Empirical Asset Pricing via Machine Learning

– Evidence from the Chinese stock market

Stockholm School of Economics

Bao Liu

Chuyue Huan

May 2023

Abstract

This thesis builds upon existing research on the application of machine learning in asset pricing in the US and European stock markets, by incorporating unique predictive indicators specific to the Chinese stock market, to explore whether machine learning can also be successfully applied in the Chinese stock market. Empirical results show that machine learning models outperform OLS significantly in predicting A-share returns, and this conclusion also applies to different portfolios we have constructed. In the analysis of feature importance, we found that the retail investors' dominating presence in the Chinese stock market makes macroeconomic variables and variables containing direct trading information, such as technical indicators, trading volume, and turnover, more influential. This is in contrast to the US market and reflects the characteristics of the Chinese stock market.

Keywords: Machine learning, Asset pricing model, Chinese stock market.

Acknowledgments: We would like to thank Dr.Tobias Rydén for his insightful guidance and support throughout our thesis writing process.

Tutor: Tobias Rydén, Senior Research Partner at Lynx Asset Management

Contents

Chapter 1

Introduction

1.1 Background

Return prediction has been a key area in the field of asset pricing. Whether it is a comparison of cross-sectional individual stock returns or a time-series forecast of index returns, the crucial issue is to determine the appropriate asset pricing model. Cross-sectional predictions usually study stock-level characteristics, such as size, value, and momentum (Fama French 1993; Lewellen 2014). While time-series predictions generally conduct regressions on a handful of macroeconomic indicators, namely interest rates and valuation ratios, and also on some technical indicators (Rapach and Zhou 2013). Most applications for cross-sectional and time-series predictions align with the established literature on asset pricing. They typically consider only linear relationships between indicators and expected returns. For example, the traditional capital asset pricing model (CAPM) based on Markowitz's mean-variance portfolio theory, assumes that the excess return of a stock or portfolio arises only from its sensitivity to systematic risk. However, the subsequent arbitrage pricing model (APT) and the classical Fama-French factor model point out that the factors affecting returns go beyond market factors and should include other factor variables in other areas such as corporate fundamentals and macro environment, i.e., multi-factor models. Afterward, many scholars start to explore nonlinear relationships in pricing factors. Bansal et al. (1993), Chapman (1997), and Asgharian and Karlsson (2008), to name a few, show that incorporating nonlinear relationships between indicators and expected stock returns add more explanatory power. Asset pricing is moving into the era of big data. Some issues, such as the importance ranking of predictive variables, cannot be explained by traditional econometric models, but need to be explored and studied with a high-dimensional perspective combined with machine learning.

In the face of the challenges posed by big data in finance, the improvement and application of machine learning algorithms have become a hot topic of academic research in recent years. Machine learning models can effectively handle high-dimensional data during the training process, avoiding the overfitting and underfitting problems of traditional regression models such as least squares regression (OLS) and other statistical tools.

Although machine learning has performed well in many fields, its application in financial markets cannot be accomplished by simple transplantation. The main reason for this is the unique characteristics of financial data. First, financial data are intrinsically noisy, making it difficult to accurately model financial phenomena. Also, models may overfit the noise, capturing the idiosyncrasies of the training data rather than generalizable patterns. In addition, the efficient market hypothesis suggests that a perfectly efficient market cannot take advantage of past information to obtain excess returns. Although such a perfect market does not exist in reality, the existence of arbitrage by market traders does absorb most of the efficient information, reducing the validity of historical data. The second characteristic is that financial data have a short sample time span. Modern stock markets are only 100 years old, and more new financial data can only be obtained over time. Most of the literature studying asset pricing use daily or even monthly data, making the data sample size limited and thus constraining the estimation and validation process of the models.

After solving the "transplantation" problem, machine learning can provide a rich set of algorithms to support asset pricing. Although the theoretical structure of some methods has been built in statistics for a long time, the generation of big data has made it possible to apply the models widely. By applying proven models to asset pricing, we can analyze the joint effect of a large number of indicators. A typical example is Gu et al. (2020), who use machine learning algorithms including LASSO, ridge regression, elastic net, PCA, PLS, random forest, and neural network models for asset return forecasting and find that non-linear models such as neural networks have significantly better cross-sectional return forecasting power than linear regression models, and overall, machine learning models are able to deliver a more significant economic return.

By far, the Chinese stock market has been particularly attractive for academic research, not only due to its increasing size but also because of its specificity. Considering the differences between the Chinese stock market and mature stock markets in various aspects such as economic environment, regulatory policies, and trading systems, some conclusions proposed before may not apply to it. For example, Liu et al. (2019) and Hu et al. (2019) similarly find that the profit-to-price factor is more applicable to asset pricing models in the Chinese stock market than the book-to-market ratio factor. The differences in pricing models highlight the significance of asset pricing studies for the Chinese stock market.

Our empirical study follows that of Gu et al. (2020) but with some key differences. First, we use Chinese stock market data and add some Chinese-specific indicators into models. Second, instead of building a comprehensive set of 94 characteristics, we use a parsimonious set of only 36 characteristics, dividing the stock-level characteristics into 6 categories according to Hou et al. (2015). To examine our model's performance for predicting aggregate portfolio returns, we build equal-weighted and value-weighted portfolios based on predicted individual stock returns and compare the Sharpe ratios of the portfolios. Finally, through comparative analysis, we find that nonlinear machine learning models outperform linear regression models in predicting the excess return of A-shares. Additionally, we are able to examine the indicators that have a greater impact on the predictions in the Chinese stock market.

The remainder of this paper is structured as follows. Chapter 2 surveys the existing literature on asset pricing theories, machine learning applications in finance, and an overview of the Chinese stock market. Chapter 3 describes the data and also covers the methodology of our study, including a brief introduction to the different models we used. Chapter 4 shows our empirical analysis results, including out-of-sample tests and feature importance demonstration. Chapter 5 conducts a study on the expected return-sorted portfolios. Finally, Chapter 6 contains a short conclusion based on the results.

1.2 Limitations

In this paper, there are several limitations that need to be acknowledged:

First, our methodology employs walk-forward validation, which has led us to select a stock pool comprising only those non-ST stocks that have been consistently trading since 2003, excluding stocks from the financial and real estate sectors. This selection criterion may limit the generalizability of our findings to other stocks and sectors.

Second, due to the absence of publicly available forums or websites in China discussing stocks, it is not feasible to mine and analyze retail investor sentiment for sentiment analysis. Consequently, our study does not incorporate any investor sentiment indicators in the selection of metrics, which may affect the comprehensiveness of our results.

Third, the Chinese stock market has a relatively short history of established rules and regulations, leading to a lack of data for many early years. Although we have attempted to fill in missing values to the best of our ability, it is possible that this data limitation may impact the accuracy of our predictions.

Last but not least, in this study, a limitation is manifested in the form of selection bias. To ensure the uniformity and coherence of our dataset, we incorporated stock data spanning from January 2003 to September 2022, deliberately excluding delisted stocks. Consequently, our data selection is inherently susceptible to forward-looking bias, given the inherent difficulty in accurately anticipating the delisting of stock prior to the event transpiring. Moreover, an array of factors may precipitate the delisting of stock, including but not limited to corporate bankruptcy, mergers, and buyouts. It is noteworthy that stocks delisted due to bankruptcy frequently demonstrate poor performance. By omitting delisted stocks from the dataset during the specified study period, an inadvertent overestimation of the overall performance of the dataset occurs when compared to real-world stock performance. This ultimately gives rise to the presence of forward-looking bias within our scholarly investigation.

Chapter 2

Literature Review

2.1 The Development of Assets Pricing Models

Introduced by William Sharpe (1964), John Lintner (1965), and Jan Mossin (1966), the Capital Asset Pricing Model formed the basis of Asset Pricing Theory. The CAPM is a valuable tool to estimate a company's cost of capital and measure the expected returns of an investor's portfolio. It effectively demonstrates the connection between risk and expected return, stating that an asset's expected rate of return is equal to the compensation it receives for its exposure to market risk.

At first, the studies were mainly based on individual stock returns to test the CAPM. However, the empirical results were not encouraging. Lintner (1969) and Douglas (1967) both discovered that the intercept in the CAPM had significantly higher values than the risk-free rate of return. Conversely, the beta coefficient had a comparatively lower value. Miller and Scholes (1972) also found the same problem when they were studying individual stocks. After then, some scholars replaced individual stock returns with portfolio returns, managing to avoid the statistical problem mentioned above. Black et al. (1972) created portfolios that included all stocks listed on the New York Stock Exchange (NYSE) between 1931 and 1965, suggesting that the expected excess return on an asset was not always directly proportional to its beta, thereby presenting significant evidence that contradicted the CAPM. Fama and MacBeth (1973) also formed 20 portfolios of all the stocks listed on the NYSE between 1935 and 1968, but the difference is that they employed monthly data for a time series regression model to estimate the beta. Their research pointed out that the coefficient of beta was significant and the value remained small over several sub-periods.

However, many scholars still held doubts about the validity of the CAPM. For example, an insignificant relationship between beta and returns was discovered by Lakonishok and Shapiro (1986). And meanwhile, they detected a significant relationship between market capitalization and returns. Tinic and West (1984) followed Fama and MacBeth (1973), applying the same NYSE data from 1935 to 1982, yet their findings were contrary. In the early 1980s, several studies identified that the single-factor CAPM was not adequate and the beta was not sufficient in clarifying the relationship between excess risk and return. Therefore, many studies turned to other potential factors that could affect the assets' risk-return relationship.

Basu (1977) proposed a hypothesis that in an efficient capital market, stock prices would immediately and impartially reflect all available information, thus providing unbiased estimates of fundamental values. He also evaluated the performance of common stocks to investigate whether their performance was linked to their P/E (price-to-earnings) ratios. The evidence demonstrated that the stocks with lower P/E ratios had higher expected returns than predicted by the CAPM. Then, Banz (1981) obtained a similar conclusion for small stocks which had lower market capitalization. Meanwhile, Stattman (1980) and Rosenberg et al. (1985) also reached a conclusion that stock returns were positively related to the B/M (book-to-market) ratio by conducting the study on U.S. stocks. Bhandari (1988) found that expected stock returns had a positive correlation with the debt-to-equity ratios when beta and firm size were under control. These studies above suggested that a single-factor CAPM did not hold and that other factors could also contribute to asset returns. Hereby, the multifactor CAPM appeared.

First, Chan et al. (1991) studied the Japanese stock market, attempting to relate cross-sectional differences in stock returns to the four variables: yield, size, cash flow returns, and B/M ratio. Their results proved these variables were significantly related to expected returns in the Japanese market. Among them, cash flow returns and the B/M ratio had the most significant positive effect. Then, Fama and French (1992) used a more indirect approach. Their arguments suggested that the higher average returns of stocks with small sizes but high B/M ratios could reflect unidentified state variables that generate undiversifiable risk, which was also called covariance, in returns that were not able to be captured by market returns and priced separately from market betas, though size and B/M ratio are not state variables. Griffin (2002) found the three-factor model more useful, especially in explaining stock returns when applied to specific countries. In order to estimate time-varying betas, Koutmos and Knif (2002) introduced a dynamic vector GARCH model, finding that there was a probability of 50% that betas are higher when the market falls. Their opinion about this result was that the market model which was static could exaggerate the unsystematic risk and that betas which were dynamic followed a stationary, mean-reverting process. Then, Fama and French (2004) further confirmed the failure of the traditional CAPM in empirical tests. Thompson et al. (2006) showed some crucial evidence against the CAPM in their study as well.

