Junior Management Science 10(3) (2025) 561-581

Junior Management Science

www.jums.academy
ISSN: 2942-1861

Predicting Stock Returns With Machine Learning: Global Versus Sector Models

Johannes Witter

Technical University of Munich

Abstract

Recent studies highlight the superior performance of non-linear machine learning models, such as neural networks, over traditional linear models in predicting cross-sectional stock returns. These models are capable of capturing complex non-linear interactions between predictive signals and future returns. This thesis researches whether sector-specific neural networks can detect sector-related relationships to outperform a global neural network. It evaluates the predictive power of these models at the stock level and in portfolios based on return forecasts, constructing long-short portfolios from the networks' sorted predictions. A global neural network model trained on the full sample of stocks dominates neural networks trained on individual GICS sectors in predicting the cross-section of US stock returns. Sector-specific neural networks fail to gain an advantage by capturing complex sector-specific interactions. They underperform the global neural network especially in the early out-of-sample period. The smaller sample size for each GICS sector requires a trade-off between model complexity and robust model estimation. Pooling the data for the global model solves this problem and supports the predictive power of neural networks for stock returns.

Keywords: cross-section of stock returns; machine learning; neural networks; return prediction; sector models

1. Introduction

I compare a global neural network with sector-specific neural networks to predict the cross-section of US stock returns. Therefore, I evaluate the predictive power of the two different models at the stock level and in portfolios constructed based on return forecasts. Recent research demonstrates the ability of non-linear machine learning models such as neural networks to outperform traditional linear models in predicting the cross-section of returns. I show how these neural networks perform better when trained on pooled data across sectors than on sector-specific data from GICS sectors. My data sample covers the sample period from July 1963 to December 2022, with an out-of-sample period from January 1994 to December 2022. The global neural network achieves a higher positive out-of-sample R^2_{OOS} than the sector models, which underperform a naive forecast of zero. For long-short portfolios based on the sorted predictions of the neural networks, the global model outperforms the sector models in terms of generated monthly returns and Sharpe ratio. The comparison of sector neural networks with simple OLS models highlights the necessary trade-off between estimating a stable model without overfitting and capturing complex sector-specific interactions.

I follow the machine learning training approach of Gu et al. (2020). I train neural networks with three hidden layers on 212 stock-specific signals from prior literature to predict the cross-section of returns for the US stock market. The neural networks are trained using a recursive scheme with increasing training samples and a fixed size rolling sample for validation. I refit all models once per year in December and predict monthly out-of-sample returns over the following year.

In total, I train eleven different neural network models. A global model is trained on the full sample of stocks. The

I would like to express my gratitude to all the people who have supported me in the realization of this thesis. In particular, I would like to thank my supervisor, Dr. Matthias Hanauer, for giving me the opportunity to work on this interesting topic and for his guidance and inspiration throughout the work on my thesis.

DOI: https://doi.org/10.5282/jums/v10i3pp561-581

sector model consists of ten sector-specific neural networks trained on subsets of the sample data filtered for each of the ten sectors defined by the Global Industry Classification Standard (GICS). To assign stocks to sectors, I use existing data on GICS classifications and a custom mapping from the Standard Industrial Classification (SIC) system to GICS sectors.

First, I evaluate the two models on their the out-of-sample predictive performance for individual stock return forecasts. The global model outperforms the sector models in all individual GICS sectors and in the full sample. The global neural network produces a monthly R^2_{OOS} of 3.37% in the full sample and a positive R^2_{OOS} for each of the ten sectors. The sector models achieve a monthly R^2_{OOS} of -6.06% in the full sample, so they underperform a naive forecast of zero for all monthly stock returns. Sector models perform particularly poorly in predicting stock returns in sectors with small sample sizes.

To understand the differences in predictive performance, I examine the importance of each input variable in predicting returns with the neural network models. Variable importance is determined by the reduction in R^2_{OOS} that results from setting all values of a particular signal to zero while holding all other model estimates fixed. The different models share most of their most important variables. Analogous to Blitz et al. (2023), Size is the most influential signal in predicting returns in the global model and in most sector models. The neural networks tend to perform well in out-of-sample return forecasting when their relative importance is skewed towards Size. Sector models with a broader set of influential characteristics underperform out-of-sample.

I compare the profitability of portfolios based on the sorted predictions of the global model with the sector models. At the end of each month, I sort the stocks into decile portfolios and calculate the value-weighted returns of holding the decile portfolios over the next month. A long-short portfolio buys stocks with the highest expected returns and sells those with the lowest. Portfolio strategies based on the out-of-sample predictions of the global model outperform the sector-specific models. A long-short portfolio based on the global model's forecasts achieves an average monthly out-of-sample return of 2.71% and an annualized Sharpe ratio of 2.07. A long-short portfolio based on the forecasts of the sector models generates an out-of-sample monthly return of 0.99% and an annualized Sharpe ratio of 1.14. The outperformance of the global neural network is particularly strong in the early years of the out-of-sample period. During this period, the global model achieves its highest returns while the sector models struggle to remain profitable.

The sector-specific neural networks continue to underperform when compared to simple ordinary least squares (OLS) models. A long-short portfolio based on the sorted predictions of the OLS sector models generates a higher value-weighted return than the sector neural networks. However, the OLS outperformance comes only from the first half of the out-of-sample period when less training data is available. Small sector sample sizes require a trade-off between stable model estimation and capturing complex sector-specific interactions. The global neural network trained on pooled data significantly outperforms the OLS models.

The global neural network shows some out-of-sample sector allocation power. In the cross-section of sectors, it correctly predicts higher relative returns for the most profitable sectors and lower returns for the least profitable sectors. As a result, the returns generated by the global model are lower with sector-neutral portfolios.

Section 2 reviews the recent literature on machine learning for return forecasting and global versus industry-specific models. Section 3 presents the sources of stock data and input signals for return prediction and explains the sector classifications. Section 4 describes the methodology used to train neural networks for return prediction and to construct portfolios based on these predictions. Section 5 presents the results of comparing the forecasting performance of the global model with the sector models. Section 6 concludes.

2. Literature review

The past decades produced a variety of literature focusing on predicting the cross-section of stock returns. Authors explore a variety of variables in linear models, but there is still a lack of consensus regarding which variables are related to expected stock returns. This problem is often referred to as factor zoo. Linear models cannot deal with many variables and their potential nonlinearities and interactions. Therefore, recent research focuses on more complex machine learning models to handle the high dimensionality in the factor zoo.

Early literature focuses on single machine learning models and their ability to outperform traditional methods. Moritz and Zimmermann (2016) propose tree-based conditional portfolio sorts as a machine learning approach. In their models, recent past returns within the last six months predict future returns and outperform linear models like Fama-MacBeth regressions. Excess returns persist even after accounting for transaction costs and common risk factors. Traditional methodologies with linear assumptions fail to capture a nonlinear relationship between past and future returns.

A nonparametric model using adaptive group least absolute shrinkage and selection operation (LASSO) isolates relevant predictors in a high-dimensional setting in Freyberger et al. (2020). A small subset of variables, including size, total volatility, and recent return-based metrics, provide unique predictive power. Their model significantly outperforms linear approaches like those of Lewellen (2015) with higher out-of-sample Sharpe ratios. The nonparametric model selects fewer variables in-sample than the linear models but captures nonlinear interactions.

Gu et al. (2020) are among the first to present a comparative analysis of machine learning models for predicting stock returns. Their models agree on a small set of dominant variables, with price trends, liquidity, and volatility as the most influential predictors. Neural networks perform the best among all machine learning models, and portfolios sorted on neural network return predictions double the

Sharpe ratios of linear models. Shallow neural networks outperform deeper ones due to the limited data and low signal-to-noise ratio in empirical asset pricing. Interactions and nonlinear relations between variables drive the outperformance of machine learning methods. Azevedo and Hoegner (2023) report similar results with tree-based and neural network approaches. Their machine learning methods uncover interaction effects challenging traditional risk-based explanations in asset pricing. Neural network models utilize 299 stock market anomalies to achieve monthly out-of-sample returns 1% higher than a linear benchmark. Linear regressions are easy to interpret but underperform their machine learning models using statistical significances and returns.

Most recent literature uses technically more advanced machine learning strategies. Azevedo et al. (2024) examine the expected returns of deep learning strategies. Their long short-term memory (LSTM) models yield net returns of up to 1.42% per month, even after accounting for the recent era of high liquidity, transaction costs, and post-publication decay. Strategies combining several machine learning models constantly achieve significant returns after cost. Cost mitigation techniques reduce turnover and trading costs but do not improve net performance. L. Chen et al. (2024) combine three deep neural network models with no-arbitrage constraints to estimate asset pricing models for US stocks. Incorporating specific domain knowledge into the technical implementation enhances prediction accuracy and out-of-sample performance. Their deep learning strategies with no-arbitrage constraints outperform other machine learning benchmarks in Sharpe ratio and identify the core variables driving asset prices.

Other research focuses on using machine learning to predict stock returns not only in the US but globally. Tobek and Hronec (2021) aggregate 153 anomalies across global markets into one mispricing signal using machine learning. Their strategy outperforms linear models out-of-sample in various international markets. Extending the training sample with international data does not improve out-of-sample performance for the US market. However, machine learning models trained on US stocks perform well in markets outside the US. Cakici et al. (2023) investigate machine learning's cross-sectional return predictability across 46 global stock markets. Combining predictions from multiple machine learning models delivers robust out-of-sample returns across diverse markets. Developed markets show higher profitability than emerging ones. Firm size and idiosyncratic risk are the most important variables for predictions, with higher returns in smaller firms and markets with more idiosyncratic risk. Azevedo et al. (2023) also focus on the out-of-sample performance of different machine learning models across an international data sample. Neural networks and composite predictors perform the best. These models achieve significant monthly long-short returns of around 2%. Portfolio returns remain significant even after transaction costs and outperform linear benchmark models. Drobetz and Otto (2021) use machine learning strategies to predict European stock returns. Like in the US market, machine learning models outperform traditional linear models by capturing nonlinearities and variable interactions. Neural networks and classification-based approaches perform best and generate significant returns even after transaction costs. Support vector machines, which classify stocks into decile portfolios, deliver even higher returns by eliminating the noise of expected returns at the stock level. Leippold et al. (2022) apply machine learning models to the Chinese stock market. In a market dominated by retail investors, liquidity and volatility indicators have predictive power over traditional variables like valuation ratios. Neural networks perform best, particularly for small-cap and non-state-owned firms. They achieve higher predictability in China than in the US due to distinct asset pricing dynamics driven by local investor behavior. Hanauer and Kalsbach (2023) assess various machine learning models for predicting stock returns in a broad sample of emerging markets. Like in developed markets, their models identify nonlinearities and interactions among variables. Tree-based methods and neural networks deliver superior long-short returns and alphas over linear models. Efficient trading rules ensure machine learning predictions outperform even after transaction costs, short-selling constraints, and limiting the sample to big stocks.

