposed to the statistic loadings in a PC model.

We do not allow for non-linear interactions between portfolio returns in determining the factor returns because we want the resulting factors to be trade-able portfolios. A factor portfolio is trade-able if defined as a linear combination of a set of trade-able portfolios. This restriction implies each of the three factors is in the span of the 58 portfolio/security returns. Similarly, we construct each long-short characteristic sorted portfolio by fixing portfolio weights as the rank-normalized characteristic realizations for some time t. We then go long one dollar and short another dollar. Consequently, the long-short portfolios are all spanned by the stocks in the cross-section.

4.1.2 The conditional beta function

The conditional beta function, $b_h^*(\cdot)$, is initialized with P-dimensional vector of rank-normalized firm characteristics, $z_{i,t}$, when predicting returns for time $t+h$. We assume that characteristic realizations at time t are informative of their time $t+h$ realizations[14].

[11] We compute all expanding window means using portfolio returns starting from January 1965.

[12] We estimate all model parameters following the same procedure as outlined in Section (2).

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

[14] Given that some characteristics are highly persistent, this is not a controversial claim. Replacing the time t realizations with rolling window averages does not change the results.

We approximate the beta function as:

$$Y_1 = \psi(z_{i,t}W_0 + b_0) \tag{15}$$

$$Y_2 = \psi(Y_1W_1 + b_1) \tag{16}$$

$$\mathbb{E}_t[\beta_{i,t+h-1}]' = Y_2W_2 + b_2 \tag{17}$$

where $W_0 \in \mathbb{R}^{56\times1024}$, $W_1 \in \mathbb{R}^{1024\times1024}$, $W_2 \in \mathbb{R}^{1024\times3}$, $b_0 \in \mathbb{R}^{1\times1024}$, $b_1 \in \mathbb{R}^{1\times1024}$ and $b_2 \in \mathbb{R}^{1\times3}$ are unknown parameters to be estimated. ψ is the relu non-linearity.

Even though we initialize all H conditional beta 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 returns for time period $t+1$ but negatively related to returns for time period $t+13$ (the reversal characteristic). Consequently, it follows that the learned relationship between the same characteristic and returns at different horizons by the neural network model will be different.

In the asset pricing literature, beta functions are usually specified as unconditional scaling functions that load on factor portfolio returns. Although this parameterization restricts the resulting model, they are still preferred to conditional variants because they are easy to estimate. Given that we estimate the parameters of predictive models in this paper using stochastic gradient descent, we 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, our predictive model is much more general in that we allow beta to change in response 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 function, which would have to be a scaler that seeks to capture the average risk loading of both the growth and value phase of the firm.

Besides allowing for time-varying beta, we also allow for nonlinear interactions between firm characteristics via the ψ non-linearities. This specification is motivated by

findings in **?** and Bryzgalova et al. (2019), that show that there exist nonlinear relations between characteristic realizations and expected returns. We further project the 56 firm characteristics into a 1024 higher dimensional space that allows the model to learn unrestricted nonlinear interactions between the base characteristics before projecting the resulting features (covariates) back into the 3-dimensional latent pricing factor space (see Johnson (2019)).

4.2 Does economic theory help?

In the previous section, we show that a neural network model may predict time-series variation in relative stock returns and still fail to robustly predict market and long-short portfolio returns. By putting more structure on the prediction problem guided by economic theory, we show in this section that the resulting prediction model is much more robust.

In addition to the reporting results for decomposed out-of-sample R^2, we also report results from regressing the demeaned forecasts on returns. The slope coefficient from this regression answers the question, "How biased are the forecasts?" An unbiased forecast will have a slope coefficient that is indistinguishable from one. A useless forecast will have a slope coefficient indistinguishable from zero.

4.2.1 Pool of stocks

Panel A of Table 5 report results for individual stocks in the pool of stocks. We show the composite and decomposed out-of-sample R^2 of the neural network model, that respect economic theory, against a zero prediction alternative and average equity market return benchmark models.

[Insert Table 5 about here]

Comparing the results in panel A of Table 5 to the results in panel B of Table 2, it is evident that the most pronounced change is in the ability of the forecasts to better predict time-series variation in relative stock returns for the short-run months. The largest (least) improvement is for period $t+1$ ($t+13$) where the out-of-sample increases

from 0.20 (0.08) to 0.71 (0.08). As was shown previously, using a zero-prediction model inflates the composite R_2 for the short-run months. We find the same effect with respect to the new forecasting model. Overall, we observe a reduction of about 30% when the alternative model is the average equity market return for the short-run months and an increase of around 10% for the long-run months.

[Insert Table 6 about here]

Panel A of Table 6 report results from regressing the demeaned relative stock returns on realized stock returns. The results generally confirm the conclusions from the decomposed out-of-sample R^2. For the short-run months, $t+1$ up to $t+13$, we can reject the null hypothesis that these forecasts fail to predict time-series variation relative stock returns because zero does not lie within the 95% confidence interval of the slope coefficient. For $t+1$, the forecasts are unbiased because the 95% confidence interval is centered around one and does not include zero. The forecasts over-predict time-series variation in relative stock returns for all other short-run months because the confidence intervals are strictly smaller than one and greater than zero.