In all, the key assumptions of CAPM have been severely criticized, including the assumption of a linear relationship between risk and return, and the assumption of a unique risk factor. Therefore, many researchers have proposed new models to overcome the shortcomings of the CAPM, especially machine learning methods.

2.2 The Application of Machine Learning in Asset Pricing

Utilizing machine learning methodologies enables us to overcome many constraints inherent in conventional asset pricing approaches, specifically addressing the challenges of predictive accuracy, feature selection, and functional form determination. Firstly, most machine learning tools are designed for prediction. Machine learning algorithms can automatically and efficiently learn patterns and relationships from existing datasets and adapt the algorithm to new data, improving their predictive accuracy over time. Secondly, technical, financial, and macroeconomic variables have customarily been considered the most influential indicators affecting stock prices. As the advancement of artificial intelligence persists, and the accumulation of signals expands over multiple decades, researchers can leverage increasing features in machine learning models to enhance asset pricing. Nevertheless, these predictors are usually highly correlated. Machine learning tools can provide dimension reduction and variable selection or shrinkage tools to optimize degrees of freedom. For example, Freyberger et al. (2020) employ the nonparametric adaptive group LASSO technique to discern vital independent variables that furnish incremental insights regarding expected returns. Kozak et al. (2017) claim that the task of estimating cross-sectional stock returns with a limited number of predictors is considered unfeasible. To address this issue, they use the joint explanatory power of the high dimensional set of predictors to construct a robust stochastic discount factor, contrasting L1-penalty (lasso) and L2-penalty (ridge). They also present economic rationales for the superior empirical performance of L2-penalty. Harvey and Liu (2016) identified that the market factor is essential in explaining the cross-section of expected returns; the study challenged the common notion that individual stocks are unsuitable as test sets due to excessive noise by operating on stock-level data. Additionally, they also offered a novel bootstrap implementation approach to test multiplicity. Giglio and Xiu (2019) employ a principal component analysis (PCA) method and find that market frictions (e.g., liquidity) have a robust and significant influence on risk premia in the application. Kelly et al. (2019) propose an Instrumented Principal Component Analysis (IPCA) method to re-estimate common predictors and identify IPCA factors that are statistically significant.

Thirdly, the association between the risk premium and the predictors may be characterized as either linear or non-linear. Through the application of effective machine learning algorithms, the optimal functional form can be ascertained with minimal computational expense. Moritz and Zimmermann (2016) conducted a comparison between excess returns derived from the Fama-MacBeth framework and tree-based models. Their research focused on relating past returns to future returns. The study revealed that tree-based models could generate more stable excess returns. This finding indicates that the current linear framework may only partially capture some relevant information in the dataset. Consequently, tree-based machine learning models demonstrate more outstanding prowess in forecasting stock returns. Messmer (2017) leverages a deep feedforward neural network (DFN) predicated on an extensive array of firm attributes to forecast the United States' cross-sectional stock return, thereby illuminating the inherent non-linear relationship between returns and firm-specific characteristics. Gu et al. (2019) conducted a comparative evaluation of eleven distinct machine learning techniques in the realm of asset pricing, utilizing a comprehensive dataset of US stocks. Their results highlight the capacity of machine learning approaches to enhance comprehension in the field of asset pricing. They advocate that neural networks and regression trees exhibit the highest performance, while all methods converge on a relatively compact array of prevailing predictive indicators. The most robust forecasting predictors identified are return reversal and momentum. Although machine learning substantially improves asset pricing, the author emphasizes that unsupervised learning cannot identify deep fundamental economic mechanisms. Some studies applied classification-based machine learning methods. Leung et al. (2000) emphasized predicting the direction of stock movements using classification-based machine learning models, as opposed to estimating the precise rate of return for individual stocks.

2.3 Overview of the Chinese Stock Market

The Chinese stock market was born in 1990 with the establishment of the Shanghai Stock Exchange. Over the past 20 years, it has experienced multiple stages of initial exploration, policy regulation, and rapid development. With the launch of the Sci-Tech Innovation Board and the implementation of the registration system in 2019, more IPOs of high-quality enterprises will further promote the prosperous development of the Chinese stock market. By the end of 2022, the total market capitalization of the Shanghai and Shenzhen stock exchanges reached 78.8 trillion yuan, accounting for more than 60% of China's total GDP and ranking second in the world.

The Chinese stock market possesses certain unique characteristics. First, due to Chinese government regulations, most foreign investors are not allowed to acquire shares in Chinese companies, and Chinese investors are restricted from participating in foreign markets. Second, Chinese stock markets are dominated by retail investors, in contrast to developed markets that are dominated by institutional investors. For example, individual retail investor transactions accounted for about 60% of the total A-shares turnover in late 2022, with the figure being less than 25% in the US. Moreover, some studies show that there is a weak connection between stock returns and macro factors. To be specific, the Chinese stock market is always with a long downturn after a spike, which is hard to explain. These distinctive market characteristics trigger many scholars' interest in predicting stock returns.

Jordan, Vivian, and Wohar (2014) presented an interesting finding that the returns of China's 15 largest trading partners could predict China's A-share index returns. Cakici, Chan, and Topyan (2017) found that predictions of stock returns in the Chinese stock market were usually out of expectations, making the predictability a little bit weak. Nevertheless, they insisted that stock returns in the Chinese stock market were still predictable if using some specific factors, such as the B/M ratio. Chen, Jiang, Liu, and Tu (2017) also believed that stock returns in China were predictable and found a useful variable, international volatility, which could predict returns in subsequent days. Lin et al. (2017) examined the empirical application of the Fama-French five-factor model and the momentum effect in the Chinese stock market. They used the data of listed companies in the Chinese stock market between July 1994 and August 2015 to test whether the five-factor model could be effectively applied in different periods and found that the size and B/M ratio are the most significant factors under the full sample. And then they divide the full sample into a two-stage subsample test according to the reform of ownership structures in China's listed companies and find that the market portfolio risk dominates before the reform, and the risk premium of three factors, profitability, investment style, and momentum factor, becomes significant after the reform. Hu et al. (2019) found that the traditional B/M ratio in the Chinese stock market was not suitable as a value factor due to the structural changes in the Chinese market during its development. Liu, Stambaugh, and Yuan (2019) confirmed that size and value factors could help explain most of the anomalies that appeared in China. However, they still doubted whether the U.S. model can be replicated in the Chinese market, considering that the two countries have very different economic and financial systems.

Looking at the contrasting results found by the researchers when trying to predict Chinese stock returns, further analysis is needed in order to assess whether returns are predictable in China and which factors should be chosen to better match the distinctive characteristics of the Chinese market.

Chapter 3

Data and Methodology

3.1 Data

The market and fundamental data of Chinese firms are collected from Wind Financial Terminal and Wind Economic Database. Our sample consists of firms that were publicly listed in the Chinese stock market before January 2003 and are still publicly listed as of September 2022, totaling 237 months. For our empirical analysis, we have exclusively incorporated firms that provide comprehensive data on monthly returns and all thirty specific characteristics. Furthermore, we have excluded stocks from the real estate and finance industries in our analysis due to their significantly higher leverage ratios compared to other industries. Finally, our sample comprises a total of 1010 stocks. To calculate excess returns, we use the Chinese 10-year treasury bond yield as the risk-free rate.

The thirty firm-level indicators we choose are based on some asset pricing literature written by Gu et al. (2020) and Hou et al. (2015). We have categorized the indicators into six groups: Basic Info, Profitability, Liquidity, Valuation, Investment, and Trading Info. Within these, four are refreshed on an annual basis, fifteen are updated every quarter, and seventeen undergo monthly updates. The detailed description of each predictor is elaborated in Table 3.1 below. In addition, we include six Chinese-specific macroeconomic characteristics following Hou et al. (2015), which are also shown in Table 3.1. Considering dummy variables used to encode different values of categorial variables, there are 56 predictors in total.

Table 3.1: Descriptive Statistics

#	Indicators	Avg	Std	Definition
target	excess_return[1]	-2.0	16.22	the monthly return of an individual stock over Chinese 10-year treasury bond yield.
1	firm_size			Categorical variable. The size of a firm, is measured by its market value. It is divided into four dummies: firm_size_large, firm_size_medium, firm_size_small, firm_size_micro.
2	firm_age	17.80	7.24	The number of years a firm has been founded
3	employ_no	5836.78	16672.51	The number of employees in a firm
4	ind_type			Categorical variable. The industry a firm belongs to, including 18 industries. It is also divided into 18 dummy variables.
5	roe	3.33	58.11	Return on Equity: net income divided by average shareholders' equity in the prior quarter
6	roa	3.08	27.04	Return on Assets: net income divided by average total assets in the prior quarter
7	roic	3.12	90.57	Return on Invested Capital: net income minus dividends divided by average shareholders' equity and debt in the prior quartor
8	gp_margin	-79.55	69933.04	Gross Profit Margin
9	op_margin	31.26	62311.14	Operating Profit Margin
10	ocf_to_a	1.58	20.08	Operating net cash flow divided by average total assets in the prior quarter. It is designed to evaluate the ability of a firm to generate cash from all of its assets.
11	z_score[2]	13.33	3218.66	Z=1.2X_1+1.4X_2+3.3X_3+0.6X_4+0.999X_5, where X_1=working capital/total assets, X_2=retained earnings/total assets, X_3=EBT/total assets, X_4=market value/total liabilities, and X_5=operating income/total assets. The score is used to analyze and predict the likelihood of financial failure or bankruptcy of a firm. The lower the Z-score, the more likely the firm is to experience bankruptcy.
12	current_ratio	1.73	9.00	Current assets divided by current liabilities at the end of the prior quarter
13	quick_ratio	1.27	8.29	Current assets minus inventories at the end of the prior quarter divided by current liabilities at the end of the prior quarter
14	cf_to_debt	0.72	7.86	Cash plus cash equivalents at the end of the prior quarter divided by current liabilities at the end of the prior quarter
15	bm	0.49	0.44	Book value of the equity divided by market value of the equity at the end of the prior month
16	pe	99.07	12688.68	Share price at the end of the prior month divided by the last four quarterly EPS
17	pb	5.00	415.14	Share price at the end of the prior month divided by the book value per share at the end of the most recent interim period
18	pcf	-36.51	18294.77	Share price divided by cash flow per share at the end of the prior month

Continued on next page

[1] Equal-weighted excess return

[2] This factor is Altman Z-Score proposed by Edward Altman to determine whether a company is headed for bankruptcy. See Altman(1968).