Despite the strong performance of machine learning models for stock market prediction, there are still problems in implementing them in practice. Rasekhschaffe and Jones (2019) focus on mitigating overfitting in machine learning models. Feature engineering and forecast combinations decrease the risk of overfitting. These techniques increase the signal-to-noise ratio and produce more robust predictions. Avramov et al. (2023) criticize high limits-to-arbitrage environments and exclude stocks like microcaps and distressed firms. Machine learning portfolios often rely on long and short positions that are impossible in practice. The profitability of machine learning strategies is reduced when trading costs are considered due to high turnover. The authors propose including trading costs in machine learning models and imposing economic restrictions. Blitz et al. (2023) report the impact of varying prediction horizons in machine learning models for stock return predictions. While one-month forecasts yield high gross returns, the net returns considering transaction costs are minimal after 2004. Machine learning models with longer prediction horizons provide significant net alpha due to reduced portfolio turnover. One-month forecasts rely on short-term price signals, whereas longer horizon predictions rely more on value-oriented signals. Aligning the design of machine learning models with trading horizons enhances profitability through reduced turnover and better after-cost performance.

In this thesis, I focus on sector-specific versus global machine learning models. Therefore, it is appropriate to consider prior research on the relation between industry-specific and market predictions. This includes traditional methods of predicting the cross-section of stock returns, like factor investing and other linear regression models. Kim et al. (2013) enhance the linear residual income model (RIM) with industry-specific factors using the value-to-book (V/B)

ratio. Decomposing V/B into industry-specific components better predicts future abnormal returns and outperforms traditional RIM implementations. Their industry-adjusted RIM model provides superior predictive accuracy for abnormal returns compared to conventional valuation measures. Liu et al. (2014) analyze the predictive power of industry effects in option-implied volatility measures on stock returns. Industry-neutral portfolios outperform full-universe portfolios with higher Sharpe ratios and lower downside risk. Cavaglia et al. (2006) explore region-neutral and industry-neutral value portfolios. Industry-neutral portfolios offer more stable returns and higher information ratios due to lower volatility and less cyclicality. They capture the global value premium more effectively and provide a better risk-return profile.

Modern machine learning models also capture economically meaningful interdependencies among industries. Rapach et al. (2019) examine cross-industry return predictability with machine learning models like LASSO. Due to gradual information diffusion, lagged returns from industries like financial and commodity sectors can predict returns in other sectors. An industry-rotation strategy based on these cross-industry signals outperforms linear methods. Ciner (2019) uses a random forest model as a machine learning strategy to predict market returns with industry returns. Industry returns provide significant out-of-sample predictive power for the market index. Random forests outperform traditional linear models due to their capacity to capture both linear and non-linear dynamics. Industry-level information forecasts market movements in a way that linear models fail to capture.

This thesis extends the existing literature on machine learning in empirical asset pricing by integrating it with industry-specific strategies. Building upon the work of Gu et al. (2020), I use neural networks as well-performing machine learning models to forecast the cross-section of stock returns for the US market. I compare long-short portfolio returns achieved by sector-specific models with those of a global model.

3. Data

3.1. Stock data

My sample includes all US stocks listed on the NYSE, AMEX, and NASDAQ. The sample period runs from July 1963 to December 2022. I source equity returns and other stock market data from CRSP. Accounting data to replicate the Fama and French (2015) five-factor model is from Compustat.

I calculate monthly excess returns as the one-month stock return from CRSP over the risk-free rate provided on Kenneth R. French's homepage.[1] To predict the cross-section of stock returns, I subtract the monthly cross-sectional median excess return from the monthly excess return for each stock. "Returns" throughout this thesis denote these relative monthly stock returns with the market component already removed.

The input variables to the machine learning models come from A. Y. Chen and Zimmermann (2022). I download the August 2023 version of signals from their Open Source Asset Pricing (OSAP) website.[2] This includes 209 firm-level characteristics replicated from the academic asset pricing literature. I use the terms variables, signals, and characteristics interchangeably throughout this thesis.

In addition to the 209 signals from the OSAP data, I define three input variables from CRSP data. Short-term reversal is the prior one-month return, Price is the natural logarithm of the CRSP price data field, and Size is the natural logarithm of the price multiplied by the shares outstanding.

All input variables are signed so that higher values correspond to higher expected returns. Following Gu et al. (2020), Blitz et al. (2023) and other recent literature, I rank all input variables cross-sectionally for each month into the interval of [-1,1]. This helps neural network models deal with varying ranges of values and different variances across the signals during training. Missing values are filled with the monthly cross-sectional median rank.

I follow Hou et al. (2020) and Blitz et al. (2023) and exclude microcaps from my sample to prevent them from driving my results. I define microcaps as all stocks with a monthly market capitalization below the 20th percentile of the NYSE market capitalization in that month.

After excluding microcaps, my full sample from July 1963 to December 2022 includes approximately 1.3 million monthly stock observations with a monthly average of 1842 stocks.

3.2. Sector data

To train sector-specific machine learning models, I assign all stocks in my sample to sectors according to the Global Industry Classification Standard (GICS) from MSCI and S&P. I categorize stocks into ten sectors defined by GICS: Energy, Materials, Industrials, Consumer Discretionary, Consumer Staples, Health Care, Financials, Information Technology, Communication Services, and Utilities. Due to the small number of observations, I include the Real Estate sector in the Financials sector. This corresponds to the GICS classification before 2016, better representing the biggest part of my sample period.

I prefer the GICS over the Standard Industrial Classification (SIC) system often used in other literature on industry specifics. The GICS is the more modern industry taxonomy with a stronger focus on new industries like the computer, software, and information technology sectors.

I source data on SIC industry classification from CRSP and data on GICS sector classification from Compustat. GICS data has weaker coverage than SIC data, particularly from

[1] See http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html (2024).

[2] See https://www.openassetpricing.com/august-2023-data-release/ (2023).

the beginning of my sample until 1985. Therefore, I develop a mapping from SIC industries to GICS sectors described in Table 1. If a monthly stock observation is not assigned to a sector under GICS but to a SIC industry, I fill in the missing GICS classification based on the mapping in Table 1.

The mapping is based on two different inputs. First, I consider overlaps between the SIC and GICS classifications for stock observations with both data points in my sample. For example, 90% of all stocks classified as Financials in SIC are also classified as Financials in GICS. Similarly, over 80% of all stocks classified as utilities in SIC are also classified as Utilities in GICS. Second, I consider the definitions of the industry taxonomy in both classification systems.[3] For example, SIC includes transportation, communications, electric, gas, and sanitary services in one Division. GICS classifies Transportation as an industry group in the sector Industrials, communications belong to the sector Communication Services, and electric, gas, and sanitary services belong to the sector Utilities. This enables me to map the corresponding SIC major groups to the GICS sectors.

Table 1 includes the ten GICS sectors used for sector-specific machine learning models. It summarizes all SIC digit codes and the corresponding industry description mapped to each GICS sector. For example, I map the SIC codes 1200-1399 (Major Group Coal Mining and Major Group Oil And Gas Extraction) and 2900-2999 (Major Group Petroleum Refining And Related Industries) to the GICS sector Energy. The SIC industries mapped to a GICS sector can be either entire Divisions (determined by a capital letter) or more detailed Major Groups (determined by the first two digits) and Industry Groups (determined by the first three digits).

If I recognize no clear relationship between a SIC industry and GICS sectors, then I don't include this SIC industry in the mapping. A stock with missing GICS data and a SIC code not contained in the mapping will not be assigned to a GICS sector. The same happens to stocks with missing data for both classification systems. After applying the mapping, 15,239 out of 1.3 million monthly stock observations in the full sample are not assigned to a GICS sector. They are included in the training data for the global neural network but not in the training data for the sector-specific neural networks.

4. Methodology

4.1. Return prediction using machine learning

My methodology follows Gu et al. (2020) and Hanauer and Kalsbach (2023), with the difference that I train one global model and one sector-specific model for each of the ten GICS sectors.

I aim to predict the cross-section of US stock returns, so I forecast the outperformance of a stock relative to the US stock market. The relative return of a stock is defined as

$$r_{i,t}^{rel} = r_{i,t} - Mkt_t, \tag{1}$$

where $r_{i,t}$ is the excess return of stock i in month t and Mkt_t is the cross-sectional median excess return across all stocks in the sample in month t.

I describe the one-month-ahead relative return of a stock $r_{i,t+1}^{rel}$ as an additive prediction error model:

$$r_{i,t+1}^{rel} = E_t\left[r_{i,t+1}^{rel} \mid x_{i,t}\right] + \epsilon_{i,t+1}. \tag{2}$$

$E_t[r_{i,t+1}^{rel}|x_{i,t}]$ is the conditional expected relative return of stock i in month t for month $t+1$. It is conditional as it depends on $x_{i,t} \in \mathbb{R}^p$, a vector of stock-specific p input variables known at month t. $\epsilon_{i,t+1}$ is the prediction error term.

I estimate the expected relative return with the unknown function $f^*, f^* : \mathbb{R}^p \to \mathbb{R}$. It estimates the expected return depending only on the vector of p stock-specific input variables available in month t:

$$E_t\left[r_{i,t+1}^{rel} \mid x_{i,t}\right] = f^*(x_{i,t}). \tag{3}$$

In the case of neural networks, the unknown function $f^*(x)$ is approximated by a nonlinear function $f(x, \theta, \rho)$. This function is parametrized by a vector of coefficients θ and a set of hyperparameters ρ. When training neural networks, the coefficients θ are estimated from the training data with respect to the hyperparameters ρ and a predefined loss function L. The hyperparameters ρ are optimized concerning the loss function L based on the estimated coefficients θ and available data.

Neural networks as a form of supervised machine learning outperform linear models in prior literature (Azevedo & Hoegner, 2023; Gu et al., 2020). Therefore, I choose three-layer neural networks as the machine learning model for this thesis. Appendix A4 describes the model architecture and hyperparameters used to train the neural networks. In addition, I later use ordinary least squares (OLS) models as a benchmark for sector-specific neural networks.

The global neural network model takes the full sample of 1.3 million monthly stock observations as input. The sector-specific machine learning models for the ten GICS sectors take all monthly stock observations assigned to the respective GICS sector as input. Therefore, the samples used to train the sector-specific neural networks differ significantly in their number of observations, depending on the sector's size.

To avoid data leakage, I divide all input samples into three disjoint time periods, which always keep the temporal ordering of the data: the training, validation, and testing samples. First, I estimate the neural network coefficients for a range of hyperparameter values on the training sample. The validation sample compares the loss function results for each set of hyperparameters based on the estimated model from the

[3] See https://www.msci.com/our-solutions/indexes/gics for GICS definitions and https://www.osha.gov/data/sic-manual for SIC definitions.