4.2.2 Value-weighted market

Panel B of Table 5 reports results for the value-weighted market and is comparable to Table 3. We find that the new forecasting model more robustly predicts time-series variation in market returns. For the short-run and intermediate months ($h \leq 37$), we find positive and statistically significant R^2_{OOS} with respect to this dimension of market returns. Panel B of Table 6 confirms these conclusions using time-series regressions. For months $t+1$ up to $t+37$, the slope coefficients from regressing demeaned market forecasts on market returns is positive and statistically different from zero.

4.2.3 Long-short portfolios

Panel C of Table 5, report results for long-short portfolios formed from the five characteristics in the Fama and French (2018) factor model. We see a pronounced improvement in the ability of the neural network model to predict returns to these portfolios. Focusing on forecasts for time $t+1$, we now report positive R^2_{OOS} for all five characteristics as against

two in Table 4. For four of these five characteristics, the R^2_{OOS} are statistically significant at least the 5 % level. The decomposition of the R^2_{OOS} shows that the forecasting power of the new model is driven by its ability to better predict time-series variation in returns to these long-short portfolios. Similar to the market, predictive regressions in panels C and D of Table 6 re-affirm the conclusions drawn from the predictive regressions.

To show how pervasive this finding is, we expand the universe of long-short portfolios to all 56 characteristics that we condition on in the beta function (Equation (15)) and focus on forecasts for month $t+1$. Figure 2 reports the results.

[Insert Figure 2 about here]

In panel A, we 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 ability of the forecasts to predict time-series variation in returns of long-short portfolios. In panel B, we find that 53 (31) of the 56 long-short portfolios have positive (and statistically significant) R^2_{OOS} pertaining to the ability of the neural network forecasts to out-perform the expanding window average return of the long-short portfolio in predicting the time-series variation in these returns.

[Insert Figure 3 about here]

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

The results in this section shows that machine learning guided by economic theory can lead to significant improvements in predicting returns.

4.3 Has predictability decayed over time?

The recent work of McLean and Pontiff (2016) shows that academic research destroys return predictability. Since most of the variables in our conditioning set came from academic research and were published at different points in our sample period, 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 a waning of the out-of-sample R^2 over time. To answer this question, we compute a two-year rolling out-of-sample R^2 from January 1997 to December 2018. We present the results for forecasts generated for month $t+1$.

[Insert Figure 4 about here]

Panel A of Figure 4 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, we can conclude that the predictive ability of the information set has waned over time. However, we cannot conclude that the forecasts have 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 we 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. By the end of the sample, the rolling estimates are negative.

5 Optimal Portfolios

This section introduces several optimal trading strategies that take advantage of the predictive content of the neural network forecasts. The point of the strategies is to show that the forecasts are robust proxies for expected returns. We come to this conclusion by showing the incremental benefits in the form of average return, Sharpe ratio, risk-adjusted return, and certainty equivalents that an agent could have enjoyed by using these forecasts in a pseudo-real-time setting over the out-of-sample period. We define the certainty equivalent with respect to an agent with mean-variance utility function and a risk aversion parameter of 2. Specifically, we compute the certainty equivalent return of a strategy as:

$$CE = \bar{r}_h^p - \frac{\gamma}{2}\sigma_{p,h}^2 \tag{18}$$

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

5.1 Optimal rotation strategies

We start by considering optimal rotation strategies that attempt to rotate across securities in the cross section of stocks and long-short characteristic sorted portfolios. For each period t, we sort all securities in a particular cross-section on their forecasted return, buy (sell) the top (bottom) 10% of securities. We repeat these sorts H times for each forecasting horizon we study in this paper. 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 7 about here]

5.1.1 Long-short stocks strategy

This is a market neutral 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. Panel A of Table 7 reports the performance statistics for this strategy.

The results are presented in panel A of Table Table 7. Re-balancing the long-short stock portfolio each month using the t+1 forecasts leads to an certainty equivalent of 11.73 which is 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. The improved performance of the portfolio therefore does not come from leveraging on the fundamental factors in either of these portfolios. Generally, the certainty equivalents, average returns and alphas fall the with horizon. The right way to think about this is that using older forecasts to re-balance the long-short portfolios comes at a cost. Remember these portfolios are re-balanced monthly, and the horizon dimension is with respect to how old the forecasts are. A monthly re-balanced long-short portfolio using forecasts for period $t + 120$ mean re-balancing with forecasts that are 10 years old.

5.1.2 Long-short characteristics strategy

This cross-section is made up of the five characteristics in the Fama and French (2018) 5 factor model. The strategy buys (sells) an equally-weighted portfolio of the two long-short characteristic sorted portfolios with the highest (lowest) expected returns. All long-short portfolios are value-weighted.

The results are presented in panel B of Table Table 7. 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 wane as we use old 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 %.