Table 3.1 – *Continued from previous page*

#	Indicators	Avg	Std	Definition
19	Itdebt_capital	18.75	274.48	Average long-term debt divided by average total capital in the prior quarter
20	a_to_e	3.39	31.35	Average total assets divided by average total equity in the prior quarter
21	ocf_to_invest	0.03	93.98	Cash Flow Adequacy Ratio: cash flow from operations divided by capital expenditure plus mandatory debt repayment plus dividends in the prior quarter
22	capex_to_da	1.99	15.58	Capital expenditure divided by depreciation and amortization in the prior quarter
23	fixed_growth	447.74	55611.43	The growth rate of fixed assets (quarterly)
24	market_beta	1.04	0.35	Market beta, estimated from weekly returns from month-25 to month-1
25	volatility	45.66	20.84	Standard deviation, estimated from weekly returns from month-25 to month-1
26	turnover_monthly	40.72	42.39	Average monthly turnover volume from month-25 to month-1
27	volume_month	25277.84	62755.98	Total trading volume in million shares
28	macd	0.27	3.33	Technical indicator. Moving Average Convergence/Divergence: subtracting the 26-period EMA from the 12-period EMA, where EMA is exponential moving averages.
29	kdj	40.11	22.65	Technical indicator. It is also known as the random index. It is used to analyze and predict changes in stock trends and price patterns in a traded asset.
30	momentum_index	0.15	11.03	Technical indicator. Moving average calculated using the closing price. It is used to determine if an asset is overbought or oversold.
31	inf_rate	2.50	1.84	We use the Consumer Price Index as an indicator of the inflation rate
32	money_multip	5.04	1.12	Money Multiplier measures the amount of new money that is created in the economy for every unit of currency held in reserve by the central bank. It can be used as a tool to predict inflation or deflation.
33	pi	99.94	11.55	The Macroeconomic Business Cycle Indicator (MEBCI) in China is a composite index that measures the overall health and growth of the country's economy. It is the Chinese prosperity index.
34	pmi	46.50	15.40	Purchasing Managers' Index is a widely used economic indicator that measures the health of a country's manufacturing sector.
35	cei	111.12	9.32	Consumer Expectations Index is an economic indicator that measures the confidence of consumers in the future performance of the economy.
36	vix	19.36	8.39	VIX is a measure of the market's expectation of volatility in the near term, based on the price of options on the S&P 500 index. We use it to measure the impact of the global markets.
	Sample size	1010		Number of firms with non-missing observations for the excess return and each of the thirty-six characteristics

Continued on next page

Table 3.1 – *Continued from previous page*

#	Indicators	Avg	Std	Definition

Table 3.1 presents the description of each indicator. #1-#4 belong to the group **Basic Info**, #5-#11 belong to the group **Profitability**, #12-#14 are in the group **Liquidity**, #15-#20 are included in the group **Valuation**, #21-#23 are in the group **Investment**, #24-#30 belong to the group **Trading Info**, and the rest are macroeconomic variables.

3.2 Sample Splitting

The dataset utilized in this paper comprises 56 independent variables (including dummy variables) and one dependent variable. Independent variables in our paper are primarily divided into stock-level data, which are mainly derived from company financial statements, and macroeconomic indicators. In real life, these data are interrelated and highly correlated, such as the fact that a company's net income is the numerator of ROE, ROA, and ROIC. When independent variables exhibit a high degree of correlation, multicollinearity issues may arise, consequently diminishing the accuracy of the model. Furthermore, forecasting models for stock data are highly susceptible to underfitting problems, as the stock market is characterized by complexity, nonlinearity, and frequently changing dynamics. The presence of a significant amount of noisy data makes it challenging to capture the underlying patterns within the model effectively.

To overcome such issues, we applied machine learning algorithms and optimized hyperparameters to control the complexity of the model to yield dependable out-of-sample predictive performance. For example, the random forest model has multiple parameters, such as n_estimators, the maximum number of features, and the maximum depth, and different parameter values can affect the model's performance. We improve the model's predictive accuracy by finding the optimal combination of parameters using the grid search method. Although this automated approach consumes a large amount of computing resources, it can avoid human bias and errors, ensuring the robustness and reliability of the model.

In the context of finance, it is crucial to retrain our models as new data becomes available over time. In our study, we employed the walk-forward validation approach, which involves tuning parameters and training models on a specific period (referred to as the training data) and evaluating their predictive power in the subsequent period (referred to as the testing data), and then rolling the fixed window forward until the last available data point is reached.

First, we divided the entire dataset into multiple segments from January 1, 2003, to September 31, 2022. Instead of gradually expanding the training window, we maintained a fixed window length and rolled it forward by one year. We set the fixed window to two years and the step length to one year.

Second, we separated the dataset within each window into two distinct subsamples: a training sample and a test sample. It is imperative to highlight that, as evidenced by Table 3.1, our independent variables' means and standard deviations display substantial disparities in their respective ranges. For example, the mean number of employees per company stands at 5,836.78, while the Cash Flow Adequacy Ratio (ocf_to_invest) exhibits a mean value of merely 0.03. Machine learning models demonstrate pronounced sensitivity to the scale of features, and such pronounced discrepancies in ranges may compromise model performance. As a result, we have undertaken the standardization of independent variables within each window to bolster the robustness and consistency of our model, thereby promoting more efficacious training. The training sample is used to fit models, tuning parameters, and evaluate the in-sample performance. The test sample is employed to evaluate the out-of-sample predictive power based on optimal parameters. We consistently allocated one and a half years for training and half a year for testing. Figure 3.1 illustrates the walk-forward validation method.

By conducting training across 19 separate windows, we generate 19 distinct models, each characterized by a unique set of parameters. The optimal model is identified according to its out-of-sample performance and subsequently employed within the entire dataset. To elaborate, we utilize the independent variables from the preceding month as input data, enabling the model to predict the excess stock returns for the ensuing month.

3.3 Machine Learning Methodology

This section covers the machine learning methods we employed in our empirical analysis. Each subsection introduces a new model and explains its fundamental

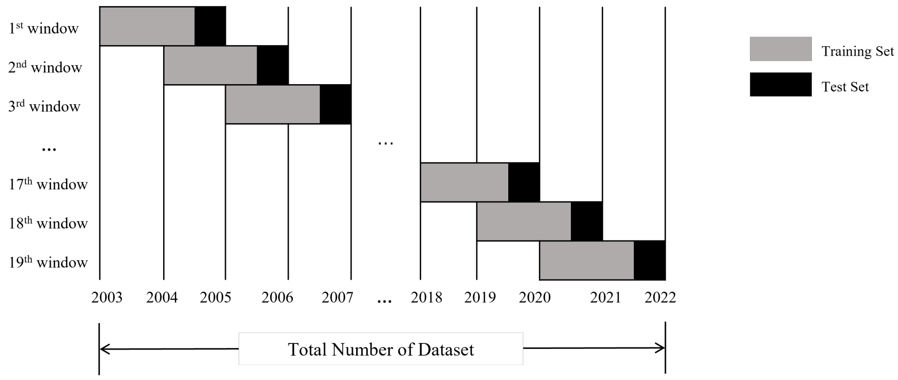


Figure 3.1: Walk-Forward validation

(a) In Figure 3.1, we demonstrate the process of walk-forward validation, a technique for validating time series data by progressively rolling the training and validation data windows.

elements. We explain the statistical model of each method, which facilitates understanding of the model structure for readers with limited background knowledge without the need for external references.

We undertake a comparative evaluation of diverse machine learning models to scrutinize their predictive performance. Following Gu et al. (2020), a stock's excess return could be described as:

$$r_{i,t+1} = E_t(r_{i,t+1}) + e_{i,t+1} \tag{3.1}$$

where $r_{i,t+1}$ is the excess return of stock $i = 1, 2, ..., N$ in month $t = 1, ..., T - 1$. The expected excess return $E_t(r_{i,t+1})$ is estimated by the function of a series of independent variables.

Our research objective is to compare the predictive accuracy of linear and nonlinear models, allowing nonlinear predictor interactions missed by traditional asset pricing models. In the context of the Chinese A-share market, we aim to identify the most effective machine learning model for predicting a stock's excess returns and to determine the dominant set of independent variables affecting the excess returns of Chinese stocks.

3.3.1 Linear Models

Ordinary Least Squares

Linear regression is a widespread statistical method commonly used in machine learning to model the relationship between a dependent variable (also known as the response or target variable) and one or more independent variables (also known as predictors or features). The central focus of linear regression is to establish the best-fit line or hyperplane that describes the underlying relationship between the variables in order to predict the value of the dependent variable of the independent variables.

In simple linear regression, we have one dependent variable Y and one independent variable X, and we assume that there is a linear relationship between the two variables. The mathematical model representing this relationship is:

$$Y_{i,t+1} = \beta_0 + \beta_1 X_{i,t} + \epsilon_{i,t+1} \tag{3.2}$$

where $Y_{i,t+1}$ is the dependent variable in month $t+1$, $X_{i,t}$ is the independent variable in month t, β_0 is the intercept, β_1 is the slope, and ϵ is the error term, accounting for the differences between the observed and predicted values of Y.

Multiple linear regression is an extension of simple linear regression. The equation for multiple linear regression can be written as follows:

$$Y_{i,t+1} = \beta_0 + \beta_1 X_{i,t}^1 + \beta_2 X_{i,t}^2 + ... + \beta_p X_{i,t}^p + \epsilon_{i,t+1}, \tag{3.3}$$

where $X_{i,t}^1$, $X_{i,t}^2$, ..., $X_{i,t}^p$ are the features and β_1, β_2, ..., β_p are the coefficients associated with each independent variable.

To maximize the out-of-sample explanatory power for realized return, OLS is designed to approximate the true forecast model by minimizing the out-of-sample mean squared error (MSE), also known as loss function:

$$MSE = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{i,t+1} - \hat{Y}_{i,t+1})^2 \tag{3.4}$$

Linear regression is used in many fields, including economics, finance, engineering, and biology, to model and predict a wide range of phenomena. However, it is important to consider the independence of the features when using OLS to estimate the coefficient values. When the features are highly correlated, multicollinearity can occur. In such cases, the design matrix becomes close to singular, and the OLS estimate becomes highly sensitive to random errors in the target variable, leading to a high variance in the results.

Penalized Regression

In order to address the potential overfitting issue of OLS, especially when dealing with high-dimensional datasets where the number of predictors is much larger than the number of observations, we employed regularized linear regression methods, including LASSO, ridge regression, and elastic net. Essentially, regularized linear regression modifies the cost function of OLS by introducing different regularization terms in the loss function, which can constrain the number and magnitude of parameters and prevent overfitting and enhance the robustness of the model, thereby increasing its out-of-sample explanatory power.

Among them, the model incorporating L1 regularization is called LASSO regression, and mathematically, its loss function is as follows:

$$J(\beta) = \frac{1}{2NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{i,t+1} - \beta_0 - \sum_{j=1}^{p} x_{ij,t}\beta_j)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \tag{3.5}$$

where NT represents the total number of samples, p represents the number of features, $Y_{i,t+1}$ represents the target variable value of the i-th sample in month $t+1$, $x_{ij,t}$ represents the value of the j-th feature of the i-th sample in month t, β_j represents the coefficient of the j-th feature, β_0 represents the intercept, and λ_1 is the L1 regularization parameter that controls the strength of the penalty term. The first term in the objective function represents the ordinary least squares (OLS) loss function, while the second term represents the L1 penalty term.