Table 1: Mapping of SIC to GICS industries

This table maps industries classified under the Standard Industrial Classification (SIC) system to sectors classified under the Global Industry Classification Standard (GICS) system used in this thesis. The first column contains the 11 different sectors of the GICS system. Throughout this thesis, the 11th sector, Real Estate, is not considered separately, but is included in the Financials sector. The second and third columns contain the SIC industries mapped to the GICS sector in the first column. The second column contains the SIC digit code, and the third includes the corresponding industry description. The SIC industries mapped to a GICS sector can be either entire Divisions (determined by a capital letter) or more detailed Major Groups (determined by the first two digits) and Industry Groups (determined by the first three digits). This mapping is used to classify individual stocks into sectors when GICS sector information is unavailable in Compustat. If a SIC industry classification is available, the GICS sector is added according to this mapping. If no SIC classification is available either, the GICS sector value is 'Missing', and the stock is not included in the training data for the sector models.

GICS Sector	SIC Code	SIC Description
10 - Energy	1200–1399	Coal Mining and Oil/Gas Extraction
	2900–2999	Petroleum Refining and Related Industries
15 - Materials	Division B 1000–1499 (excluding 1200–1399)	Mining (excluding Coal Mining and Oil/Gas Extraction)
	2400–2499	Lumber and Wood Products, Except Furniture
	2600–2699	Paper and Allied Products
	3300–3399	Primary Metal Industries
20 - Industrials	Division C 1500–1799	Construction
	Division E 4000–4999 (excluding 4800–4999)	Transportation (excluding Communications and Utilities)
	Division J 9100–9999 (excluding 9900–9999)	Public Administration (excluding Nonclassifiable Establishments)
	3400–3499	Fabricated Metal Products, Except Machinery and Transportation Equipment
	3500–3599	Industrial and Commercial Machinery
	7320–7329	Credit Reporting and Collection
	7340–7349	Services to Dwellings and Other Buildings
	7360–7369	Personnel Supply Services
	7390–7399	Miscellaneous Business Services
	7500–7599	Automotive Repair Services and Parking
	7600–7699	Miscellaneous Repair Services
	8710–8719	Engineering Architectural and Surveying Services
	8740–8749	Management and Public Relations
	8900–8999	Services Not Elsewhere Classified
25 - Consumer Discretionary	Division G 5200–5999 (excluding 5400–5499)	Retail Trade (excluding Food Stores)
	Division F 5000–5199 (excluding 5140–5189, 5180–5189)	Wholesale Trade (excluding Groceries and Beer, Wine, and Distilled Alcoholic Beverages)
	1500–1599	Building Construction General Contractors and Operative Builders
	2200–2299	Textile Mill Products
	2300–2399	Apparel and Other Finished Products Made from Fabrics and Similar Materials
	2500–2599	Furniture and Fixtures
	3100–3199	Leather and Leather Products

Table 1 — continued

GICS Sector	SIC Code	SIC Description
25 - Consumer Discretionary	3900–3999	Miscellaneous Manufacturing Industries
	7000–7099	Hotels, Rooming Houses, Camps, and Other Lodging Places
	7200–7299	Personal Services
	7800–7899	Motion Pictures
	7900–7999	Amusement and Recreation Services
30 - Consumer Staples	Division A 0100–0999	Agriculture, Forestry, and Fishing
	2000–2199	Food Products and Tobacco Products
	5140–5159	Wholesale Trade - Groceries
	5180–5189	Wholesale Trade - Beer, Wine, and Distilled Alcoholic Beverages
	5400–5499	Food Stores
35 - Health Care	2800–2899	Chemicals and Allied Products (including Drugs)
	3840–3849	Surgical Medical and Dental Instruments and Supplies
	3850–3859	Ophthalmic Goods
	8000–8099	Health Services
	8300–8399	Social Services
	8730–8739	Research Development and Testing Services
	9900–9999	Nonclassifiable Establishments
40 - Financials	Division H 6000–6799	Finance, Insurance, and Real Estate
45 - Information Technology	3570–3579	Computer and Office Equipment
	3600–3699	Electronic and Other Electrical Equipment and Components
	3820–3829	Laboratory Apparatus and Analytical Optical Measuring and Controlling Instruments
	7370–7379	Computer Programming Data Processing
50 - Communication Services	4800–4899	Communications
55 - Utilities	4900–4999	Electric, Gas, and Sanitary Services
60 - Real Estate	Included in GICS sector 40 - Financials	

training sample. The optimal hyperparameter set minimizes the loss function on the validation sample and is then used to retrain five different neural networks on the training sample. I use these five models to predict the monthly returns for the test sample. The final prediction for each stock is the average over the five individual model predictions to reduce the variance in single forecasts.

Following the training approach as in Gu et al. (2020) and Blitz et al. (2023), I retrain the models once at the end of every year but predict every month using the latest model and data. The first 18 years of my sample (July 1963 to December 1981) are the first training sample, and the next 12 years (January 1982 to December 1993) the first validation sample. The first one-year test sample is the following 12 months, so the first out-of-sample (OOS) prediction is made for January 1994. To predict the monthly returns from January 1995 to December 1995, I extend the training sample by one year (July 1963 to December 1982) and roll forward the validation sample by one year (January 1983 to December 1994). I repeat this procedure for each year in my sample. No future information is leaked from a previous period.

To evaluate the predictive performance for individual stock return forecasts on the test sample, I use the pooled out-of-sample R^2_{OOS} defined by Gu et al. (2020):

$$R^2_{OOS} = 1 - \frac{\sum_t^T \sum_i^N \left(r^{rel}_{i,t} - \hat{r}^{rel}_{i,t}\right)^2}{\sum_t^T \sum_i^N \left(r^{rel}_{i,t}\right)^2}. \tag{4}$$

This metric compares the out-of-sample forecasts with a naive forecast of zero, better suited to individual return predictions than the typical forecast with mean returns.

4.2. Variable importance

As a primary measure to interpret the results of the machine learning models, I rank the respective input variables according to their variable importance. This aims to identify characteristics that influence the cross-section of expected returns. Following Gu et al. (2020) and Blitz et al. (2023), I define variable importance as the reduction in panel predictive out-of-sample R^2_{OOS}. For each annually trained neural network, I iteratively set the values for each input variable to zero while holding the model estimates fixed. In each iteration, I predict new monthly returns for the respective one-year test sample and calculate the change in out-of-sample R^2_{OOS}. I use the average variable importance across each annually trained model to rank each feature where rank one is the most important characteristic. To determine the relative importance of individual variables for the performance of each model, I normalize variable importance within a model to sum to one. If setting a signal to zero increases the panel predictive out-of-sample R^2_{OOS}, the variable importance measure for that signal is negative. Therefore, the normalized variable importance of other signals within a model can be greater than 1.

4.3. Machine learning portfolios

Portfolio performance is my primary metric for evaluating the forecast performance of machine learning models. At the end of each month, each model produces a prediction of a stock's next month's relative return $\hat{r}^{rel}_{i,t+1}$. Based on these forecasts, I sort stocks from highest to lowest predicted return and assign them into decile portfolios using NYSE breakpoints.[4] I reassign and rebalance portfolios at the end of each month. I compute value-weighted returns from holding the decile portfolios over the next month to avoid small stocks driving the results. Finally, I construct a zero-net investment (long-short) portfolio that goes long in the highest decile portfolio (decile 10) and short in the lowest decile portfolio (decile 1).

I evaluate the predictive performance of three different machine learning strategies. The first strategy forms decile portfolios based on the predictions of the global neural network model. The second strategy forms decile portfolios with sector-neutral portfolio sorts based on the global neural network model predictions. This means each sector gets individual breakpoints for the decile sorts. The third strategy forms decile portfolios based on the predictions from the sector-specific neural networks. First, I perform portfolio sorts for each sector individually based on the respective model forecasts. Then, I combine the sector-specific decile portfolios into single decile portfolios. For example, the top decile portfolio for month t contains all stocks from the ten top decile portfolios across all sectors in month t.

[4] Breakpoints for the decile sorts are first determined using only stocks listed on the NYSE. All stocks are then sorted into decile portfolios based on these breakpoints, regardless of which exchange they are listed on. As the NYSE contains stocks with larger average market capitalizations, this reduces the influence of small stocks on the portfolio sorts.

To compare the results of the three neural network portfolio sorts, I provide each decile portfolio's average predicted returns, realized returns, and Sharpe ratios. I compute mean returns and associated t-statistics for the long-short portfolios of each machine learning strategy. To benchmark the long-short returns, I consider the adjusted R^2-value and alphas from the Capital Asset Pricing Model (CAPM) and Fama and French (2015) five-factor model with their associated Newey and West (1987) adjusted t-statistics using six lags. The factors are based on the same sample of 1.3 million monthly stock observations as the neural network portfolios.

Finally, I benchmark the predictive performance of the sector-specific neural networks with OLS models in Section 5.4. The portfolio sorts work similarly but are based on predictions from sector-specific OLS models. All sector models mentioned outside of Section 5.4 always refer to sector-specific neural networks.

5. Empirical results

5.1. Prediction performance

The global neural network dominates the sector-specific neural networks in out-of-sample predictive performance for individual stock return forecasts. On the full sample over the out-of-sample period from January 1994 to December 2022, the global model achieves a monthly R^2_{OOS} of 3.37%. The sector models achieve a monthly R^2_{OOS} of -6.06%, so they underperform a naive forecast of zero for all monthly stock returns. Across all ten GICS sectors, the global model outperforms the respective sector model. Sector models perform particularly worse for sectors with only a small sample size.

Table 2 compares the monthly out-of-sample stock-level prediction performance across all ten GICS sectors between the global model and the ten sector-specific models. The sector performance for the global model is determined by filtering the out-of-sample predictions of the global neural network for the respective sector stocks. In addition, Table 2 includes the average monthly observations per sector over the full sample period from July 1963 to December 2022. This demonstrates the sample size for each sector.