5.2 Optimal timing strategies

We consider a strategy that tries to time a risky asset by leveraging up and down this security based on whether expected returns are high or low. For each month t, we use the conditional expected return forecast from the neural network model to calculate the Markowitz optimal weight to be invested in the risky asset as:

$$w_{t,h} = \frac{\tilde{r}_{t,h} - r^f_{t+1}}{\gamma \sigma(\tilde{r}_{1:t-1,h})} \tag{19}$$

where γ is a risk aversion coefficient which we fix as 2. Given that our model does not produce a conditional standard deviation estimate ($\sigma(\tilde{r}_{1:t-1,h})$), we fix this value at an annualized value of 15 % across securities. At the end of each month, we compute the timing portfolio return as:

$$r^p_{t,h} = w_{t,h} r_{t+1,h} - (1 - w_{t,h}) r^f_{t+1} \tag{20}$$

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

5.2.1 The optimal market timing portfolio

The first trading strategy we consider seeks to time the value-weighted market return by deciding on how much to invest between the market and a risk free asset using the neural network forecast for the market. We 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 8 about here]

Panel A of Table 8 reports the results. A buy and hold strategy that is fully invested in the market over the sample period makes a return of 10.39 % with a certainty equivalent of 8.18. The returns to this strategy are fully explained by the CAPM and the Fama and French (2018) 5 factor model. A timing strategy that uses the most recent aggregate forecasts, $t+1$, earns an annualized average return of 17.21 % with a certainty equivalent

of 12.86. Timing the market with forecasts that a month old, $t+1$ to two years old, $t+25$ all out-perform the buy and hold strategy. Similar to what we found for the rotation strategies, the more timely the forecasts are, the more informative they are about expected return. We see this from the higher certainty equivalents and average returns for periods $t+1$ and lower estimates for much older forecasts such as $t+121$.

5.2.2 The optimal characteristic timing portfolio

The second timing strategy we consider seeks 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.

Panel B of Table 8 reports the result for the timing strategy. Similar to previous results, we find that the spread in returns generated by the timing strategy is strongest when the forecast is much closer to the trading 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 timely forecast $t+1$ almost tripples the buy and hold certainty equivalent, average returns and alphas.

6 Variable importance across horizons

In this section, we ask the question seek to show which covariates matter the most in generating return forecasts across horizons. We use the notion of Shapley values from Lundberg and Lee (2017). The authors show that Shapley values generalize a number of competing measures of model explain-ability 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 that represents 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:

$$\phi_i = \sum_{S\subseteq F\backslash k} \frac{|S|!(|F|-|S|-1)!}{|F|!}\Big[f_{S\cup\{k\}}(x_{S\cup\{k\}}) - f_{S\cup\{k\}}(x_S)\Big] \tag{21}$$

Shapley additive explanation (SHAP) values represent an easily computable approximation of Eq. (21). These values provide the unique covariate importance measure we use therein [15].

[Insert Figure 5 about here]

Figure 5 reports the overall ranking of characteristics for select horizons. We 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. We analyze in the results in relation to the market given the results in the previous chapters. Across, horizons, the forecasts generated by the neural network model overwhelming explain variations in the market factor as against cross-sectional variation.

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. The third important variable is

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

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 these estimates and their impact on return forecasts with respect to the cross-section of stocks, it is important to remember that there exists a factor structure inherent in this cross-section. And so, it is not clear how exactly to decompose variable importance along similar lines as in Eq. (8).

7 Conclusion

This paper primarily shows that incorporating economic theory in a neural network model's architectural design 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 better.

I show that the model that strictly adheres to economic theory produces forecasts that robustly generalize beyond the cross-section of stocks compared to a simpler neural network 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 I consider. 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. Forecasts for months up to one year into the future predict well both cross-sectional variations in returns and the level factor in returns. Contrast this with forecasts for periods beyond two years into the future, which only explains well variations in the level factor in stocks as against explaining any variation in the cross-section.

The results in this paper answer a lot of questions but also raise a few. The most important being, "Why does cross-sectional return predictability decay so quickly?" 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. We leave the question of which of these two hypotheses we can reject for future research.

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 .

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.

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 .

Bülhmann, P., van der Geer, S., 2011. Statistics for high dimensional data. Statistics (New York). Springer-Verlag, Berlin .

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

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., 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.

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.

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. Empirical asset pricing via machine learning. The Review of Financial Studies 33, 2223–2273.

Gu, S., Kelly, B. T., Xiu, D., 2020b. Empirical asset pricing via machine learning. The Review of Financial Studies - Forthcoming .

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., 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 .

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., 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.

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.

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

Wolpert, D. H., 1996. The lack of a priori distinctions between learning algorithms. Neural computation 8, 1341–1390.