The L1 regularization can generate a sparse weight matrix, which means a matrix with many zero elements and only a few non-zero values. LASSO regression has several advantages over traditional linear regression. It can handle high-dimensional datasets with many potential predictors and can perform the variable selection by shrinking irrelevant coefficients to zero. This can improve the interpretability of the model and reduce its complexity. Additionally, the L1 penalty can lead to sparse solutions, meaning that only a subset of the predictors is used in the model, which can reduce the risk of overfitting.

Ridge regression is a type of linear regression that adds an L2 regularization term to the OLS loss function. Similar to LASSO, ridge regression can also address overfitting and mitigate the effects of multicollinearity. However, unlike LASSO, ridge regression (L2) does not have the ability to generate sparse solutions, which means that the parameters do not actually become many zeros. Mathematically, its loss function is as follows:

$$J(\beta) = \frac{1}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{i,t+1} - \beta_0 - \sum_{j=1}^{p} x_{ij,t}\beta_j)^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \tag{3.6}$$

where λ_2 is the L2 regularization parameter that controls the strength of the penalty term. The first term in the objective function represents the ordinary least squares (OLS) loss function, while the second term represents the L2 penalty term.

Elastic net is a type of linear regression that combines the L1 (LASSO) and L2 (ridge regression) regularization parameters in the loss function. Elastic net can be seen as a compromise between the LASSO and ridge regression methods, providing a more general and flexible approach to regularization. The loss function can be formulated as follows:

$$J(\beta) = \frac{1}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{i,t+1} - \beta_0 - \sum_{j=1}^{p} x_{ij,t}\beta_j)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \tag{3.7}$$

where λ_1 and λ_2 are the L1 regularization parameter and the L2 regularization parameter that controls the strength of the penalty term.

Penalized Regression Models have been widely used in the field of asset pricing. Rapach et al. (2015) employed the adaptive LASSO technique from statistical learning literature, aiming to identify economically linked industries within a comprehensive framework that accounts for intricate interdependencies among various industries. Chinco et al. (2017) found rare, short-lived, 'sparse' variables that predict returns using the LASSO.

Principal Component Regression and Partial Least Squares

As the number of dimensions increases, the amount of data required to represent the space accurately increases exponentially. This leads to several problems, such as the sparsity of data samples, increased computational complexity, difficulty in visualization, and overfitting. When addressing high-dimensional data, there are usually two kinds of methods: the first is shrinkage, as demonstrated by the penalized regression section above, which shrinks some parameter values near or to exactly zero; the second method is dimension reduction, such as Principal Component Regression (PCR) and Partial Least Squares (PLS). These methods can transform the data and map the high-dimensional data into a lower-dimensional space while preserving as much original information as possible.

PCR is a multivariate regression method that examines the correlation among multiple variables. It consists of two procedures: first, using principal component analysis (PCA), multiple independent variables are transformed into a small set of principal components that preserve as much information from the original variables as possible while being uncorrelated with each other. Then, these principal components are used for regression analysis.

First, we convert Equation 3.3 into a vectorized form:

$$R = Z\theta + E \tag{3.8}$$

where R is the $NT \times 1$ matrix, representing the forecasting excess return. Z is the $NT \times P$ matrix of stacked predictors, θ is the vector of the predictive coefficient, and E is the vector of residuals.

PCR and PLS condense the dimensionality of predictors to a smaller number of K. Thereafter, the predictive model could be written as:

$$R = (Z\Omega_K)\theta_K + \widetilde{E} \tag{3.9}$$

where Ω_K is the $P \times K$ matrix with columns $\omega_1, \omega_2, \ldots, \omega_K$. $Z\Omega_K$ is the dimension-reduced version of the set of predictors. Each column ω_j in the matrix represents a linear weight or feature vector used to construct the new dimensions. The objective is to find the coordinates with maximum variance among the variables for dimensionality reduction while maintaining orthogonality and eliminating duplicate information among different feature variables. The objective function of the feature vectors for PCR can be expressed as:

$$w_j = \arg\max_w \operatorname{Var}(Zw), s.t. w'w = 1, \operatorname{Cov}(Zw, Zw_l) = 0, l = 1, 2, ..., j-1. \quad (3.10)$$

It is important to note that the selection of components is not influenced by the forecasting objective. Rather, PCR places emphasis on identifying components that capture as much common variation within the predictor set as possible. PCA, as an early and classical algorithm, has been widely used in the academic field. In the field of asset pricing, Giglio et al. (2016) used PCA to extract potential omitted predictors when constructing factor pricing models. Lettau and Pelger (2020) propose a new estimator that can find asset-pricing factors by generalizing PCA with accounting for pricing error in a large-dimensional panel of financial data. Overall, there are relatively few research studies that solely use PCA as the algorithm, and most studies combine it with other algorithms for comparative analysis.

Unlike PCA, PLS's objective is seeking K linear combinations of Z that have a maximal predictive association with the forecast target. The objective function can be expressed as:

$$w_j = \arg\max_w Cov(R, Zw), s.t. w'w = 1, Cov(Zw, Zw_l) = 0, l = 1, 2, ..., j-1. \quad (3.11)$$

From the formula, we can see that compared to PCA, which only considers the internal correlation of variables, PLS introduces the research objective-stock returns and examines the correlation between variables and returns. PLS then reconstructs and reduces the coordinate system according to the dimensions most correlated with returns.

Recent studies on the PLS algorithm can be found in Kelly and Pruitt (2012), where the algorithm is applied to the study of return prediction, and it is found that the covariance between variables and predicted values is more important than the relationships among the variables. The authors suggest that some effective predictors may be ignored by the PCA algorithm due to their small variances, while PLS can improve prediction by increasing the weights of these variables. Similarly, Huang et al. (2015) introduce the PLS algorithm into the measurement of investor sentiment, using stock return volatility as an instrumental variable to extract investor sentiment and giving higher weights to proxy variables that are sensitive to investor sentiment and have strong predictive power for stock returns, and constructing a new investor sentiment index accordingly.

3.3.2 Non-Linear Models

Decision Tree and Random Forest

A decision tree is a hierarchical tree structure, where each node in the tree represents a specific condition related to the input data, and each branch represents a possible outcome based on the answer to that condition. As the tree progresses, each subsequent condition is based on the previous condition, until a final decision or outcome is reached. A path from the root node of the decision tree to the leaf nodes forms a prediction of the category of the corresponding object. There are many algorithms for decision trees, such as ID3, C45, CART, etc. These algorithms all use a top-down greedy algorithm, where each node selects the attribute with the best classification and splits the node into two or more sub-nodes, continuing the process until the tree can accurately classify the training set, or all attributes have been used.

However, the decision tree model has significant drawbacks. First, this model is prone to overfitting, leading to a low generalization ability. Second, decision trees are susceptible to sample imbalance. We need to balance the samples before training the model to avoid the situation where one category is in the absolute majority of the dataset. Third, the stability of the decision tree is low. A very small change to the dataset may result in training a completely different tree. When the number of features is large, it is difficult to solve the above problems. According to the researches, bagging can improve the model performance, so we introduce the random forest model.

Random forest is an ensemble learning method that combines multiple weak and diverse models to make the overall model performance with high accuracy by averaging or taking the majority vote. The "forest" of random forest is built in a random way consisting of many decision trees. Each tree is trained on a subset of the original dataset and each tree is unrelated to the other. The procedure of bagging for trees can be explained as follows: Given a training set $X = x_1, ..., x_N$ with responses $Y = y_1, ..., y_N$, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples: For $b = 1, ..., B$: Select, with replacement, N training examples from X, Y, and they are called X_b, Y_b. Train a classification or regression tree f_b on X_b, Y_b. After training, predictions can be made by averaging the results from all the regression trees or taking the majority vote of all the classification trees. Compared with the decision tree model, the random forest model is more robust and better able to handle noise and outliers in the data. That is to say, it can decrease the variance of the model, without increasing the bias.

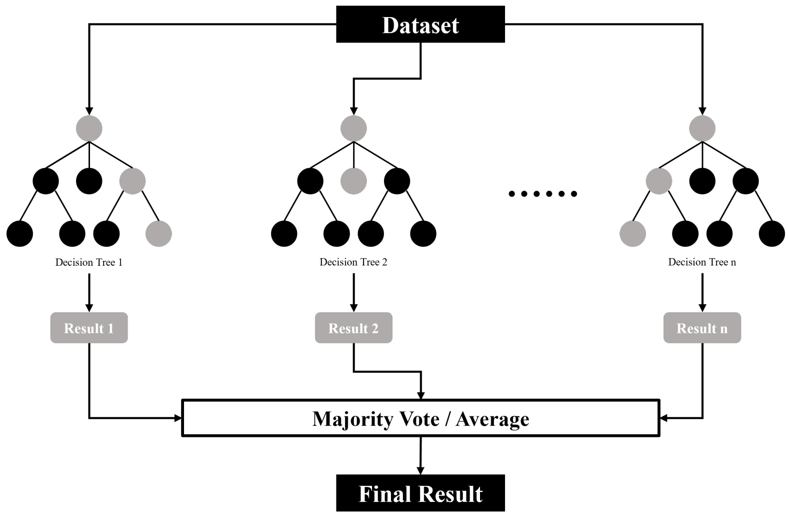
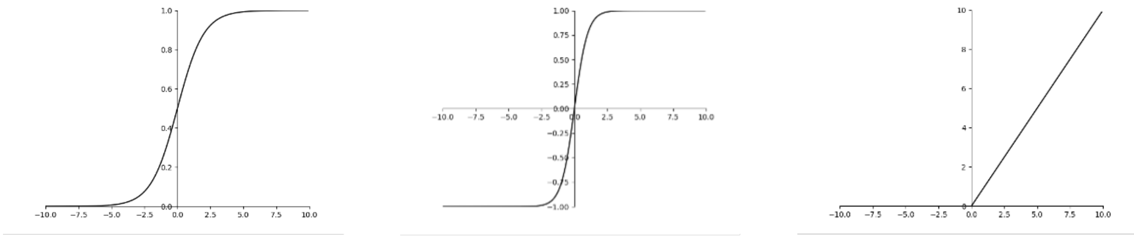


Figure 3.2: Decision Trees

(a) In the Figure 3.2, we present an example of a random forest, an ensemble learning method that improves model accuracy and stability by constructing multiple decision trees and combining their predictions. During the training process, the random forest uses bootstrap sampling to generate different training sets and randomly selects feature subsets to build each decision tree. The final prediction is derived from the average of the predictions made by all the decision trees (for regression problems)

Deep Neural Networks

Deep neural network (DNN) is a type of artificial neural network (ANN). DNN can learn and perform complex tasks by using multiple layers of interconnected nodes or "neurons". Therefore, DNN is also known as Multi Layer Perceptron (MLP). These layers allow the network to process and analyze large amounts of data, enabling it



(a) Sigmoid (b) Tanh (c) ReLu

Figure 3.3: Common activation functions

(a) In the Figure 3.3, we showcase three commonly used activation functions: Sigmoid, Tanh, and ReLU. Activation functions play a crucial role in neural networks, as they introduce non-linearity, enhancing the model's expressive power.

to recognize patterns and make predictions with a high degree of accuracy. The specific model we apply in the empirical study is feedforward neural network.