The global model produces positive R^2_{OOS} statistics across all individual sectors. Taking the full sample of stocks, the global model achieves a R^2_{OOS} of 3.37%. Therefore, the predictions consistently outperform a naive forecast of zero to all stocks in all months over the out-of-sample period. Except for Utilities, the R^2_{OOS} statistics for all sectors are larger than 2%. Over six out of ten sectors the global model produces R^2_{OOS} above 3%, with the highest value at 4.57% for the sector Health Care. Utilities appear to be an outlier with 0.11%, more than an order of magnitude smaller than the R^2_{OOS} for all other sectors. There is no correlation between the sectors' sample size and the predictive performance of the global neural network. The model achieves a R^2_{OOS} of 3.11% on Communication Services, the sector with the smallest sample size and only an average of 58 stocks per month in the sample. This is more than the 2.26% for Financials, the biggest sector

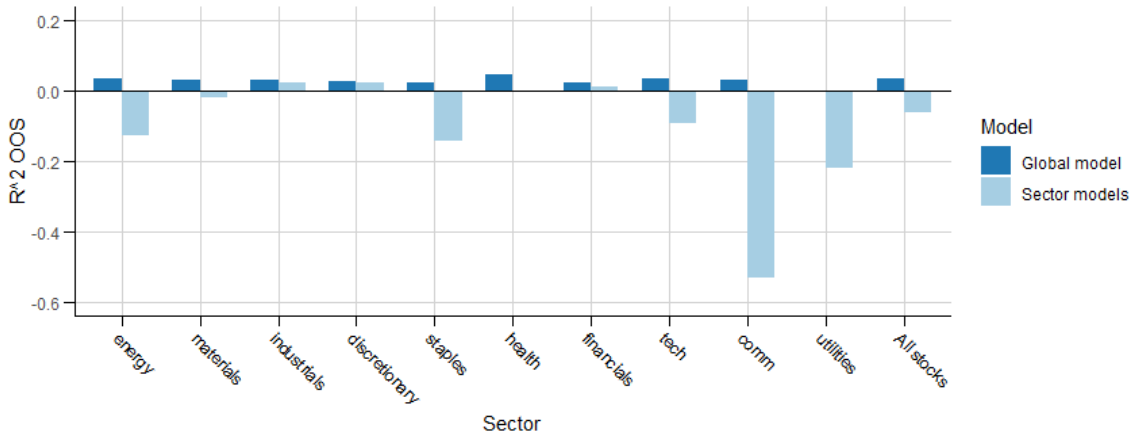
Table 2: Monthly out-of-sample stock-level prediction performance

This table summarizes the monthly out-of-sample stock-level prediction performance across all GICS sectors using the global model and the respective neural network sector models. Panel A reports the monthly R^2_{OOS} statistics of the global model and the respective neural network sector models for the full sample over the out-of-sample period from January 1994 to December 2022. The third column in Panel A reports the average monthly stocks per sector in the full sample over the full sample period from July 1963 to December 2022. The sample consists of US CRSP stocks, excluding microcap stocks with a market capitalization smaller than the 20th percentile of stocks listed on the NYSE. Panel B provides a visual comparison of the monthly R^2_{OOS} statistics in Panel A.

Panel A: Percentage R^2_{OOS}

Sector	Global Model	Sector Models	Avg. monthly observations
Energy	3.49	-12.79	107
Materials	3.23	-1.92	136
Industrials	3.07	2.21	293
Consumer Discretionary	2.86	2.17	286
Consumer Staples	2.47	-14.11	113
Health Care	4.57	-0.36	184
Financials	2.26	1.18	306
Information Technology	3.43	-9.17	223
Communication Services	3.11	-53.07	58
Utilities	0.11	-21.85	113
All stocks	3.37	-6.06	1842

Panel B: Visual comparison of R^2_{OOS} statistics



with a mean of 306 monthly stock observations in the sample.

The sector-specific neural networks perform significantly worse than the global model across the whole sample of stocks. Sector models produce negative R^2_{OOS} in seven out of ten sectors. The R^2_{OOS} statistic on the full sample is also negative with -6.06%. This means a naive forecast of zero to all stocks in all months dominates the sector models over the full sample. The sector neural networks achieve positive R^2_{OOS} only in the sectors Industrials, Consumer Discretionary and Financials. These are the three biggest sectors by sample size, with around 300 stock observations per month. The models perform poorly in all sectors with only a small sample size. Sectors like Energy, Consumer Staples, Communication Services, and Utilities are the smallest sectors, with monthly stock observations in the range from 58 to 113, and all produce two-digit negative R^2_{OOS}. Communication Services has only half of the sample size compared to the next bigger sector, and the respective sector model performs by far the worst with a R^2_{OOS} of -53.07%. Information Technology is an exception with 223 average monthly stock observations and a negative R^2_{OOS} of -9.17

Based on the correlation between sector size and sector model performance, complex nonlinear machine learning models like neural networks might not be suited for small samples. The number of coefficients estimated in a neural network rapidly expands with 212 input signals. Therefore, the observations-to-parameters ratio is particularly small in sectors with a small sample size. Pooling all sectors as input for the global neural network improves the observations-to-

parameters ratio. This motivates benchmarking the sector neural networks against sector OLS models in Section 5.4.

Panel B of Table 2 visualizes the negative outliers in the predictive performance of the sector models. The global neural network achieves a constant positive R^2_{OOS} in a similar range over the whole sample. The sector neural networks never dominate the global model and produce apparent downward deviations that stand out from the range of other values.

5.2. Variable importance

The neural network models with the best out-of-sample prediction performance share most of their most important variables. Size is the most influential signal in predicting returns in the global model and across most sector models. Neural networks tend to perform well in out-of-sample return forecasts if their relative importance is skewed towards Size. Risk measures, liquidity variables, and recent price trends are other important characteristics in well-performing models. Sector models with a broader set of influential characteristics underperform out-of-sample.

I examine the importance of individual input signals in predicting returns with neural network models. To determine variable importance, I calculate the reduction in R^2_{OOS} from setting all values of a particular signal to zero while keeping all other model estimates fixed. This process is repeated iteratively over all variables for each model. The average over all annual test samples results in a single importance measure per model for each variable.

Figure 1 visualizes each neural network model's variable importance of the top 20 stock-level signals. This includes the global model and sector-specific models for each of the ten GICS sectors. Variable importances within a model are normalized to sum to one, giving them the interpretation of relative importance for that model.

Figure 2 reports overall rankings of variable importance across the 100 most influential characteristics for all neural network models. I rank the importance of individual signals for each model (with rank one as the most important) and then sum up their ranks. Variables in Figure 2 are ordered so that the highest total ranks are at the top, and the lowest-ranking characteristics are at the bottom. A darker color in a column indicates a higher importance of the respective variable for the individual model.

The most essential variables are similar among many of the neural networks. Analogous to Blitz et al. (2023), Size is the most influential variable overall. It is first in overall ranks and has the highest relative importance for the predictive performance of nine models. It dominates all other variables by a large margin. Except for Consumer Staples (*staples*), the relative importance is highly skewed towards Size in all models where it is the most influential variable. Sector models for Communication Services (*comm*) and Utilities (*utilities*) are the only ones without Size as the most important signal. They are more democratic and draw information from a broader set of predictive variables. However, they underperform all other models in predictive performance, as seen in Table 2. Neural networks tend to perform better out-of-sample in Table 2 when the relative importance in Figure 1 is skewed towards Size.

The best-performing sector models like Industrials (*industrials*), Consumer Discretionary (*discretionary*), and Financials (*financials*) focus on the same set of influential variables as the global model (global). Besides the most important signal Size, this includes Idiosyncratic risk (*IdioVolAHT, IdioVol3F*) and Realized Volatility (*RealizedVol*) as risk measures; Amihud (2002) illiquidity (*Illiquidity*) and Volume Variance (*VolSD*) as liquidity variables; Short-term Reversal (*STreversal*) and Trand Factor (*TrendFactor*) as recent price trends; and to a lesser extent valuation ratios and fundamental signals like earnings-to-price (*EP*) and sales-to-price (*SP*).

Looking at Figure 2, the highest-ranked characteristics are the ones that achieve solid relative importance across all models. The four top-ranked signals are Size, Institutional Ownership and Idiosyncratic Volatility from Nagel (2005) (*RIO_Volatility*), Short-term Reversal, and Price. They are among the top-ranked variables in many individual models. However, the following top 25 ranked characteristics do not belong to top-ranked variables in any particular model but instead stay at least relatively important throughout all models. Forecasting power for the rest of the 100 most influential variables in Figure 2 is more heterogeneously distributed. The signals achieve high importance for the predictive power of some individual models but are not relevant in others. For example, risk measures like Idiosyncratic risk and Realized Volatility are among the top signals in the global model and parts of the sector models but irrelevant to the R^2_{OOS} of other sector models.

5.3. Neural network portfolios

Portfolio strategies based on the global model's sorted out-of-sample predictions outperform the sector-specific models. A long-short portfolio based on the global model's forecasts achieves an average monthly out-of-sample return of 2.71% and an annualized Sharpe ratio of 2.07. The portfolio strategy is profitable throughout the out-of-sample period and generates significant alphas over established factor models. The outperformance of the global neural network is particularly strong from the beginning of the out-of-sample period through to the early 2000s. The sector neural networks predict very extreme portfolio returns but achieve significantly lower returns than the global model. A long-short portfolio based on the forecasts of the sector models has an out-of-sample monthly return of 0.99% and an annualized Sharpe ratio of 1.14. The portfolio strategy generates small alphas on top of the factor models but performs unreliably in the first half of the out-of-sample period and only starts to recover thereafter.

After comparing the predictive ability of the global model versus the sector models for individual stock returns in Section 5.1, I assess the profitability of portfolios based on the sorted predictions from my neural networks trained to forecast out-of-sample returns. Each neural network produces