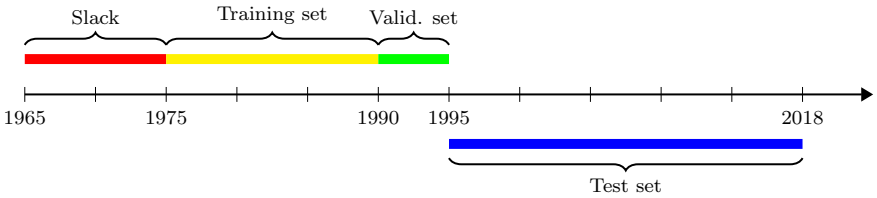

Figure 1: **Sample Splitting time-line**
This figure presents a time-line for sample splitting scheme in the paper.

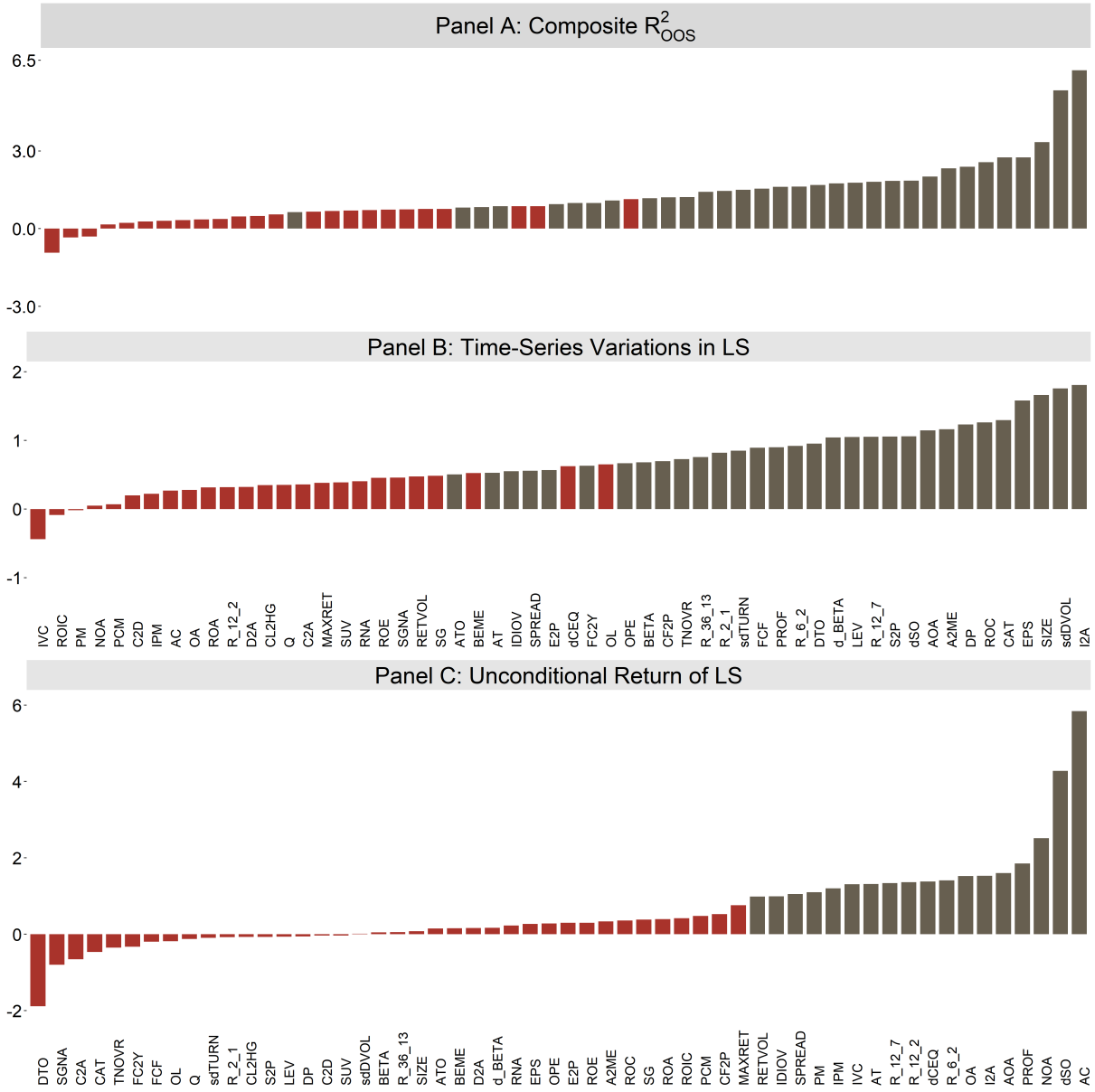

Figure 2: **Forecasting returns of long-short portfolios**
This figure reports the out-of-sample R^2 (R^2_{OOS}) of monthly re-balanced long-short portfolios formed from quintile sorts on 56 characteristics listed in Table A.1 using forecasts from time $t-h$, where $h \in \{1, 2, 3, 13, 25, 37, 49, 61, 121\}$. 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. Red bars represent statistically significant R^2_{OOS} using the the Clark-West (2007) test at the 5% level. The sample period is from January 1995 to December 2018.