The neuron is the basic computation unit, also called a node. The node receives input, where each input has a weight, from other nodes and then computes the output. It applies a function to the weighted sum of all the inputs. The function is called the Activation Function. Some common activation functions are as Figure 3.3

Sigmoid: scale the output to $[0, 1]$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.12}$$

Tanh: scale the output to $[-1, 1]$

$$tanh(x) = 2\sigma(2x) - 1 \tag{3.13}$$

ReLu:

$$f(x) = \max(0, x) \tag{3.14}$$

In a deep neural network, each layer of neurons processes the input data, and the output from one layer serves as the input for the next layer. The input layer receives the input data and processes it, passing it to the hidden layer, which processes the data further. This process continues until the data has passed through all the hidden layers. Finally, the network generates an output or prediction based on the input.

Figure 3.5 is an example of the feedforward neural network with three hidden layers, where each hidden layer has 32, 16, and 8 neurons.

We follow Gu et al. (2020) in constructing our feedforward neural network with up to five hidden layers. The first neural network, which is the shallowest one, has a hidden layer with 32 nodes. We call it FFN1. Then the second neural network FNN2 has two hidden layers with 32 and 16 nodes separately. The third neural network FNN3 has three hidden layers with 32, 16, and 8 nodes, respectively. The fourth neural network FNN4 has four hidden layers with 32, 16, 8, and 4 nodes, respectively. And the last neural network FFN5 has five hidden layers with 32, 16, 8, 4, and 2 nodes separately. By comparing the five neural networks, we can detect the trade-offs of network complexity in the predicting problem.

To ensure computational feasibility and prevent overfitting, we employ additional types of regularization in conjunction with a ReLU activation function. Our approach for estimating the weight parameters of the neural network is to use L2 regularization by penalizing weights with large magnitudes. To train the neural networks, we apply the stochastic gradient descent (SGD) algorithm which can divide the whole training sample into smaller random subsamples.

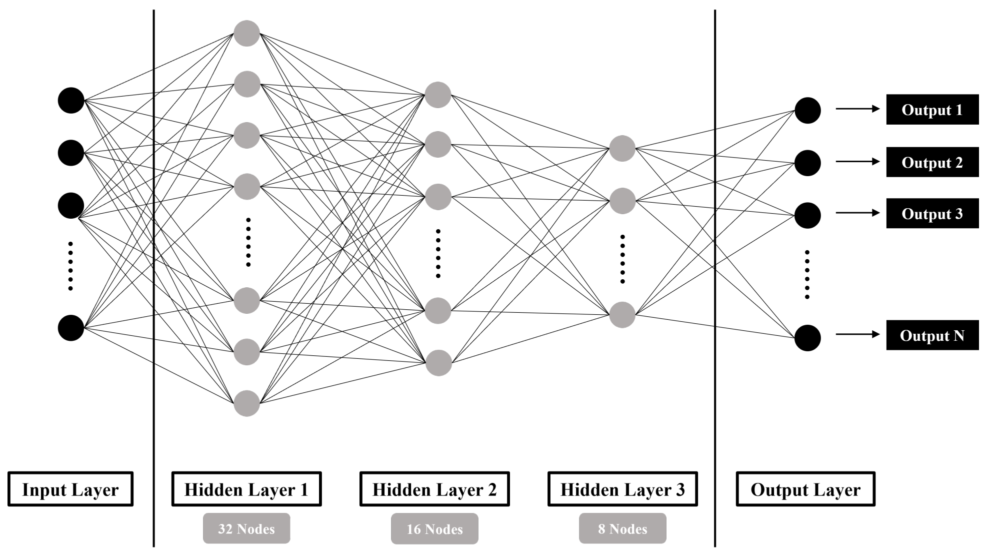


Figure 3.5: Neural Network

(a) In Figure 3.5, we present the structure of a Neural Network with three hidden layers, consisting of 32, 16, and 8 neurons, respectively. Neural networks are computational models that mimic biological neural systems for recognizing and processing complex patterns and data. These networks comprise input, hidden, and output layers, with nodes (neurons) connected by weights between the layers.

3.3.3 Performance Evaluation

We conduct whether these models differ in predictive performance from two aspects: First, we used the out-of-sample R^2, also known as the coefficient of determination, to assess predictive performance for individual stock excess returns. Then, we use Sharpe Ratio to evaluate predictive performance at the portfolio level.

R-squared

In machine-learning models, the variance of the dependent variable (Y) can be referred to as the total sum of squares (TSS), which consists of two parts: the model sum of squares (MSS), representing the information that can be explained by the model, and the residual sum of squares (RSS), representing the information that cannot be explained by the model. The definition of R^2 is the proportion of the variance of the dependent variable that is explained by the predicted values, which measures the goodness of fit of the predicted values to the true values. The formula for R^2 is:

$$R^2_{oos} = 1 - \frac{\sum_{i,t}^{n_{test}} (y_{i,t+1} - \hat{y}_{i,t+1})^2}{\sum_{i,t}^{n_{test}} (y_{i,t+1} - \bar{y}_{i,t+1})^2} \tag{3.15}$$

where $y_{i,t+1}$ and $\hat{y}_{i,t+1}$ are the actual and predicted values of the dependent variable for the i^{th} observation in the test dataset, respectively. $\bar{y}_{i,t+1}$ is the mean of the dependent variable in the test dataset. The range of the R^2 value is from $(-\infty, 1]$, where a higher value indicates a better predictive ability of the model. When the model predicts the returns of stocks for all periods perfectly, R^2 equals 1.

We also follow Gu et al. (2020) to evaluate the forecasting performance of machine learning models with a different kind of out-of-sample R^2; the formula is:

$$R^2_{oos} = 1 - \frac{\sum_{i,t}^{n_{test}} (y_{i,t+1} - \hat{y}_{i,t+1})^2}{\sum_{i,t}^{n_{test}} (y_{i,t+1})^2} \tag{3.16}$$

The difference of the R^2 is the denominator is the sum of squared excess returns without demeaning.

Sharpe Ratio

In this paper, we use the Sharpe Ratio to assess the predictive accuracy of the machine learning models employed for portfolio analysis. Sharpe Ratio is a widespread measure of risk-adjusted return used in finance regions. By using Sharpe Ratio, we can evaluate the risk-adjusted performance of an investment. The function is shown as:

$$Sharpe\ Ratio = \frac{R_p - R_f}{\sigma_p} \tag{3.17}$$

where R_p is the mean return of portfolio, R_f is the risk-free rate, and σ_p is standard deviation of the portfolio's excess return.

A higher Sharpe Ratio indicates better risk-adjusted performance, as it implies that the portfolio earns a higher return per unit of risk taken.

Chapter 4

Empirical Analysis

In this chapter, we start by comparing the stock-level prediction performance of both linear and non-linear models we elaborated above via out-of-sample predictive R^2 and exploring the contribution of each variable within each model. Then, we employ portfolio analysis to examine the portfolio-level prediction performance for each model. Please be informed that our empirical analysis was conducted within the Python environment, utilizing packages such as numpy, pandas, matplotlib, and scikit-learn among others.

4.1 Empirical Test Results

Table 4.1 presents the comparison of each model by means of their predictive R^2. Totally, we include 12 models within the comparison: ordinary least squares (OLS) with all variables, ridge regression (RR), LASSO, elastic net (ELNT), PCR, PLS, random forests (RF), and feedforward neural network with different layers from 1 to 5 (FFN1, FFN2, FFN3, FFN4, FFN5).

Table 4.1: Monthly prediction performance (Percentage R^2_{oos} and R^2_{is})

	OLS	RR	LASSO	ELNT	PCR	PLS	RF	FFN1	FFN2	FFN3	FFN4	FFN5
R^2_{is}	29.26	26.90	24.56	24.57	24.64	24.64	82.06	49.34	52.49	55.15	55.37	51.89
R^2_{is-dm}	45.46	27.64	24.60	24.61	24.67	24.67	85.01	57.91	60.52	62.73	62.91	60.02
R^2_{oos}	-269	6.52	23.84	23.84	23.81	23.78	35.84	39.50	33.99	42.51	42.07	37.15
R^2_{oos-dm}	-127	6.73	24.00	24.00	23.98	23.94	41.51	47.83	43.08	50.42	50.04	45.80

(a) In Table 4.1, R^2_{is} represents the in-sample R^2 of each model and R^2_{is-dm} represents the in-sample R^2 with demeaning. Similarly, R^2_{oos} and R^2_{oos-dm} are out of sample R^2 without and with demeaning respectively of each model.

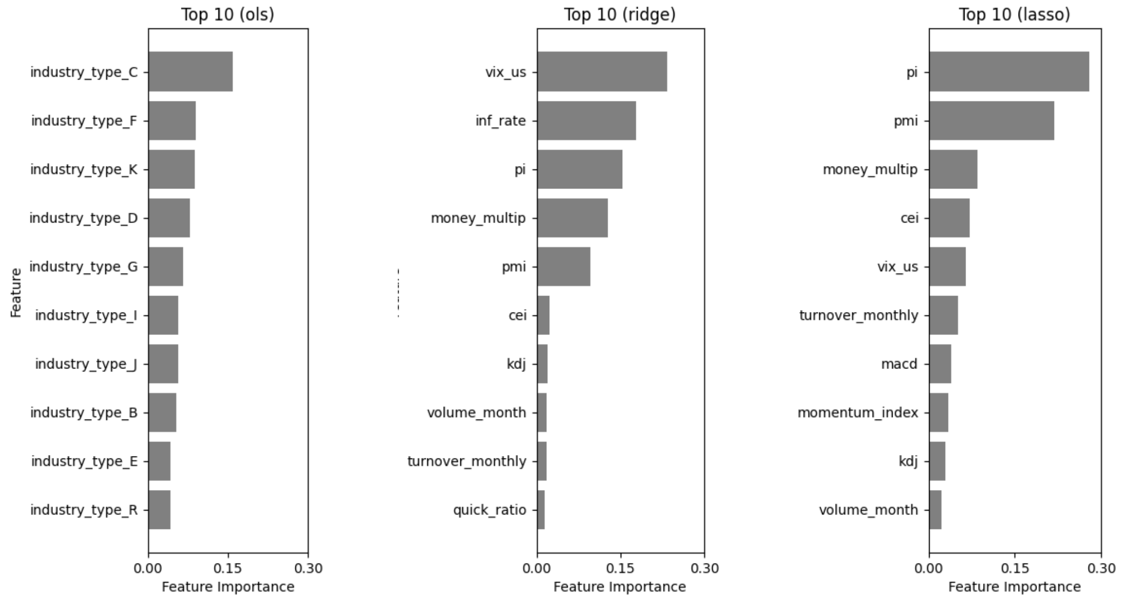
OLS model generates positive in-sample R^2 values, which are similar to other linear models, but negative out-of-sample R^2 values of -269% and -127%, showing a naive prediction of a negative value to all stocks. It is unsurprising since the absence of regularization in OLS can make it highly vulnerable to overfitting. Therefore, regularization via dimension reduction can significantly improve prediction performance. First, by shrinking the coefficients of the less important features towards zero and reducing the impact of highly correlated features, LASSO and elastic net reduce the number of factors to 38 and 40 respectively, and raise the out-of-sample R^2 to more than 23%. Second, by forming a few linear combinations of predictors, PCR and PLS also raise the out-of-sample R^2 to over 23%. In detail, PCR uses 30 to 50 components and PLS finds a smaller group of 9 components.