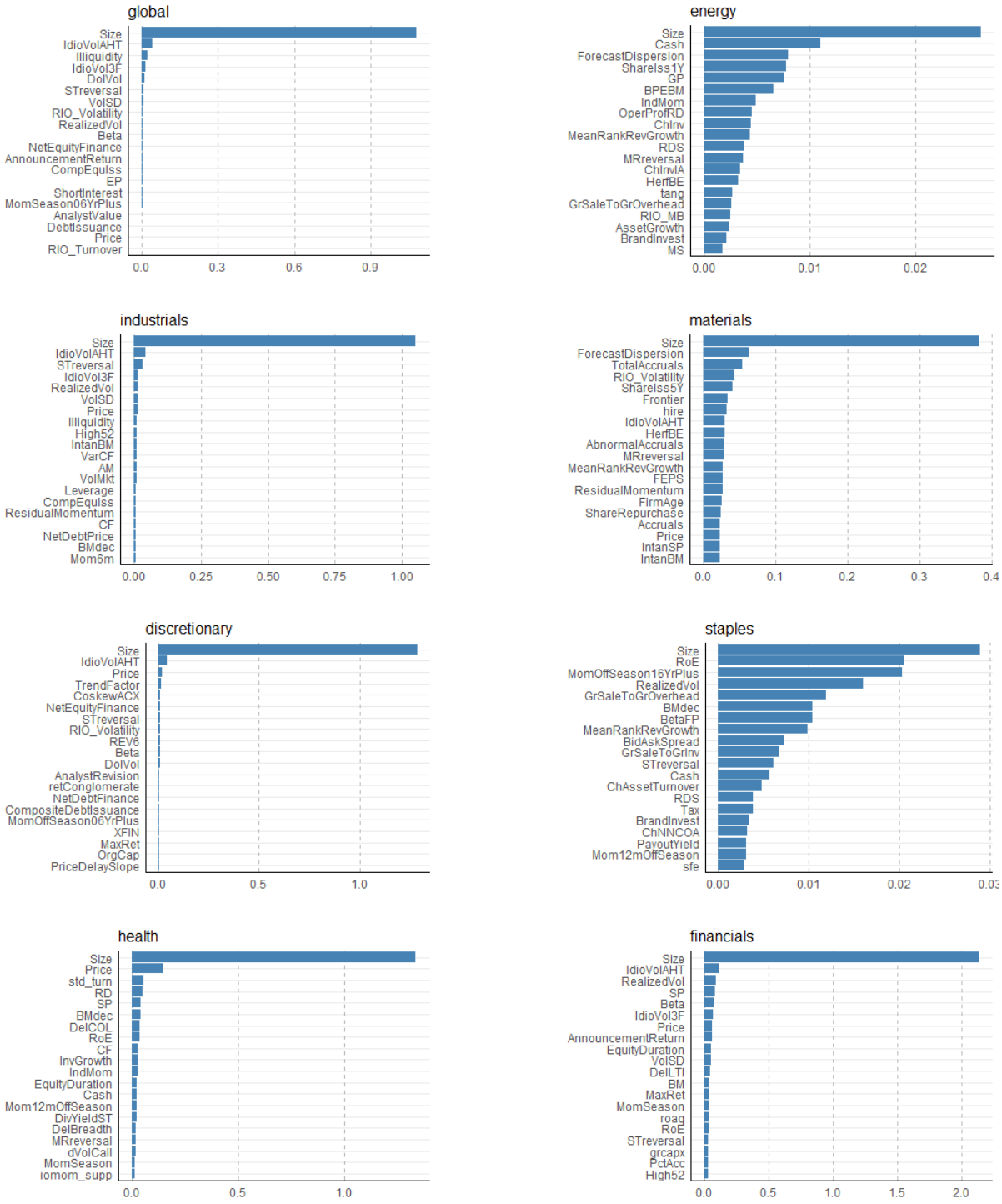
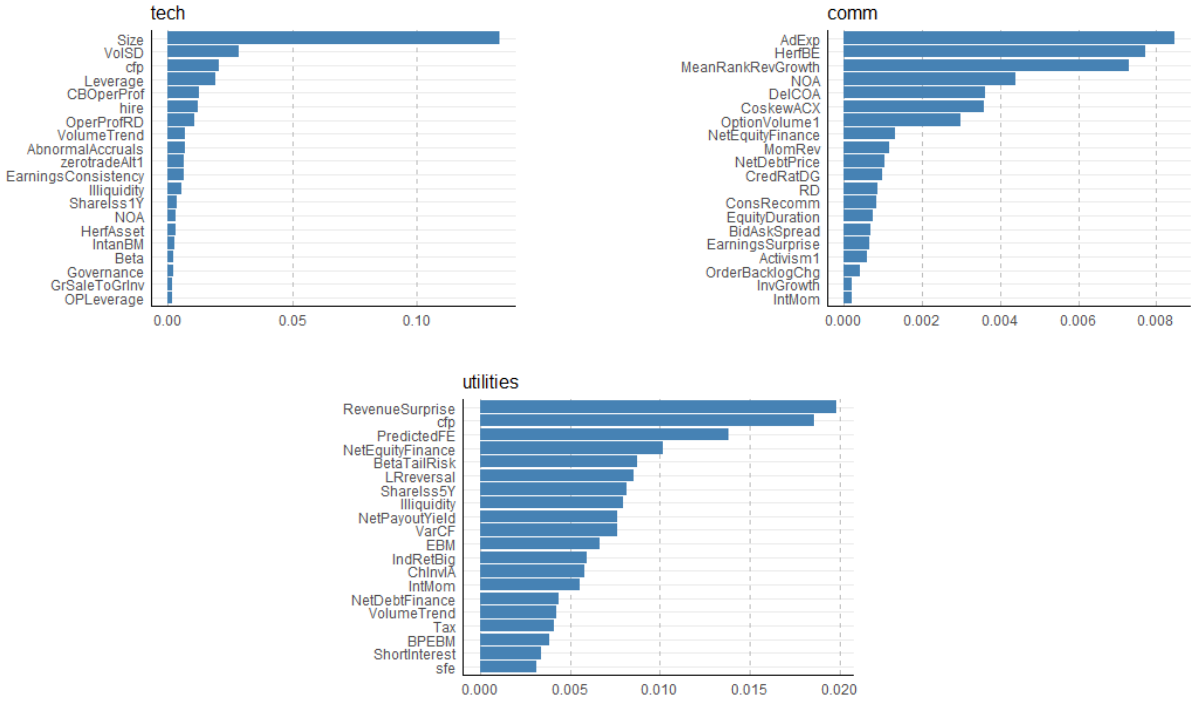


Figure 1: Variable importance by model

This figure plots the variable importance for each model's top 20 most influential characteristics. This includes the global model (*global*) and all ten sector models (GICS sectors Energy (*energy*), Materials (*materials*), Industrials (*industrials*), Consumer Discretionary (*discretionary*), Consumer Staples (*staples*), Health Care (*health*), Financials (*financials*), Information Technology (*tech*), Communication Services (*comm*) and Utilities (*utilities*)). Variable importance is an average over all training samples. Variable importances within a model are normalized to sum to one, giving them the interpretation of relative importance for that model. As variable importances within a model can be negative, the normalized variable importance for other characteristics within that model can be greater than one. The sample consists of US CRSP stocks, excluding microcap stocks with a market capitalization smaller than the 20th percentile of stocks listed on the NYSE. The sample runs from January 1994 to December 2022.

Figure 1 — continued



one-month-ahead stock return predictions at the end of each month. Based on each model's forecasts, I sort stocks into decile portfolios using NYSE breakpoints and reassign the portfolios each month. I compute the value-weighted relative returns of holding the decile portfolios over the next month. In the end, I construct a long-short portfolio with zero net investment. The long-short portfolio buys the highest expected return stocks (decile 10) and sells the lowest (decile 1).

I evaluate three portfolio strategies to compare the profitability of the global model versus the sector models. The first sorts stocks into decile portfolios based on the global model's relative return forecasts. The second strategy performs sector-neutral portfolio sorts on the global model's return predictions. The third strategy produces decile portfolios based on the forecasts of the ten sector models.

Table 3 reports the out-of-sample performance of all value-weighted decile portfolios for the three neural network strategies. The results align with the out-of-sample predictive performance of the models on individual stock-level in Section 5.1. The global model in Panel A again dominates the sector models in Panel C in out-of-sample portfolio performance.

For the global model in Panel A, the average realized relative returns (*Avg*) generally increase monotonically throughout all decile portfolios from lowest (decile 1) to highest (decile 10). Only the difference between decile nine and the top decile is significantly larger, as the return is more than triple. The lowest decile portfolio earns a monthly value-weighted return of -0.97%, and the highest decile portfolio earns 1.73%. The long-short portfolio (*H-L*) based on the global neural network forecasts returns on average 2.71% per month (32.52% on an annualized basis). Its monthly volatility (*Std*) is 4.52% (15.66% annualized), resulting in an annualized out-of-sample Sharpe ratio (*SR*) of 2.07.

The global model with sector-neutral portfolio sorts in Panel B produces similar results but is slightly less profitable. Realized returns increase monotonically throughout all decile portfolios from lowest to highest, with a larger gap between decile nine and the highest decile. The most extreme deciles produce lower average returns than the global model with unrestricted portfolio sorts in Panel A but also experience lower monthly volatility. The lowest decile portfolio earns a monthly value-weighted return of -0.82%, and the highest decile portfolio earns 1.26%. The long-short portfolio based on the global model with sector-neutral portfolio sorts returns on average 2.08% per month (24.96% on an annualized basis). Its monthly volatility is 3.74% (12.96% annualized), resulting in an annualized out-of-sample Sharpe ratio of 1.93.

Analogous to the predictive performance of the sector neural networks out-of-sample, the portfolio strategy based on the sector models in Panel C is significantly less profitable than the global model. Realized relative returns no longer monotonically increase throughout all decile portfolios from lowest to highest. The highest decile portfolios stand out. The ninth decile should contain the stocks with the second-highest expected returns, but it generates a negative average monthly return of -0.08%. The top decile portfolio earns 0.53%, less than a third of the global model's top decile return. The lowest decile portfolio earns a monthly value-weighted return of -0.46%, half of the global model in Panel

Figure 2: Ranked variable importance across models

This figure presents the ranked variable importance for the 100 most influential characteristics across all models. Characteristics are ordered based on the sum of their ranks over all models, with the most influential characteristics on top and the least influential on bottom. Columns correspond to individual models, and color gradients within each column indicate the model's most influential (dark blue) to least influential (white) characteristics. The sample consists of US CRSP stocks, excluding microcap stocks with a market capitalization smaller than the 20th percentile of stocks listed on the NYSE. The sample runs from January 1994 to December 2022.

Table 3: Performance of neural network portfolios

This table summarizes the out-of-sample performance of decile portfolios based on three different neural network strategies. All stocks are sorted into decile portfolios based on their predicted relative returns (with the market component removed) for the next month. The table reports value-weighted returns for all decile portfolios and a zero-net-investment portfolio (*H-L*) that buys the highest expected return stocks (decile 10) and sells the lowest (decile 1). For each portfolio, the table presents the average predicted monthly returns in percentages (*Pred*), the average realized monthly returns in percentages (*Avg*), their standard deviations in percentages (*Std*), and annualized Sharpe ratios (*SR*). Panel A reports results for the global model. Panel B reports results for decile portfolios based on the global model's return predictions but with sector-neutral portfolio sorts. Panel C reports results for decile portfolios based on return predictions from the sector models. The sample consists of US CRSP stocks, excluding microcap stocks with a market capitalization smaller than the 20th percentile of stocks listed on the NYSE. The sample runs from January 1994 to December 2022.

Panel A: Global model											
	Low	2	3	4	5	6	7	8	9	High	*H–L*
Pred	-1.22	-0.57	-0.27	-0.03	0.18	0.39	0.63	0.89	1.27	2.69	3.91
Avg	-0.97	-0.44	-0.25	-0.08	-0.02	0.04	0.14	0.22	0.51	1.73	2.71
Std	2.85	2.13	2.25	2.10	2.26	2.17	2.32	2.84	3.20	3.35	4.52
SR	-1.18	-0.71	-0.39	-0.13	-0.02	0.07	0.20	0.26	0.55	1.79	2.07
Panel B: Global model with sector-neutral portfolio sorts											
	Low	2	3	4	5	6	7	8	9	High	*H–L*
Pred	-1.11	-0.53	-0.26	-0.04	0.16	0.36	0.57	0.83	1.15	2.30	3.41
Avg	-0.82	-0.50	-0.25	-0.21	0.00	0.07	0.11	0.21	0.39	1.26	2.08
Std	2.45	2.02	2.06	2.11	2.35	2.12	2.53	2.70	2.88	3.00	3.74
SR	-1.16	-0.85	-0.42	-0.34	0.01	0.11	0.15	0.26	0.47	1.45	1.93
Panel C: Sector models											
	Low	2	3	4	5	6	7	8	9	High	*H–L*
Pred	-2.85	-1.32	-0.64	-0.08	0.42	0.89	1.43	2.21	3.33	6.29	9.14
Avg	-0.46	-0.25	-0.15	-0.03	-0.04	-0.13	0.00	0.23	-0.08	0.53	0.99
Std	2.49	2.19	2.12	2.18	2.14	2.29	2.36	2.27	2.52	2.61	3.00
SR	-0.64	-0.40	-0.25	-0.04	-0.07	-0.20	0.00	0.34	-0.12	0.70	1.14

A. The long-short portfolio based on the sector neural networks returns on average 0.99% per month (11.88% on an annualized basis). Its monthly volatility is 3.00% (10.39% annualized), resulting in an annualized out-of-sample Sharpe ratio of 1.14.