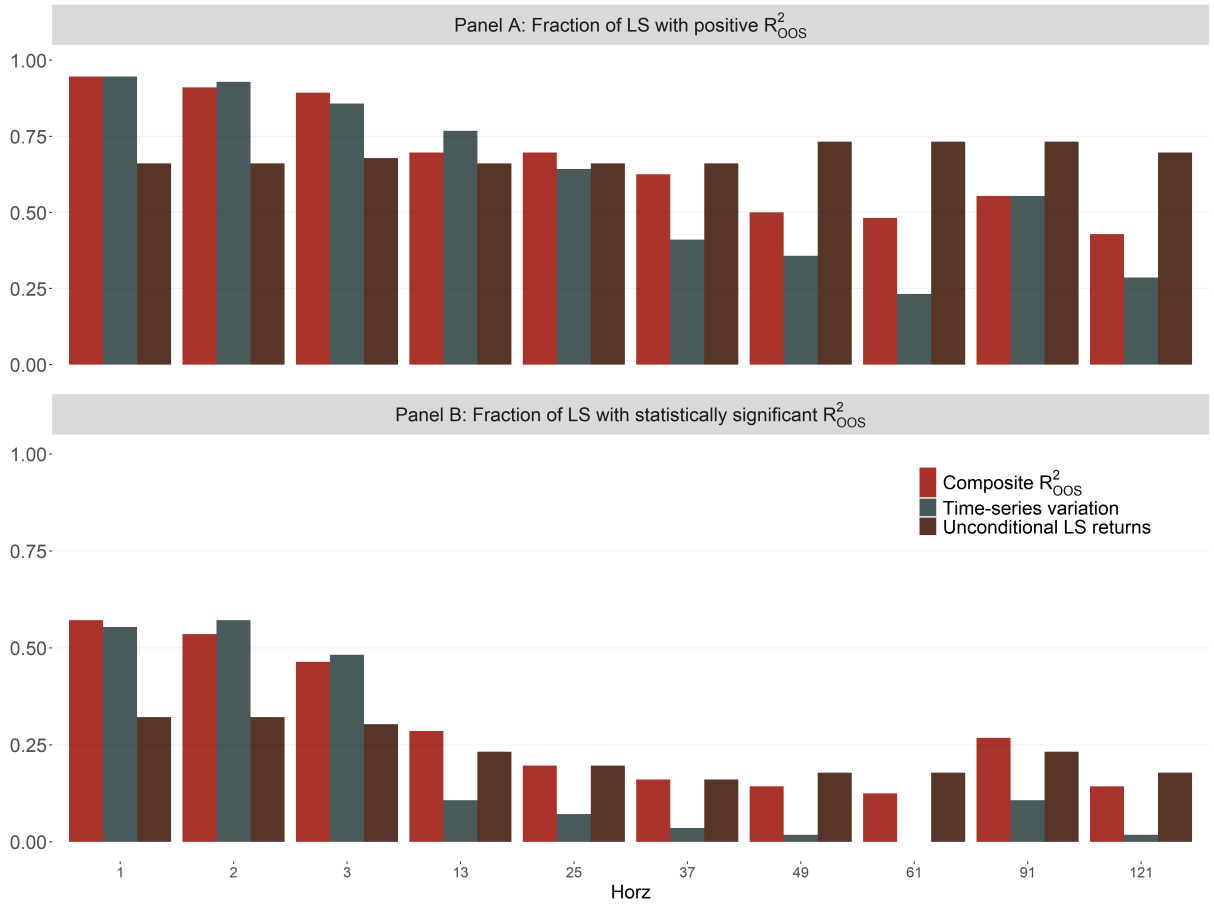

Figure 3: **Forecasting returns of long-short portfolios using forecasts for different horizons.**
This figure reports statistics for long-short portfolios formed from quintile sorts on 56 characteristics listed in Table A.1. We report in panel A the fraction of 56 long-short portfolios where the neural network forecast has a positive R^2_{OOS} with respect 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 forecast $t-h$, 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.