Neural networks are the best-performing models overall. The out-of-sample R^2 ranges from 30% to 50%, among which FFN3 has the highest R^2 of 50.42%. The results suggest that a more complex neural network model does not necessarily lead to better performance, which is a common phenomenon because the additional layers may introduce noise or overfit the model to the training data, leading to poorer performance on new or unseen data. Random forest is competitive with neural networks with the out-of-sample R^2 of 35.84% and 41.51%. The forests generally estimate deep trees, with about 20 layers on average.

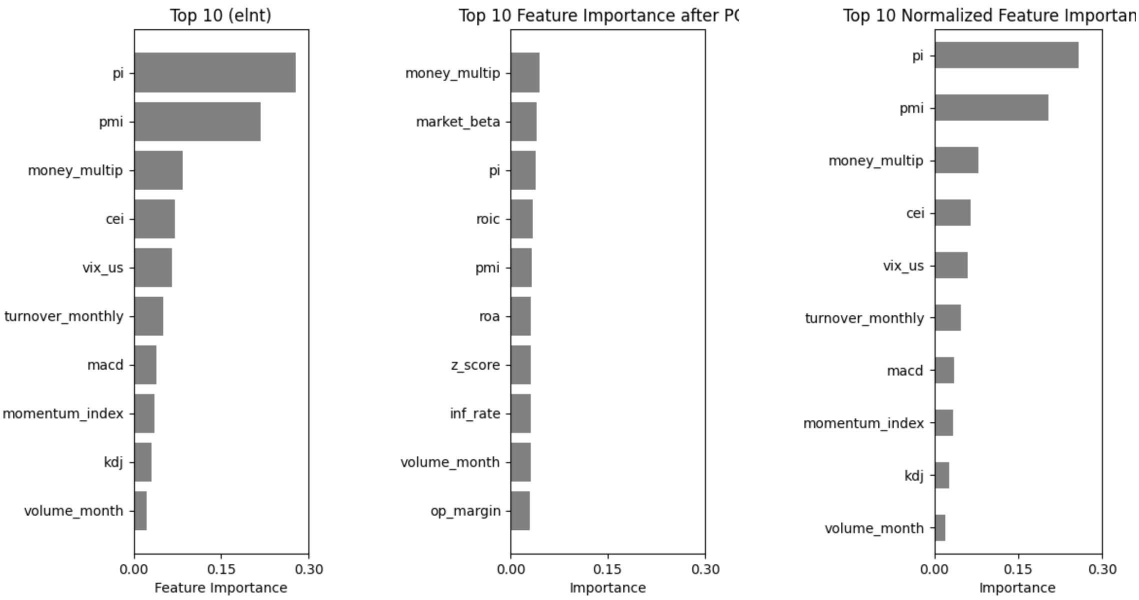
4.2 Feature Importance

Following Drobetz et al. (2021), we utilize a two-step method to calculate the variable importance matrix for each model. In the first step, we determine the absolute variable importance by evaluating the reduction in R^2 that results from setting all values of a specific predictor to zero within the training sample. In the second step, we normalize the absolute variable importance measures so that they add up to 1, indicating the proportional contribution of each variable to a model. Figure 4.1 and Figure 4.2 depict the top 10 relative variable importance measures for each model[1]. We find that the different forecasting models identify almost the same variables as the most informative, except the OLS model.

[1]For the whole relative variable importance measures, please see Appendix A



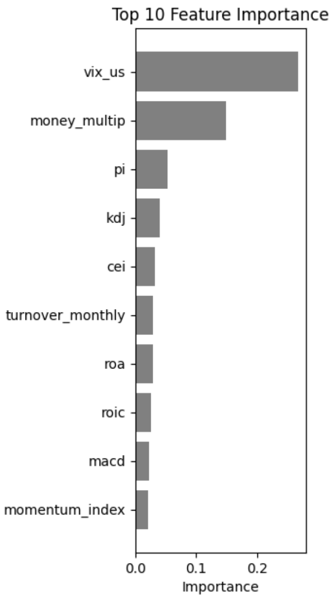
(a) ols(10) (b) ridge(10) (c) LASSO(10)



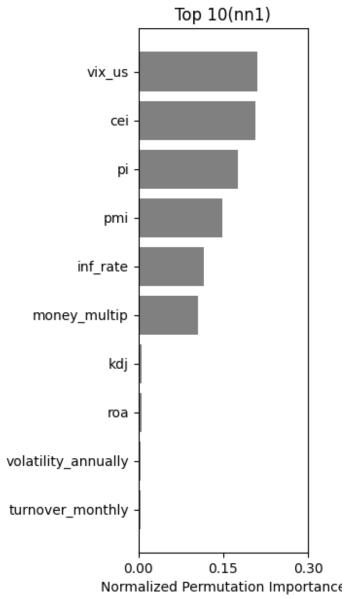
(d) elnt(10) (e) pcr(10) (f) pls(10)

Figure 4.1: Top 10 Feature Importance(a)

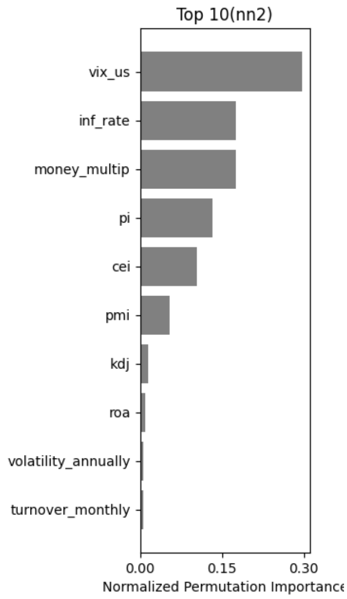
(a) Figure 4.1 contains the top 10 feature importance obtained from six different machine learning models. (totaling 12 models)



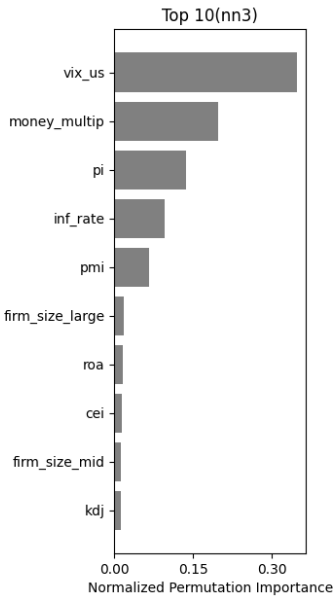
(a) rf(10)



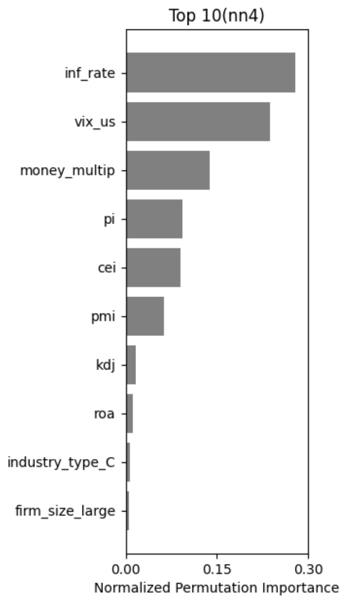
(b) nn1(10)



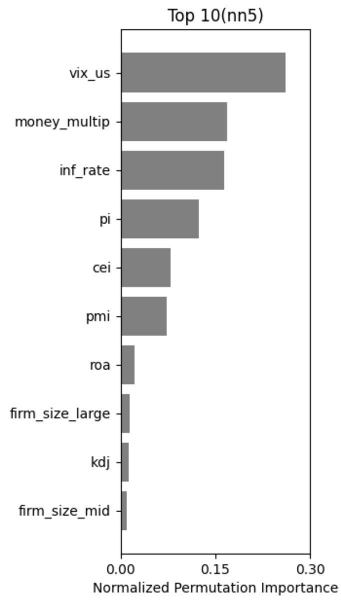
(c) nn2(10)



(d) nn3(10)



(e) nn4(10)



(f) nn5(10)

Figure 4.2: Top 10 Feature Importance(b)

(a) Figure 4.2 contains the top 10 feature importance obtained from six different machine learning models. (totaling 12 models)

Most of the forecasting models favor the macroeconomics predictors, including ridge regression (RR), PCR, PLS, random forest (RF), and neural networks (FFN1-FFN5). For RR, RF, and FNNs, *vix_us* has the largest variable importance. Meanwhile, these models also put significant weight on Chinese inflation rate (*inf_rate*) and money multiplier (*money_multi*), suggesting these macroeconomic indicators are also significant. In addition, PLS, ELNT, and LASSO strongly favor the Chinese prosperity index (*pi*) and Chinese Purchasing Managers' Index (*pmi*). Considering the characteristics of Chinese stock market, we think it is reasonable that macroeconomic variables are more influential. First, China is an economy with a strong dependence on the global economy. Therefore, factors such as global economic fluctuations, trade frictions and changes in international financial markets can directly or indirectly affect the performance of China's stock market. Second, the structural characteristics of China's stock market also increase the likelihood that it will be affected by macroeconomic factors. Many companies in the Chinese stock market are state-owned or have close relationships with the government, which can lead to changes in stock prices based on government policies or announcements. For instance, on December 4, 2012, the Central Committee of the Chinese government issued the 'Eight-point Regulation', which strictly stipulated the standards for official receptions, specifically prohibiting the consumption of high-end alcohol during such events. As Kweichow Moutai (Stock Index: 600519.SS), the highest-grade baijiu brand characteristic of China, was significantly affected by this policy, its stock price fell nearly 50% within a year. Third, the investor structure of the Chinese stock market is also an important factor contributing to its exposure to macroeconomic factors. The Chinese stock market is dominated by a large proportion of retail investors, who are more sensitive to news and macroeconomic events compared to institutional investors. Finally, the Chinese stock market is highly speculative, with many investors relying on rumors and market gossip to make investment decisions, which can further amplify the impact of macroeconomic factors.

For stock-level characteristics, variables *turnover_monthly*, *macd*, *momentum_index*, *kdj*, and *volume_month*, which all belong to the category **Trading Info** have high priority in RR, LASSO, ELNT, and PLS. However, our findings in machine learning models contrast those in Gu et al. (2020) for the US market. Random trees and neural networks are highly skewed toward macroeconomic characteristics. Compared with other stock-level characteristics, variables in the group **Profitability** and **Trading Info**, namely *kdj*, *roa*, and *turnover_monthly* are relevantly significant. We believe that this outcome is in line with the characteristics of the Chinese stock market, primarily because it is dominated by retail investors who tend to make decisions based on short-term trends and news rather than long-term fundamentals. As a result, indicators that can reflect trading information are more influential on the market. This is particularly important given the lack of mature regulations and institutions in the market, which can make it more volatile and susceptible to sudden changes in sentiment.