Across the three strategies, the predicted out-of-sample returns (*Pred*) do not match the realized returns (*Avg*) very closely, except for the pattern of increasing returns throughout the deciles. The neural networks produce extreme return forecasts for the lowest and highest decile portfolios and overstate realized returns. This is particularly true for the sector models in Panel C. Their predicted monthly long-short (*H-L*) portfolio return based on neural network forecasts amounts to 9.14%, almost an order of magnitude larger than the realized return of 0.99%. The effect is less pronounced in the global model in Panel A. It predicted a monthly return of 3.91% for the long-short portfolio, compared to a realized return of 2.71%. The overly extreme portfolio return predictions by the sector models can be caused by overfitting. As demonstrated by the results in Section 5.1, the complex neural networks struggle with sector-specific out-of-sample predictions on small sample sizes and low observations-to-parameter ratios.

Table 4 reports further statistics on the out-of-sample performance of the three neural network strategies, focusing on the long-short portfolio returns. Panel A compares long-short portfolio returns across different out-of-sample periods. The average monthly returns over the full out-of-sample (OOS) period from January 1994 to December 2022 are the same as reported in Table 3. I further show mean returns over two subsamples: the first half of the out-of-sample period from January 1994 to December 2008 and the second half from January 2009 to December 2022. Panel A additionally reports associated t-statistics (*t-stat*).

The global model achieves statistically significant mean returns over the entire out-of-sample period and both subsamples. Long-short portfolios for the global model with unrestricted portfolio sorts outperform sector-neutral portfolio sorts over both subsamples. More importantly, the global neural network dominates the sector-specific models in all samples, both in terms of mean return and risk-return profile based on the associated t-statistics. However, its outperfor-

Table 4: Statistics of neural network long-short portfolios

This table summarizes the out-of-sample statistics of the value-weighted long-short portfolios formed from different neural network model return predictions. All stocks are sorted into decile portfolios based on their predicted relative returns (with the market component removed) for the next month. A long-short portfolio buys the highest expected return stocks (decile 10) and sells the lowest (decile 1). Results are reported for the global model, the global model with sector-neutral portfolio sorts, and the sector models. Panel A presents the average value-weighted monthly full sample mean return and average monthly sub-sample mean returns with associated t-statistics (*t-stat*). Panel B reports the average CAPM alphas and average Fama and French (2015) five-factor model (FF5) alphas, corresponding Newey and West (1987) adjusted t-statistics with six lags ($t\text{-}stat_\alpha$), and corresponding R^2. The sample consists of US CRSP stocks, excluding microcap stocks with a market capitalization smaller than the 20th percentile of stocks listed on the NYSE. The sample runs from January 1994 to December 2022.

	Global model	Global model + sector-neutral sorts	Sector models
Panel A: Percentage returns			
Mean 1994–2022	2.71	2.08	0.99
t-stat	*11.16*	*10.36*	*6.14*
Mean 1994–2008	3.24	2.66	0.70
t-stat	*9.55*	*9.08*	*3.06*
Mean 2009–2022	2.14	1.46	1.29
t-stat	*6.24*	*5.49*	*5.80*
Panel B: Risk-adjusted performance			
CAPM alpha (%)	2.57	2.00	0.98
$t\text{-}stat_\alpha$	*8.80*	*9.55*	*5.78*
R^2	*0.03*	*0.02*	*-0.003*
FF5 alpha (%)	2.44	1.90	0.85
$t\text{-}stat_\alpha$	*8.80*	*8.78*	*5.98*
R^2	*0.14*	*0.12*	*0.04*

mance is significantly more substantial in the first half of the out-of-sample period, when it earns more than four times as much as the sector neural networks.

The global model performs better from 1994 to 2008 with a 3.24% monthly long-short return and an associated t-statistic of 9.55 compared to a 2.14% return from 2009 to 2022 with an associated t-statistic of 6.24. This is analogous to the results of Blitz et al. (2023), who find a weaker out-of-sample performance for machine learning models in their later subsample after 2004. They base this result partly on the strength of the Size factor over different time periods. Size is the most important predictor for returns in my global neural network. According to Blitz et al. (2023), the stronger performance of the global model in the earlier subsample can be attributed to the excellent performance of the Size factor in earlier periods. The Size factor starts to perform worse in the last 20 years, which weakens the profitability of the global neural network in the second subsample.

The long-short portfolio based on the sorted predictions of the sector models underperforms the global model, particularly in the first half of the out-of-sample period. It achieves a monthly mean return of 0.7% with an associated t-statistic of 3.06. Although still less profitable than the global model, the sector-specific neural networks improve in the second subsample with an average return of 1.29%. Associated t-statistics are not far apart, with 5.80 for the sector models and 6.24 for the global model. Therefore, both strategies are similar in risk-return profile over the second half of the out-of-sample period.

Two factors can drive the improved performance of the sector models in the later period of the sample. First, not all sector neural networks rely on Size as an essential signal to the same extent as the global model. Therefore, their profitability is not as dependent on the performance of the Size factor. Second, in the earlier part of the sample, the training samples for the sector-specific neural networks are tiny, with a very low observations-to-parameters ratio. This makes it more challenging to estimate coefficients in a neural network without overfitting on the training data.

Table A1 in the Appendix reports the out-of-sample performance of individual GICS sector long-short portfolios based on the sorted predictions from the ten sector-specific neural networks.

Panel B of Table 4 summarizes the risk-adjusted performance of the long-short portfolios for each strategy based on factor pricing models. I report alphas on top of the Capital Asset Pricing Model (CAPM) and the Fama and French (2015) five-factor model (FF5) with associated Newey and

West (1987) adjusted t-statistics with six lags ($t\text{-}stat_\alpha$) and adjusted R^2 with respect to each factor model.

The risk-adjusted performance yields similar results as the raw long-short returns with a superior global model. All three neural network strategies achieve statistically significant alphas with t-statistics ranging from 5.78 for the sector models on the CAPM to 9.55 for the global model with sector-neutral portfolio sorts on the CAPM. The global model produces the highest alphas, with 2.57% on the CAPM and 2.44% on FF5. The sector-neutral portfolio sorts slightly lower the alphas of the global model to 2.00% on top of the CAPM and 1.90% on top of FF5. The sector-specific models span the lowest alphas, with 0.98% on the CAPM and 0.85% on FF5. The CAPM barely has any explanatory power on the average long-short returns of neural network forecasts, with R^2 never exceeding 0.03 for the global model. The five-factor model explains as much as 14% of the variation in the long-short portfolio based on the global model's forecasts. Unsurprisingly, the Size factor is the statistically most significant factor in regressions of portfolio returns on the five-factor model for all neural network strategies.

The results of Tables 3 and 4 are illustrated in Figure 3. It plots the cumulative log returns of the value-weighted long and short sides for the three neural network strategies in the out-of-sample period. The long side buys the stocks in the highest decile portfolio and the short side sells the lowest decile portfolio. Therefore, returns on the short portfolio are the relative returns of the lowest decile stocks multiplied by -1. The cumulative performance is cut off in December 2008 and restarted in January 2009 to present differences in cumulative returns between the two halves of the out-of-sample period.

The global model consistently dominates the sector models over time. However, its outperformance is mainly in the first half of the out-of-sample period and tapers off thereafter. The cumulative returns of all three strategies follow similar patterns in the second half of the out-of-sample period. The long-short spreads are still larger for the global neural network after 2008, but the magnitude of the returns relative to the sector models is smaller. Sector-neutral portfolio sorting prevents the global model from outperforming the sector neural networks in the second half of the out-of-sample period. The performance of long-short portfolios for the global model is not predominantly based on the short side, which would raise questions about practical implementation due to shorting frictions.

The return series of the global model's long portfolio is strong initially and starts to shift after 2000. It still delivers positive results, but with higher volatility, the overall magnitude of relative returns is lower. The stocks in the top decile of the global neural network's forecasts cumulate double the returns in the eight years from 1994 to 2001 than in the following seven years. The second half of the out-of-sample period from 2009 to 2023 accumulates roughly the same returns as the first eight years. Apart from the shift in the return series during the dot-com bubble crash in 2001, global shocks such as the financial crisis of 2008 and 2009 or the COVID-19 pandemic in early 2020 did not cause significant portfolio downturns.

The short side of the global model generates positive returns but underperforms the long portfolio in both subsamples. Its positive performance is mainly due to the dot-com bubble crash from 2000 to 2002. Apart from this period, cumulative returns increase only slightly or move sideways over extended periods in the plot.

The long and short sides of the portfolios for the global model with sector-neutral portfolios generally follow the same pattern. The magnitude of returns and the long-short spread are smaller. The shift in the return series after 2000 is more pronounced, and the dot-com bubble crash causes a portfolio downturn.

The top and bottom decile portfolios based on the sorted predictions of the sector neural networks do not perform well in the first half of the out-of-sample period. The long portfolio generates no significant returns after 1999 and wipes out all accumulated returns in the dot-com bubble crash. The short portfolio benefits from this crisis but otherwise does not generate any substantial returns. It is the only portfolio to accumulate negative returns in parts of the out-of-sample period. After 2008, the cumulative returns of the long and short portfolios recover but never reach the magnitude of the global model.

Appendix A2 plots the cumulative log returns of top and bottom decile portfolios sorted on the out-of-sample return forecasts of individual sector-specific neural network models.

5.4. Benchmarking with OLS sector models

Sector-specific neural networks underperform a benchmark in the form of sector-specific ordinary least squares (OLS) models in the full sample. A long-short portfolio based on the sorted predictions of OLS sector models generates a higher value-weighted relative return than the sector neural networks. However, the OLS outperformance comes only from the first half of the out-of-sample period. The trade-off between stable model estimation and capturing sector-specific complex interactions is particularly relevant for small sector samples. Forecasts from the global neural network on the pooled data across sectors remain more profitable than the OLS models.

Sector models underperform the global model regarding stock-level forecasting and long-short portfolio performance, even when sector-specific breakpoints sort the global model predictions. As a first step towards understanding the reasons for this observation, I compare the sector-specific models with another machine learning technique. The sector models used in this thesis build on a neural network architecture with three hidden layers (NN3). I compare these with sector-specific models based on a simple linear predictive regression model estimated by ordinary least squares. The methodology remains the same; only the machine learning model for predicting returns from stock-specific signals changes.