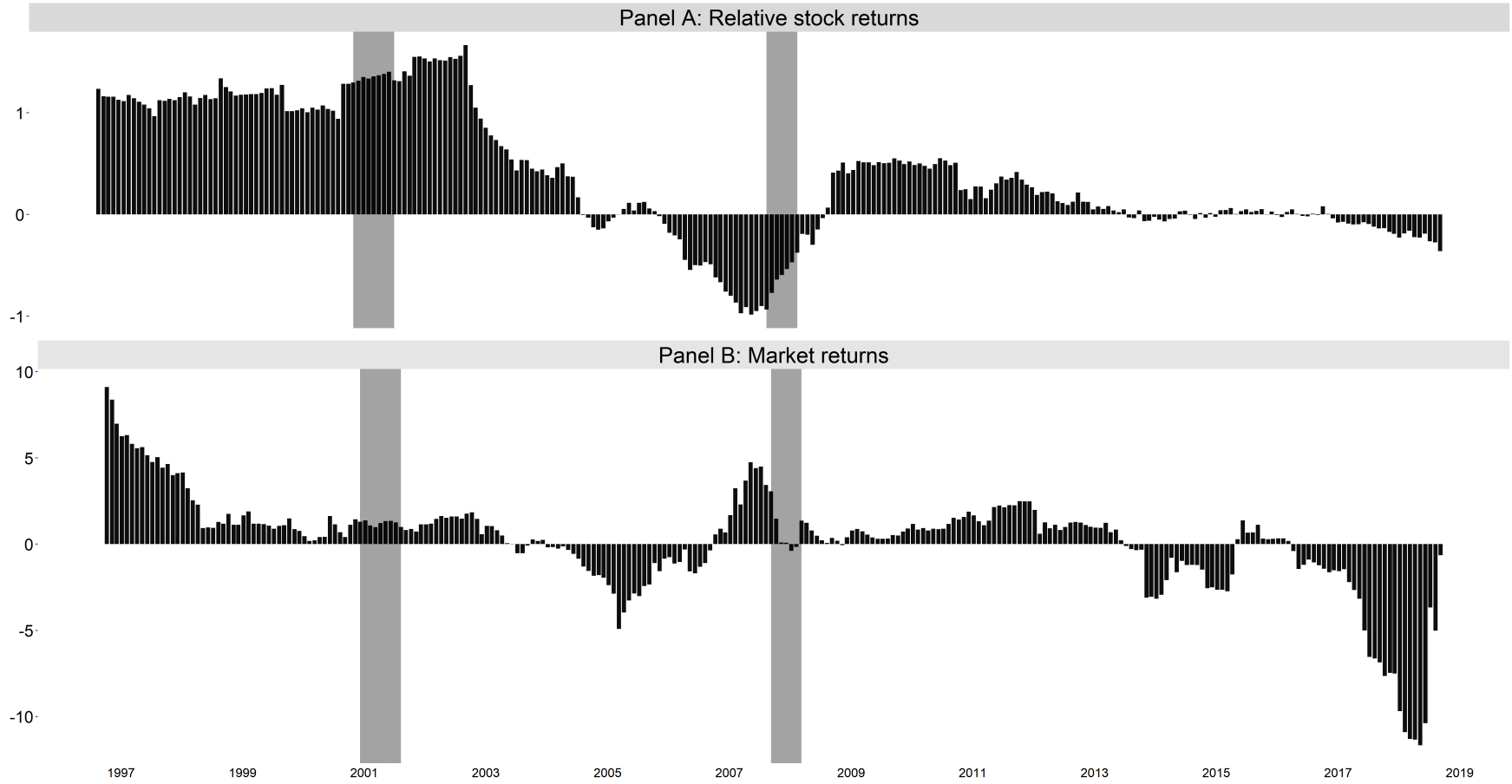

Figure 4: **Two-year rolling out-of-sample R^2**
This figure reports a two-year rolling out-of-sample R^2 (R^2_{OOS}) for demeaned monthly equity return forecasts from a neural network model. The rolling estimates capture the changes in the neural network forecasts' ability to predict time-series variations 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 5: **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 out-of-sample R^2 (R^2_{OOS}) estimates for firm level return forecasts generated by a neural network model at time t for month $t+h$ where $h \in \{1, 2, 3, 13, 25, 37, 49, 61, 121\}$. We present the R^2_{OOS} for the neural network model against three alternative models; a five-year rolling window firm average return model, an expanding window firm average return model, and a zero prediction 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.

$\mathbb{E}_t[R_{t+h}]$	1	2	3	13	31	61	91	121
Panel A: Predicting stock returns								
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**
Zero Prediction	**0.58**	**0.41**	**0.40**	**0.37**	**0.35**	**0.34**	**0.27**	**0.18**

Table 2: **Disentangling the out-of-sample R^2_{OOS}**
This table presents R^2_{OOS} estimates for firm level return forecasts generated by a neural network model at time t for month $t+h$ where $h \in \{1, 2, 3, 13, 31, 61, 121\}$. Panel A reports the R^2_{OOS} for the neural network model against a zero prediction model and decomposes this composite R^2_{OOS} into three parts. Panels B and C report R^2_{OOS} against the equally-weighted market average return computed from an expanding window using data from 1926. Panel B contains all firms that were present at time t whereas panel C drops all firms that fall out of the sample at time t. The time-series variation dimension captures the ability of the neural network model 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 ability of the neural network model to improve upon on forecasts from the benchmark model with respect to predicting unconditional relative stock returns. Finally, the cross-sectional mean return Cross-sectional mean dimension captures the ability of the neural network to improve upon forecasts from the benchmark model with respect to predicting both time series and unconditional cross-sectional mean 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.