4.3 Portfolio Analysis

So far, we have discussed using multiple machine-learning models to predict stock-level excess returns. However, most rational investors tend to hold investment portfolios (such as mutual funds and ETFs). The advantage of holding portfolios is that they can minimize and diversify the idiosyncratic risk caused by noise in individual stock predictions as much as possible. To analyze the forecasting power of machine learning models at the portfolio level, we employed portfolio analysis in this section. Last but not least, another reason for doing so is to test the predictive accuracy using the portfolio approach.

From January 2003 to September 2022, we conduct an analysis on a monthly basis in which we sort the predicted returns of various models in the entire sample in ascending order. Subsequently, we partition the sorted returns into ten equal parts, resulting in the formation of ten investment portfolios. These portfolios comprise p_low, which consists of the bottom 10% of stocks by predicted return, and p_high, which consists of the top 10% of stocks by predicted return. The intermediate portfolios range from the 2^{nd} to the 9^{th}. Meanwhile, we build a long-short investment portfolio, designated as H-L, by longing stocks in the p_high portfolio and short-selling those in the p_low portfolio. As depicted in Table 4.2, we compute equal-weighted and value-weighted monthly predicted excess returns, monthly realized excess returns, standard deviations of actual returns, and annualized Sharpe ratios for each of the distinct portfolios.

We construct the equal-weighted portfolio return with:

$$\hat{r}^{p}_{t+1} = \sum_{i=1}^{N} \frac{1}{N} * \hat{r}_{i,t+1} \tag{4.1}$$

The value-weighted portfolio return with:

$$\hat{r}^{p}_{t+1} = \sum_{i=1}^{N} w^{p}_{i,t} * \hat{r}_{i,t+1} \tag{4.2}$$

Table 4.2: portfolio analysis result

	equal-weighted				value-weighted			
	Pred(%)	Avg(%)	Std(%)	SR	Pred(%)	Avg(%)	Std(%)	SR
ols								
Low	-3.13E+11	-2.16	16.62	-0.45	1.5E+11	-3.10	12.65	-0.85
2	-7.71E+10	-2.20	15.11	-0.50	4.92E+11	-2.56	12.08	-0.74
3	-4.87E+10	-3.00	16.03	-0.65	5.9E+11	-3.13	14.35	-0.76
4	-4.87E+10	-3.00	16.03	-0.65	4.94E+11	-2.55	13.49	-0.65
5	-4.87E+10	-1.71	14.69	-0.40	5.17E+11	-2.53	12.78	-0.69
6	-4.87E+10	-1.50	15.27	-0.34	5.61E+11	-2.72	12.14	-0.78
7	-1.90E+10	-2.07	15.26	-0.47	5.37E+11	-3.11	12.60	-0.85
8	2.64E+10	-2.05	14.40	-0.49	4.59E+11	-2.89	12.54	-0.80
9	1.69E+11	-2.17	22.34	-0.34	7,72E+11	-3.45	12.91	-0.93
High	4.09E+11	-2.03	16.63	-0.42	9.39E+11	-3.08	13.31	-0.80
H-L	7.22E+11	0.13	13.41	0.03	7.88E+11	0.03	15.35	0.01
ridge								
Low	-5.79	-3.46	16.27	-0.74	-14.33	-3.33	13.60	-0.85
2	-3.62	-2.68	15.26	-0.61	-5.12	-3.05	13.67	-0.77
3	-2.88	-2.34	15.29	-0.53	-4.93	-3.05	13.36	-0.79
4	-2.37	-2.22	14.43	-0.53	-4.24	-2.81	12.96	-0.75
5	-1.95	-2.10	14.59	-0.49	-4.04	-2.97	12.57	-0.81
6	-1.58	-1.94	13.77	-0.48	-3.48	-3.03	12.41	-0.85
7	-1.22	-1.81	14.75	-0.43	-2.16	-2.77	12.67	-0.76
8	-0.87	-1.70	14.10	-0.41	-1.49	-2.69	12.35	-0.76
9	-0.45	-1.49	14.00	-0.36	-1.71	-2.31	11.79	-0.68
High	-0.44	-1.12	26.06	-0.15	0.27	-2.73	11.79	-0.80
H-L	5.35	2.14	16.48	0.45	14.60	0.60	16.44	0.13
lasso								
Low	-6.93	-3.43	15.93	-0.74	-9.03	-3.16	13.56	-0.80
2	-2.91	-2.68	14.97	-0.62	-3.56	-3.32	13.61	-0.84
3	-1.65	-2.38	14.64	-0.56	-3.01	-3.17	13.07	-0.84
4	-0.82	-2.08	14.35	-0.50	-2.67	-2.90	12.95	-0.77
5	-0.18	-2.09	14.57	-0.50	-2.78	-3.29	12.61	-0.90
6	0.36	-2.01	13.87	-0.50	-2.19	-2.86	12.36	-0.80
7	0.88	-1.92	13.49	-0.49	-1.71	-2.66	12.31	-0.74
8	1.42	-1.64	13.64	-0.41	-1.60	-2.89	12.71	-0.78
9	2.09	-1.67	13.75	-0.42	-1.16	-2.60	12.47	-0.72
High	3.84	-0.97	27.81	-0.12	-0.27	-2.50	13.03	-0.66
H-L	10.77	2.46	17.08	0.49	8.76	0.66	17.30	0.13
elnt								
Low	-6.95	-3.40	15.92	-0.74	-23.46	-3.22	13.23	-0.84

Table 4.2: portfolio analysis result

	equal-weighted				value-weighted			
	Pred(%)	Avg(%)	Std(%)	SR	Pred(%)	Avg(%)	Std(%)	SR
2	-2.91	-2.68	14.98	-0.62	-1.73	-3.08	13.28	-0.80
3	-1.65	-2.38	14.62	-0.56	0.39	-2.86	13.19	-0.75
4	-0.82	-2.08	14.35	-0.50	1.64	-2.70	12.69	-0.74
5	-0.18	-2.10	14.61	-0.49	2.06	-3.15	11.87	-0.92
6	-0.37	-2.01	13.85	-0.50	2.10	-2.85	12.44	-0.79
7	-0.88	-1.92	13.49	-0.49	3.24	-3.00	12.32	-0.84
8	1.42	-1.64	13.65	-0.42	3.38	-2.51	12.19	-0.71
9	2.09	-1.68	13.75	-0.42	4.55	-2.61	12.56	-0.72
High	3.86	-0.96	27.81	-0.12	8.67	-2.63	13.73	-0.66
H-L	10.81	2.44	17.09	0.49	32.13	0.59	17.51	0.12
pcr								
Low	-7.31	-3.33	15.88	-0.72	-24.52	-3.10	13.14	-0.82
2	-3.09	-2.70	14.98	-0.62	-1.60	-3.18	13.50	-0.82
3	-1.77	-2.43	14.55	-0.58	0.19	-3.15	13.12	-0.83
4	-0.88	-2.03	14.32	-0.49	1.63	-2.68	12.76	-0.73
5	-0.20	-2.12	14.40	-0.51	2.02	-2.96	11.93	-0.86
6	0.38	-2.00	13.63	-0.50	2.67	-3.05	12.49	-0.84
7	0.93	-1.87	13.92	-0.47	3.12	-2.86	12.15	-0.82
8	1.52	-1.77	13.53	-0.45	3.79	-2.55	12.20	-0.72
9	2.27	-1.59	14.04	-0.39	4.42	-2.56	12.69	-0.70
High	4.25	-0.96	27.79	-0.12	9.05	-2.68	13.61	-0.68
H-L	11.56	2.37	17.10	0.48	33.57	0.42	17.31	0.08
pls								
Low	-7.29	-3.35	15.86	-0.73	-24.56	-3.16	13.15	-0.83
2	-3.09	-2.71	14.99	-0.62	-1.57	-3.16	13.49	-0.81
3	-1.77	-2.40	14.62	-0.57	0.31	-3.03	13.21	-0.79
4	-0.89	-2.06	14.31	-0.50	1.54	-2.75	12.74	-0.75
5	-0.21	-2.10	14.36	-0.50	1.96	-2.89	12.12	-0.83
6	0.36	-2.04	13.59	-0.52	2.73	-2.90	12.26	-0.82
7	0.92	-1.89	13.75	-0.48	3.26	-3.05	12.40	-0.85
8	1.51	-1.73	13.75	-0.44	3.89	-2.56	12.00	-0.74
9	2.27	-1.65	13.94	-0.41	4.19	-2.60	12.76	-0.71
High	4.30	-0.93	27.90	-0.11	9.14	-2.64	13.53	-0.68
H-L	11.59	2.42	17.09	0.49	33.7	0.70	17.32	0.14
rf								
Low	-7.58	-3.81	18.93	-0.70	-3.75	-3.58	12.55	-0.99
2	-6.03	-3.24	14.36	-0.78	-2.71	-3.32	12.68	-0.91
3	-5.26	-2.57	15.09	-0.59	-2.36	-3.41	13.05	-0.94
4	-4.65	-2.45	14.65	-0.58	-1.96	-2.83	13.02	-0.75
5	-4.08	-2.26	13.65	-0.57	-1.67	-2.76	13.48	-0.71
6	-3.51	-2.00	14.01	-0.49	-1.24	-3.11	13.49	-0.80
7	-2.87	-1.84	13.75	-0.46	-0.96	-2.56	12.77	-0.69
8	-2.07	-1.71	13.86	-0.43	-0.80	-2.31	12.56	-0.64
9	-0.85	-1.31	14.08	-0.32	-0.61	-2.73	12.68	-0.75
High	3.40	0.34	25.62	0.05	-0.22	-2.49	13.01	-0.66
H-L	10.98	4.15	16.08	0.89	3.53	1.09	16.04	0.24
ffn1								
Low	-3.10	-4.03	17.05	-0.61	-17.18	-2.72	12.27	-0.77
2	0.17	-2.81	14.55	-0.67	1.10	-3.38	13.57	-0.86
3	1.28	-2.44	14.43	-0.58	2.09	-3.01	13.39	-0.78
4	2.06	-2.20	16.33	-0.47	2.64	-3.05	13.03	-0.81
5	2.72	-2.25	14.17	-0.55	3.19	-2.90	13.06	-0.77
6	3.30	-1.99	14.82	-0.47	3.67	-2.83	12.64	-0.78
7	3.88	-1.73	15.12	-0.40	3.72	-2.55	13.22	-0.67