Table 5 compares the out-of-sample performance of long-short portfolios based on the sorted predictions from the two

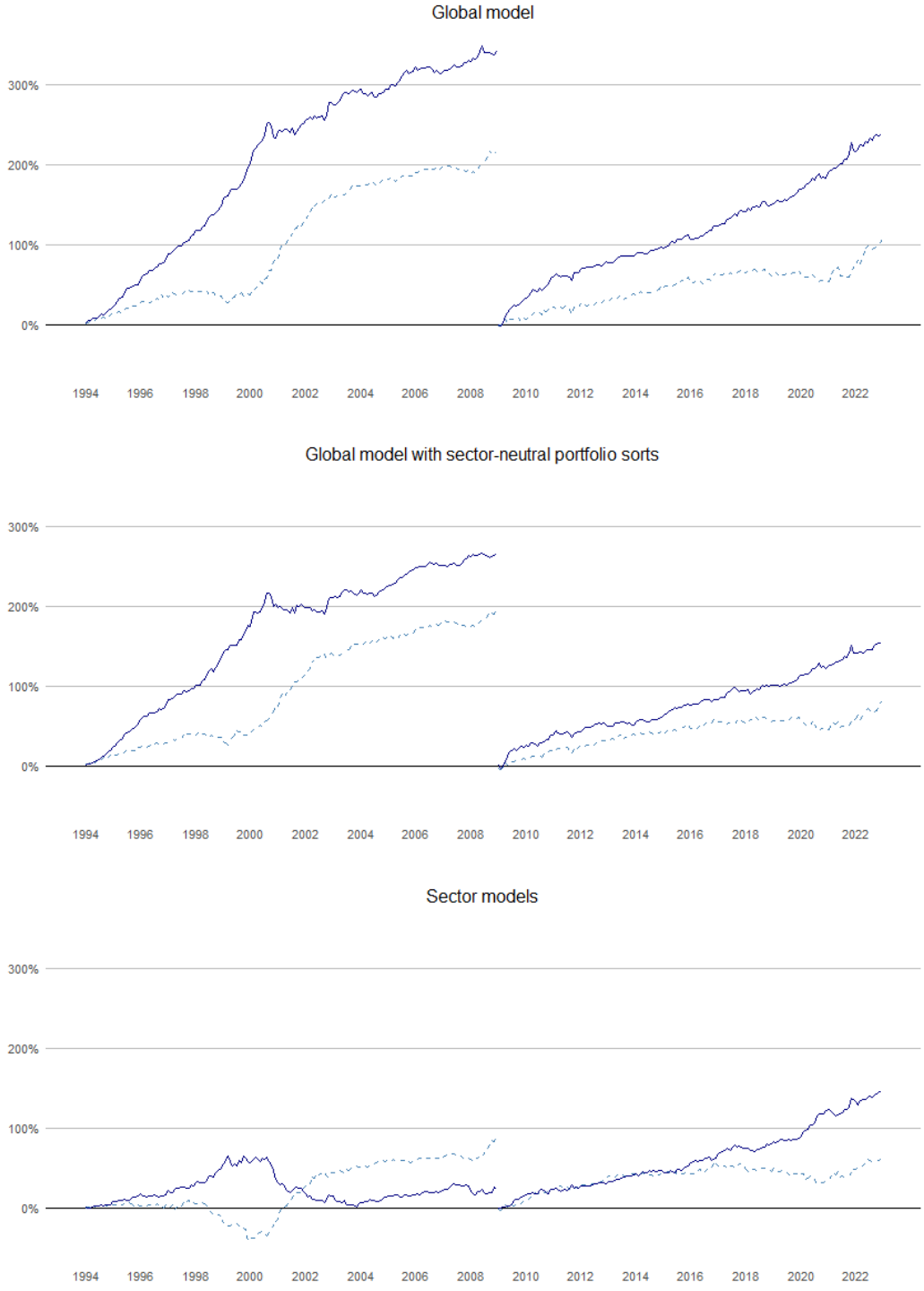


Figure 3: Cumulative performance of top and bottom decile portfolios

The figure plots the cumulative log returns of top and bottom decile portfolios sorted on the out-of-sample neural network return forecasts. Each month, stocks are sorted into value-weighted decile portfolios based on the predicted relative returns (with the market component removed) from the three neural network strategies. The solid and dash lines represent long (top decile) and short (bottom decile) positions, respectively. For the short position, the monthly relative returns of the bottom portfolio are multiplied by -1. The figure includes plots for the cumulative returns of the global model, the global model with sector-neutral portfolio sorts, and the sector models. The sample consists of US CRSP stocks, excluding microcap stocks with a market capitalization smaller than the 20th percentile of stocks listed on the NYSE. The sample runs from January 1994 to December 2022. The cumulative performance is cut off in December 2008 and restarted in January 2009 to present differences in cumulative returns between the two sub-samples.

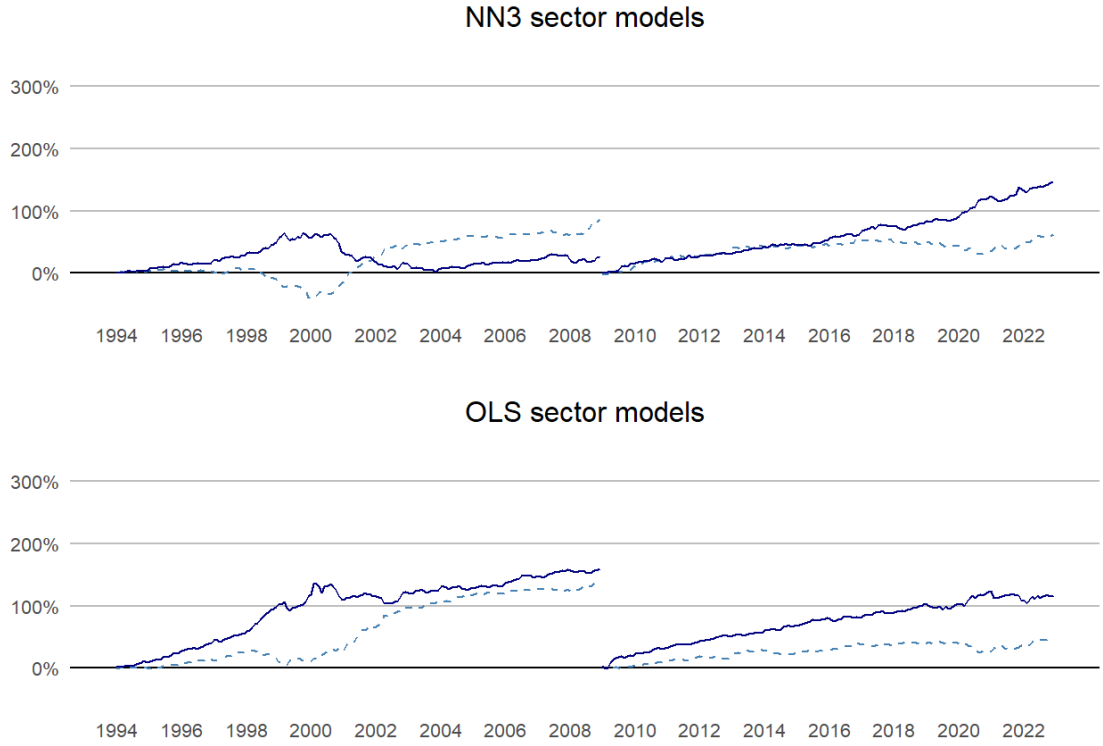
Table 5: Performance of NN3 sector models versus OLS sector models

This table compares the out-of-sample performance of the value-weighted long-short portfolios formed from different machine learning sector model return predictions. Results are reported for the NN3 sector models (same sector models as in Table 4 and Figure 3) and OLS sector models. All stocks are sorted into decile portfolios based on their predicted relative returns (with the market component removed) for the next month. A long-short portfolio buys the highest expected return stocks (decile 10) and sells the lowest (decile 1). Panel A compares the average value-weighted monthly full sample mean return and average monthly sub-sample mean returns with associated t-statistics (t-stat). Panel B visualizes the cumulative log returns of top and bottom decile portfolios sorted on the out-of-sample return forecasts of the two machine learning strategies. The solid and dash lines represent long (top decile) and short (bottom decile) positions, respectively. The sample consists of US CRSP stocks, excluding microcap stocks with a market capitalization smaller than the 20th percentile of stocks listed on the NYSE. The sample runs from January 1994 to December 2022.

Panel A: Percentage returns

Sector Models	Mean 1994-2022	*t-stat*	Mean 1994-2008	*t-stat*	Mean 2009-2022	*t-stat*
NN3	0.99	*6.14*	0.70	*3.06*	1.29	*5.80*
OLS	1.43	*7.80*	1.78	*6.41*	1.05	*4.53*

Panel B: Cumulative performance of top and bottom decile portfolios



different machine learning methods. Panel A reports average monthly value-weighted out-of-sample returns with associated t-statistics (*t-stat*). The sector neural networks (NN3) returns are the same as in Panel A of Table 4. Panel B visualizes the cumulative out-of-sample log returns of the value-weighted long and short sides for the two machine learning models.

The *OLS* sector models outperform the neural network sector models over the whole out-of-sample period from January 1994 to December 2022. They achieve average monthly long-short portfolio returns of 1.43% with an associated t-statistic of 7.80. However, the results differ for the two different subsamples. The outperformance of the OLS models is based solely on the first half of the out-of-sample period from 1994 to 2008. The OLS sector models generate significantly higher long-short returns than the NN3 sector models during this period. The monthly relative returns of 1.78% for the OLS models are more than double the 0.70% achieved by the sector models. The picture changes when looking at the later subsample from 2009 to 2022. In this period, OLS models underperform NN3 models, with average monthly returns of 1.05% and an associated t-statistic of 4.53. Panel B illustrates these results. The cumulative log returns for the OLS long and short portfolios remain positive throughout the out-of-sample period. They perform more reliably in the early years, with only a tiny long-short spread. As with all other machine

learning models, the gains for the short portfolio are mainly due to the dotcom bubble crash. The biggest difference between the OLS and NN3 sector models is the long portfolio in the first half of the out-of-sample period. The cumulative returns of the top decile stocks based on OLS forecasts are about an order of magnitude higher than their NN3 counterparts. In the second subsample from 2009 to 2022, the returns of the OLS long and short portfolios follow a similar pattern to the NN3 returns, with a slight underperformance.

These results demonstrate a trade-off between robust model estimation and complex model architecture to capture sector-specific nonlinearities and variable interactions. A simple linear regression model such as OLS requires only a single parameter for each of the 212 signals. The parameters for the optimization problem are derived from a closed-form solution. Therefore, OLS models benefit from a higher ratio of observations to parameters and more stable model parameters. Their obvious disadvantage is the inability to capture non-linearities and variable interactions. The number of estimated coefficients in each of my neural network models is 7,489. These numbers result in a low observations-to-parameter ratio on small data sets, such as the individual sector samples, especially at the beginning of my sampling period when the training samples are the smallest. In low signal-to-noise problems like stock return prediction, complex machine learning models such as neural networks tend to overfit noise rather than extract signals. The NN3 sector models cannot exploit their advantage of being able to capture complex sector-specific relationships between signals and future returns. This can lead to the poor predictive performance of neural networks for small sectors and low portfolio returns in the first half of the out-of-sample period. The pooling of data across sectors for the global model improves the ratio of observations to parameters. This explains why it does not suffer from the same problems as the sectoral models. The global neural network is still able to capture nonlinearities and variable interactions across sectors and thus significantly outperforms OLS models in the full sample, in line with previous research (e.g., Azevedo and Hoegner (2023), Blitz et al. (2023), and Gu et al. (2020)).