$\mathbb{E}_t[R_{t+h}]$	1	2	3	13	31	61	91	121
Panel A: Zero Prediction								
Composite R^2_{OOS}	**0.58**	**0.41**	**0.40**	**0.37**	**0.35**	**0.34**	**0.27**	**0.18**
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**
Panel B: Historical Market Return (Survivorship Biased)								
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**
Panel C: Historical Market Return								
Composite R^2_{OOS}	**0.20**	**0.03**	**0.01**	-0.00	-0.00	-0.00	-0.04	-0.09
Time-series variation	**0.20**	**0.01**	**0.02**	**0.03**	-0.03	-0.02	-0.03	-0.03
Unconditional rel. stock ret.	-0.04	-0.01	-0.02	-0.04	-0.01	-0.01	-0.02	-0.01
Cross-sectional mean return	**0.03**	**0.03**	**0.02**	**0.01**	**0.04**	**0.03**	**0.01**	-0.05

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 $t+h$, $h \in \{1, 2, 3, 13, 31, 61, 121\}$, conditional on information observed at time t. We define the market forecasts as the value-weighted cross-sectional average stock forecasts. We similarly define the market return as the value-weighted cross-sectional average stock return. The first benchmark model is a zero-prediction model that predicts a return of zero for all horizons. The second alternative model is the historical equity market model that forecasts market return as the average market return using data from July 1926 upto time t. We decompose this R^2_{OOS} into contributions coming from the neural network forecasts ability to better explain 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.

$\mathbb{E}_t[R_{t+h}]$	1	2	3	13	31	61	91	121
Panel A: Zero prediction model								
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**
Panel B: Historical equity return								
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**
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). We report the estimates for short-run months ($t+h$ where $h \in \{1, 2, 3, 13\}$). The alternative 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.

$\mathbb{E}_t[R_{t+h}]$	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: **Regularization using economic theory**
This table presents out-of-sample R^2 (R^2_{OOS}) estimates for time t return forecasts for month $t+h$ returns where $h \in \{1, 2, 3, 13, 25, 37, 49, 61, 91, 121\}$. The forecasting model, as defined in Equation (13) adheres more to economic theory than the much simpler model in the previous tables. Panel A shows the decomposition of the R^2_{OOS} against a zero-prediction benchmark and the equally-weighted market average return benchmark, as defined in Table 2. Panel B shows the decomposition of the R^2_{OOS} for the value-weighted market return, just as in Table 3. In panel C, the decomposition is the same as in Table 4, and the alternative model is the average portfolio return computed from an expanding window mean with data from 1965. Bold fonts highlight horizons where 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 is from 1995 to 2018.

$\mathbb{E}_t[R_{t+h}]$	1	2	3	13	25	37	49	61	91	121
Panel A: Predicting returns in the pool of stocks										
	Zero Prediction Alternative									
Composite R^2_{OOS}	**0.99**	**0.49**	**0.40**	**0.35**	**0.29**	**0.28**	**0.31**	**0.28**	**0.24**	**0.13**
Time-series variation	**0.71**	**0.16**	**0.09**	**0.08**	-0.04	-0.06	-0.06	-0.05	-0.07	-0.09
Unconditional rel. stock ret.	-0.16	-0.03	-0.02	-0.08	-0.00	0.03	0.02	0.02	0.04	0.03
Cross-sectional mean return	**0.43**	**0.36**	**0.33**	**0.36**	**0.34**	**0.32**	**0.34**	**0.31**	**0.27**	**0.20**
	Historical Equity Return Alternative									
Composite R^2_{OOS}	**0.64**	**0.14**	**0.04**	0.00	-0.01	**0.05**	**0.12**	**0.16**	**0.37**	**0.62**
Time-series variation	**0.71**	**0.16**	**0.09**	**0.08**	-0.04	-0.06	-0.06	-0.05	-0.07	-0.09
Unconditional rel. stock ret.	-0.16	-0.03	-0.02	-0.08	-0.00	0.03	0.02	0.02	0.04	0.03
Cross-sectional mean return	**0.08**	**0.01**	-0.02	**0.01**	**0.03**	**0.08**	**0.16**	**0.19**	**0.40**	**0.69**
Panel B: Predicting market returns										
	Zero Prediction Alternative									
Composite R^2_{OOS}	**5.35**	**5.30**	**5.17**	**4.90**	**4.97**	**4.42**	**3.97**	**3.05**	**3.48**	**2.77**
Time-series variation	**1.92**	**1.82**	**1.69**	**1.51**	**1.59**	**1.10**	0.76	-0.03	0.19	-0.42
Uncond. market ret.	**3.42**	**3.49**	**3.49**	**3.39**	**3.38**	**3.31**	**3.21**	**3.08**	**3.30**	**3.19**
	Historical Equity Return Alternative									
Composite R^2_{OOS}	**2.05**	**2.03**	**1.89**	**1.64**	**1.69**	1.09	0.64	-0.40	0.13	-0.63
Time-series variation	**2.06**	**1.97**	**1.84**	**1.68**	**1.75**	**1.21**	0.86	-0.04	0.26	-0.39
Uncond. market ret.	-0.01	0.06	0.05	-0.04	-0.06	-0.12	-0.22	-0.36	-0.13	-0.24
Panel C: Predicting LS portfolio returns										
	BEME									
Composite R^2_{OOS}	0.68	0.80	0.73	0.81	-0.13	-0.30	-0.40	-0.17	-0.34	-0.23
Time-series variation	0.52	0.62	0.51	-0.03	-0.13	-0.30	-0.38	-0.13	-0.20	0.01
Unconditional LS Ret.	0.15	0.18	0.22	0.84	0.00	-0.00	-0.02	-0.04	-0.14	-0.24
	SIZE									
Composite R^2_{OOS}	**1.74**	**1.48**	**1.32**	**0.71**	0.33	0.14	0.11	0.01	-0.45	-0.56
Time-series variation	**1.66**	**1.43**	**1.28**	**0.78**	0.33	0.10	0.05	-0.06	-0.41	-0.44
Unconditional LS Ret.	0.08	0.05	0.04	-0.08	-0.00	0.04	0.05	0.07	-0.04	-0.12

Continued

$\mathbb{E}_t[R_{t+h}]$	1	2	3	13	25	37	49	61	91	121
					PROF					
Composite R^2_{OOS}	**0.95**	**0.97**	0.75	0.55	0.24	-0.06	0.18	0.07	0.24	-0.12
Time-series variation	**0.67**	**0.72**	**0.51**	0.28	0.14	-0.03	0.04	-0.05	0.15	-0.14
Unconditional LS Ret.	0.28	0.24	0.25	0.27	0.10	-0.03	0.14	0.12	0.09	0.01
					INV					
Composite R^2_{OOS}	**3.34**	**3.49**	**3.23**	**1.93**	**2.95**	0.30	0.08	0.17	0.00	0.36
Time-series variation	**1.81**	**1.95**	**1.67**	0.56	0.37	0.07	-0.18	-0.13	-0.09	-0.01
Unconditional LS Ret.	**1.53**	**1.55**	**1.56**	**1.37**	**2.58**	**0.23**	**0.26**	0.30	0.10	0.37
					MOM					
Composite R^2_{OOS}	**1.34**	**0.84**	0.43	**0.41**	0.08	-0.21	-0.20	-0.10	-0.25	-0.50
Time-series variation	0.22	0.18	0.22	0.28	0.14	0.02	0.04	-0.02	-0.04	-0.14
Unconditional LS Ret.	**1.12**	**0.66**	0.20	**0.14**	-0.06	-0.24	-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 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.

$\mathbb{E}_t[R_{t+h}]$	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

$\mathbb{E}_t[R_{t+h}]$	1	2	3	13	25	37	49	61	91	121
					Panel D: 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: **Rotation strategies**
This table reports performance statistics for market-neutral portfolios formed from sorts on neural network forecasts. We report the annualized average return (Avg. ret), annualized Sharpe ratio (Sharpe), the annualized certainty equivalent (Utility) in percentages for an agent with mean-variance utility function and a risk aversion parameter of 2, the CAPM and Fama-French 6 factor model alphas. In panel A, we go long (short) the value-weighted portfolio of the top (bottom) 10% of stocks with the highest (lowest) forecasted return for each month t. We restrict the cross-section of stocks to the 500 largest market-cap firms. The buy and hold strategy buys all 500 stocks in this restricted cross-section for each month t. In panel B, we buy (sell) an equally-weighted portfolio of the two characteristic-sorted long-short portfolios with expected return above (below) the median forecast for each month t. The buy and hold strategy buys all five portfolios in each month t. We restrict the characteristic-sorted long-short portfolios to the 5 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.

$\mathbb{E}_{t-h}[R_t]$	Utility	Avg. ret	Sharpe	α_{capm}	α_{FF6}
		Panel A: Stocks			
Buy & hold	8.18	**10.39**	0.70	0.23	0.04
Rotation					
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					
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

Table 8: **Timing Strategies**
This table reports performance statistics, as in Table 7, for the monthly returns of two timing strategies that invest: $w_{t,h} = \frac{\tilde{r}_{t,h} - r^f_{t+1}}{\gamma \sigma^2(\tilde{r}_{1:t-1,h})}$ in a risky security and $1 - w_{t,h}$ in the risk-free asset. We fix $\sigma(\tilde{r}_{1:t-1,h})$ as $\frac{0.15}{\sqrt{12}}$ 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 formed from the charactersiticcs 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.

$\mathbb{E}_{t-h}[R_t]$	Utility	Avg. ret	Sharpe	α_{capm}	α_{FF6}
		Panel A: Market			
Buy & hold	8.18	**10.39**	0.70	0.23	0.04
Rotation					
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: Long-short portfolios			
Buy & hold	3.02	**3.88**	0.42	**5.73**	1.70
Rotation					
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**

Internet Appendix

A Variable Construction

Table A.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).

B 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: $y_i = BN_{\gamma,\beta(x_i):i=1,2,3,...,N}$;

C Decomposing the out-of-sample R^2

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

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

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:

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

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

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:

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

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}:

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

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

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

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

Expanding R^2_{UN}:

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

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

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

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

Expanding R^2_{CS}:

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

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

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

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

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:

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

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

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.