Table A3 in the Appendix summarizes the out-of-sample statistics of the value-weighted long-short portfolios formed from different sector-specific OLS model return predictions.

5.5. Sector allocation as a return driver

The global neural network demonstrates some out-of-sample sector allocation power. In the cross-section of sectors, it correctly predicts higher relative returns for the most profitable sectors and lower returns for the least profitable sectors. This allows the global model to generate higher returns in the long (top decile) portfolio when portfolio sorting is not sector-neutral.

As seen in Table 4, sector-neutral portfolio sorts worsen the profitability of long-short portfolios sorted on the out-of-sample return forecasts of the global neural network. To better understand this difference in performance, I briefly evaluate the global neural network's potential out-of-sample sector allocation power. Table 6 summarizes my results. For each of the ten GICS sectors, Panel A reports the average value-weighted realized monthly relative return compared to the average predicted return from the global model. In addition, I report the average monthly allocation to the respective sector in the top decile portfolio sorted based on forecasts from the global model. I focus on the long portfolio, the stronger driver of the global model's profitability than the short portfolio. Panel B plots the cumulative value-weighted relative log returns per sector over the out-of-sample period from January 1994 to December 2022.

Information Technology (*tech*) is the best-performing sector over the entire out-of-sample period, with a monthly relative return of 0.23%. Health Care (*health*) follows in second place with a return of 0.09%, and Energy (*energy*) is in third place with a return of 0.01%, thanks to solid gains after 2020. All other sectors generate negative value-weighted relative returns out-of-sample.[5] The worst-performing sectors are Utilities (*utilities*), with an average monthly return of -0.24%, and Communication Services (*comm*), with a return of -0.35%. Apart from Information Technology, the global model's return predictions are not very close to the realized returns, and they overstate expected returns. Still, it performs relatively well in cross-sectionally classifying the sectors into the correct extremes. The global model correctly identifies the two best-performing sectors and predicts Utilities and Communication Services to be among the three worst-performing sectors.

As a result, the global model shows some (limited) out-of-sample sector allocation power. The long portfolio based on global neural network forecasts has high average allocations to Information Technology (22.93%) and Health Care (15.09%). It has an average allocation of less than 5% to each of the three worst-performing sectors: Materials, Utilities, and Communication Services. The global model generates higher returns than sector-neutral portfolios by overweighting more profitable sectors and underweighting less profitable sectors in the top decile portfolio.

6. Conclusion

I examine the difference in predictive power for the cross-section of US stock returns between a global machine learning model and sector-specific models. Based on their strong performance in previous research, I use neural networks as the machine learning models. The global neural network is trained on the full sample of stocks, while the sector neural networks are trained on ten different GICS sectors. The global model consistently outperforms the sector models out-of-sample in terms of predictive accuracy and profitability. It

[5] Equally-weighted results (not reported) show that almost all sectors have positive relative returns. On average, stocks with smaller market capitalization generate higher relative returns. This further demonstrates the strong performance of the Size factor in the sample and justifies the high variable importance of Size in the global neural network.

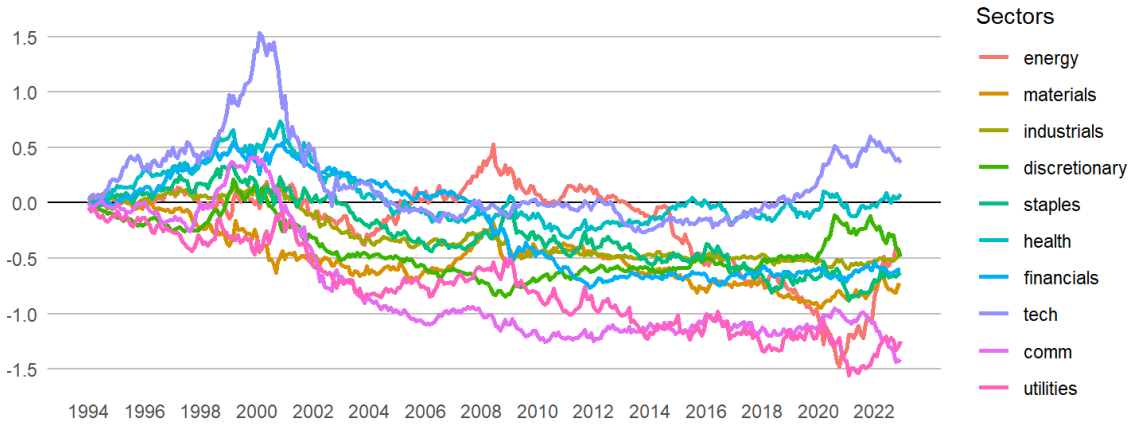
Table 6: Sector allocation power of global neural network model

This table summarizes parts of the out-of-sample sector allocation power of the global neural network model. Panel A compares the average monthly realized relative returns (with the market component removed) per sector to the average monthly predicted relative returns from the global model over the out-of-sample period. The sectors in Panel A are ranked in descending order of realized returns. Panel A additionally reports the average monthly allocation to each sector of the long (top decile) portfolio sorted based on the global model's return predictions for the next month. Panel B plots the cumulative relative log returns per sector over the out-of-sample period. All returns are value-weighted. The out-of-sample period runs from January 1994 to December 2022. The sample consists of US CRSP stocks, excluding microcap stocks with a market capitalization smaller than the 20th percentile of stocks listed on the NYSE.

Panel A: Percentage returns and long portfolio allocation

Sector	Realized returns	Predicted returns	Long portfolio allocation
Information Technology	0.23	0.26	22.93
Health Care	0.09	0.17	15.09
Energy	0.01	0.06	7.21
Consumer Discretionary	-0.10	0.14	14.52
Consumer Staples	-0.11	-0.03	3.74
Industrials	-0.11	0.12	11.54
Financials	-0.13	0.14	12.96
Materials	-0.16	0.11	4.50
Utilities	-0.24	-0.02	3.96
Communication Services	-0.35	0.05	4.36

Panel B: Cumulative relative log returns per sector



derives most of its predictive power from Size as an input signal. A long-short portfolio based on the sorted predictions of the global model generates significant returns and Sharpe ratios over the entire out-of-sample period. The sector models generate negative out-of-sample R^2_{OOS} and their long-short portfolio returns are lower, especially in the early out-of-sample period. Complex models such as non-linear neural networks struggle to exploit their advantages on small sector-specific samples and underperform simple OLS models. The results for the global model support the recent literature on the strong predictive power of neural networks for the cross-section of stock returns.

References

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets*, *5*(1), 31–56.

Avramov, D., Cheng, S., & Metzker, L. (2023). Machine Learning vs. Economic Restrictions: Evidence from Stock Return Predictability. *Management Science*, *69*(5), 2587–2619.

Azevedo, V., & Hoegner, C. (2023). Enhancing stock market anomalies with machine learning. *Review of Quantitative Finance and Accounting*, *60*(1), 195–230.

Azevedo, V., Hoegner, C., & Velikov, M. (2024). The Expected Returns on Machine-Learning Strategies. *SSRN Electronic Journal*.

Azevedo, V., Kaiser, G. S., & Mueller, S. (2023). Stock market anomalies and machine learning across the globe. *Journal of Asset Management*, *24*(5), 419–441.

Blitz, D., Hanauer, M. X., Hoogteijling, T., & Howard, C. (2023). The Term Structure of Machine Learning Alpha. *SSRN Electronic Journal*, 1–40.

Cakici, N., Fieberg, C., Metko, D., & Zaremba, A. (2023). Machine learning goes global: Cross-sectional return predictability in international stock markets. *Journal of Economic Dynamics and Control*, *155*, 1–32.

Cavaglia, S. M. F. G., Sefton, J., Scowcroft, A., & Smith, B. (2006). Global Style Investing. *The Journal of Portfolio Management*, *32*(4), 10–22.

Chen, A. Y., & Zimmermann, T. (2022). Open Source Cross-Sectional Asset Pricing. *Critical Finance Review*, *27*(2), 207–264.

Chen, L., Pelger, M., & Zhu, J. (2024). Deep Learning in Asset Pricing. *Management Science*, *70*(2), 714–750.

Ciner, C. (2019). Do industry returns predict the stock market? A reprise using the random forest. *Quarterly Review of Economics and Finance*, *72*, 152–158.

Drobetz, W., & Otto, T. (2021). Empirical asset pricing via machine learning: evidence from the European stock market. *Journal of Asset Management*, *22*(7), 507–538.

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

French, K. R. (2024). Current Research Returns. Retrieved November 1, 2024, from http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Freyberger, J., Neuhierl, A., & Weber, M. (2020). Dissecting Characteristics Nonparametrically. *The Review of Financial Studies*, *33*(5), 2326–2377.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Hanauer, M. X., & Kalsbach, T. (2023). Machine learning and the cross-section of emerging market stock returns. *Emerging Markets Review*, *55*, 101022.

Hou, K., Xue, C., & Zhang, L. (2020). Replicating Anomalies. *The Review of Financial Studies*, *33*(5), 2019–2133.

Kim, K. J., Lee, C., & Tiras, S. L. (2013). The effects of adjusting the residual income model for industry and firm-specific factors when predicting future abnormal returns. *Asia-Pacific Journal of Financial Studies*, *42*(3), 373–402.

Leippold, M., Wang, Q., & Zhou, W. (2022). Machine learning in the Chinese stock market. *Journal of Financial Economics*, *145*(2), 64–82.

Lewellen, J. (2015). The Cross-section of Expected Stock Returns. *Critical Finance Review*, *4*(1), 1–44.

Liu, X., Pong, E. S. Y., Shackleton, M. B., & Zhang, Y. (2014). Option-Implied Volatilities and Stock Returns: Evidence from Industry-Neutral Portfolios. *The Journal of Portfolio Management*, *41*(1), 65–77.

Moritz, B., & Zimmermann, T. (2016). Tree-Based Conditional Portfolio Sorts: The Relation between Past and Future Stock Returns. *SSRN Electronic Journal*, 1–81.

MSCI Inc. (2024). The Global Industry Classification Standard (GICS®). Retrieved November 1, 2024, from https://www.msci.com/our-solutions/indexes/gics

Nagel, S. (2005). Short sales, institutional investors and the cross-section of stock returns. *Journal of Financial Economics*, *78*(2), 277–309.

Newey, W. K., & West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *The Econometric Society*, *55*(3), 703–708.

Rapach, D. E., Strauss, J. K., Tu, J., & Zhou, G. (2019). Industry Return Predictability: A Machine Learning Approach. *The Journal of Financial Data Science*, *1*(3), 9–28.

Rasekhschaffe, K. C., & Jones, R. C. (2019). Machine Learning for Stock Selection. *Financial Analysts Journal*, *75*(3), 70–88.

Tobek, O., & Hronec, M. (2021). Does it pay to follow anomalies research? Machine learning approach with international evidence. *Journal of Financial Markets*, *56*.

U.S. Department of Labor. (2025). Standard Industrial Classification (SIC) Manual. Retrieved November 1, 2024, from https://www.osha.gov/data/sic-manual