Table 4.2: portfolio analysis result

	equal-weighted				value-weighted			
	Pred(%)	Avg(%)	Std(%)	SR	Pred(%)	Avg(%)	Std(%)	SR
8	4.53	-1.68	14.32	-0.41	4.23	-3.04	13.17	-0.80
9	5.40	-1.58	15.96	-0.34	4.36	-3.07	12.54	-0.85
High	7.92	-1.57	23.43	-0.17	6.01	-3.05	12.75	-0.82
H-L	11.02	2.46	15.68	0.54	23.19	-0.33	16.03	-0.07
ffn2								
Low	-11.89	-2.84	16.02	-0.61	-23.27	-3.03	12.84	-0.82
2	-8.59	-2.50	14.99	-0.58	-8.82	-2.56	13.47	-0.66
3	-7.27	-2.43	14.60	-0.58	-7.70	-3.08	12.77	-0.84
4	-6.31	-2.20	14.37	-0.53	-6.75	-2.88	12.51	-0.80
5	-5.51	-1.98	14.50	-0.47	-6.41	-2.66	12.86	-0.71
6	-4.76	-1.96	14.76	-0.46	-5.93	-2.92	12.96	-0.78
7	-4.00	-1.75	16.53	-0.36	-5.59	-2.80	13.43	-0.72
8	-3.11	-1.98	13.96	-0.49	-5.25	-3.18	12.91	-0.85
9	-1.90	-1.79	14.81	-0.42	-4.37	-3.18	12.57	-0.88
High	1.07	-1.40	24.89	-0.19	-2.25	-3.08	12.73	-0.84
H-L	12.96	1.44	16.08	0.31	21.02	-0.05	16.18	-0.01
ffn3								
Low	-12.83	-2.97	15.25	-0.67	-21.33	-2.78	12.92	-0.75
2	-8.97	-2.49	16.38	-0.52	-8.12	-2.95	13.25	-0.77
3	-7.59	-2.38	14.68	-0.56	-6.54	-3.03	13.03	-0.81
4	-6.62	-2.21	13.98	-0.54	-5.90	-3.20	12.84	-0.86
5	-5.79	-2.12	15.48	-0.47	-5.28	-2.92	12.75	-0.79
6	-5.02	-1.89	14.65	-0.44	-4.72	-2.86	12.99	-0.76
7	-4.20	-1.87	16.65	-0.38	-3.80	-2.77	13.39	-0.72
8	-3.27	-1.59	16.81	-0.32	-3.27	-2.92	13.23	-0.76
9	-2.03	-1.84	14.78	-0.43	-2.16	-2.94	12.63	-0.81
High	0.85	-1.46	22.15	-0.22	0.80	-3.20	12.12	-0.91
H-L	13.68	1.51	14.73	0.36	22.13	-0.42	15.71	-0.09
ffn4								
Low	-11.78	-2.83	15.70	-0.62	-12.69	-3.19	13.59	-0.81
2	-8.10	-2.39	14.93	-0.55	-8.19	-2.81	13.67	-0.71
3	-6.61	-2.05	15.14	-0.47	-6.73	-2.80	13.34	-0.73
4	-5.51	-2.16	14.78	-0.51	-5.86	-2.57	13.45	-0.66
5	-4.53	-1.97	15.00	-0.45	-5.08	-2.44	13.41	-0.63
6	-3.59	-2.01	15.08	-0.46	-4.09	-3.38	13.14	-0.89
7	-2.64	-1.99	14.90	-0.46	-23.09	-2.95	12.91	-0.79
8	-1.60	-1.83	14.87	-0.42	-2.13	-2.91	12.68	-0.80
9	-0.30	-1.76	21.45	-0.28	-1.38	-2.91	12.25	-0.82
High	2.26	-1.84	19.05	-0.33	0.21	-3.19	11.63	-0.95
H-L	14.04	0.99	13.85	0.25	12.90	0.00	15.38	0.00
ffn5								
Low	-11.35	-2.67	18.61	-0.49	-28.82	-2.96	12.55	-0.81
2	-5.58	-2.57	14.80	-0.60	-6.89	-3.47	13.39	-0.89
3	-3.74	-2.23	14.72	-0.52	-4.94	-2.97	13.11	-0.78
4	-2.67	-2.12	14.82	-0.49	-3.16	-2.92	12.93	-0.78
5	-1.87	-2.13	14.74	-0.50	-2.36	-2.74	13.20	-0.71
6	-1.17	-1.84	15.49	-0.41	-1.19	-2.89	12.77	-0.78
7	-0.49	-1.90	14.34	-0.45	-0.75	-3.04	12.78	-0.82
8	0.24	-1.93	14.35	-0.46	0.61	-2.98	13.03	-0.79
9	1.20	-1.86	16.03	-0.40	0.93	-2.81	12.89	-0.75
High	2.71	-1.59	22.55	-0.24	2.24	-2.52	12.59	-0.69
H-L	14.06	1.08	15.93	0.23	31.06	0.44	15.99	0.10

In general, the patterns for equal-weighted and value-weighted decile portfolios are similar, with their realized excess returns being nearly monotonically increasing in relation to the average predicted excess returns (except for a few value-weighted investment portfolios constructed based on FFN models, which is similar to the findings obtained by Gu et al. (2020) in the empirical analysis of the U.S. stock market). Evaluating the performance based on the Sharpe Ratio reveals that the equal-weighted schemes for all machine learning models significantly surpass the value-weighted schemes. We believe the primary reason for this observation is that smaller companies predominantly contribute to the excess returns.

According to Table 3.1, the average equal-weighted realized monthly excess return for this database is -2% with a standard deviation of 16.22%, resulting in an annualized Sharpe ratio of -0.43. Therefore, we can conclude that the H-L investment portfolios constructed based on machine-learning models are capable of generating statistically and economically significant profits, with the exception of a few value-weighted investment portfolios based on neural networks.

When comparing the equal-weighted H-L investment portfolios for different models, the similar Sharpe Ratio performance of linear regression models suggests a limited relationship between statistical measurement and financial profitability. The results further reveal that the random forest model-based H-L portfolio delivers the best performance, boasting an annualized Sharpe ratio of 0.89. Therefore, it is posited that incorporating non-linear relationships into asset pricing models can lead to improved predictions of stock excess returns and the construction of superior-performing investment portfolios.

Chapter 5

Conclusion

Utilizing market and fundamental data of Chinese A-shares, we examine the predictive performance of machine learning methods in forecasting stock excess returns. Our investigation encompasses variable selection, dimensionality reduction techniques, tree-based models, and neural networks. To account for the distinctive characteristics of the Chinese stock market, we adjust the set of indicators based on Gu et al. (2020) and Hou et al. (2015). Our empirical research indicates that, regardless of whether it is at the stock-level or portfolio-level, non-linear machine learning models outperform linear models in predicting stock excess returns. This demonstrates that incorporating non-linear relationships into asset pricing models can effectively improve the accuracy of predictions. However, we note that ”deep” learning does not consistently outperform ”shallow” learning due to the limited amount of data and low signal-to-noise ratio in asset pricing. Moreover, our findings indicate that macroeconomic variables are the most prominent factor group among almost all the models, followed by **Trade Info**. This can be viewed as a reflection of the unique traits of the Chinese stock market, which is dominated by retail investors who tend to be speculative in their decision-making process.

Reference

[1] Edward I Altman. “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy”. In: *The journal of finance* 23.4 (1968), pp. 589–609.

[2] Hossein Asgharian and Sonnie Karlsson. “Evaluating a non-linear asset pricing model on international data”. In: *International Review of Financial Analysis* 17.3 (2008), pp. 604–621.

[3] Ravi Bansal, David A Hsieh, and S Viswanathan. “A new approach to international arbitrage pricing”. In: *The Journal of Finance* 48.5 (1993), pp. 1719–1747.

[4] Rolf W Banz. “The relationship between return and market value of common stocks”. In: *Journal of financial economics* 9.1 (1981), pp. 3–18.

[5] Sanjoy Basu. “Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis”. In: *The journal of Finance* 32.3 (1977), pp. 663–682.

[6] Laxmi Chand Bhandari. “Debt/equity ratio and expected common stock returns: Empirical evidence”. In: *The journal of finance* 43.2 (1988), pp. 507–528.

[7] Nusret Cakici, Kalok Chan, and Kudret Topyan. “Cross-sectional stock return predictability in China”. In: *The European Journal of Finance* 23.7-9 (2017), pp. 581–605.

[8] Louis KC Chan, Yasushi Hamao, and Josef Lakonishok. “Fundamentals and stock returns in Japan”. In: *The journal of finance* 46.5 (1991), pp. 1739–1764.

[9] David A Chapman. “Approximating the asset pricing kernel”. In: *The Journal of Finance* 52.4 (1997), pp. 1383–1410.

[10] Jian Chen et al. “International volatility risk and Chinese stock return predictability”. In: *Journal of International Money and Finance* 70 (2017), pp. 183–203.

[11] Alex Chinco, Adam D Clark-Joseph, and Mao Ye. “Sparse signals in the cross-section of returns”. In: *The Journal of Finance* 74.1 (2019), pp. 449–492.

[12] Francis X Diebold and Robert S Mariano. “Comparing forecast accuracy”. In: *Journal of Business and* (1995).

[13] George Warren Douglas. *Risk in the equity markets: An empirical appraisal of market efficiency.* Yale University, 1967.

[14] Wolfgang Drobetz and Tizian Otto. “Empirical asset pricing via machine learning: evidence from the European stock market”. In: *Journal of Asset Management* 22 (2021), pp. 507–538.

[15] Eugene F Fama and Kenneth R French. “Common risk factors in the returns on stocks and bonds”. In: *Journal of financial economics* 33.1 (1993), pp. 3–56.

[16] Eugene F Fama and Kenneth R French. “The capital asset pricing model: Theory and evidence”. In: *Journal of economic perspectives* 18.3 (2004), pp. 25–46.

[17] Eugene F Fama and Kenneth R French. “The cross-section of expected stock returns”. In: *the Journal of Finance* 47.2 (1992), pp. 427–465.

[18] Eugene F Fama and James D MacBeth. “Risk, return, and equilibrium: Empirical tests”. In: *Journal of political economy* 81.3 (1973), pp. 607–636.

[19] Joachim Freyberger, Andreas Neuhierl, and Michael Weber. “Dissecting characteristics nonparametrically”. In: *The Review of Financial Studies* 33.5 (2020), pp. 2326–2377.

[20] Stefano Giglio and Dacheng Xiu. “Asset pricing with omitted factors”. In: *Journal of Political Economy* 129.7 (2021), pp. 1947–1990.

[21] John M Griffin and Michael L Lemmon. “Book-to-market equity, distress risk, and stock returns”. In: *The Journal of Finance* 57.5 (2002), pp. 2317–2336.

[22] Shihao Gu, Bryan Kelly, and Dacheng Xiu. "Empirical asset pricing via machine learning". In: *The Review of Financial Studies* 33.5 (2020), pp. 2223–2273.

[23] Campbell R Harvey, Yan Liu, and Heqing Zhu. "… and the cross-section of expected returns". In: *The Review of Financial Studies* 29.1 (2016), pp. 5–68.

[24] Kewei Hou, Chen Xue, and Lu Zhang. "Digesting anomalies: An investment approach". In: *The Review of Financial Studies* 28.3 (2015), pp. 650–705.

[25] Grace Xing Hu et al. "Fama–French in China: size and value factors in Chinese stock returns". In: *International Review of Finance* 19.1 (2019), pp. 3–44.

[26] Dashan Huang et al. "Investor sentiment aligned: A powerful predictor of stock returns". In: *The Review of Financial Studies* 28.3 (2015), pp. 791–837.

[27] Michael C Jensen, Fischer Black, and Myron S Scholes. "The capital asset pricing model: Some empirical tests". In: (1972).

[28] Steven J Jordan, Andrew J Vivian, and Mark E Wohar. "Forecasting returns: new European evidence". In: *Journal of Empirical Finance* 26 (2014), pp. 76–95.

[29] Bryan Kelly and Seth Pruitt. "Market expectations in the cross-section of present values". In: *The Journal of Finance* 68.5 (2013), pp. 1721–1756.

[30] Gregory Koutmos and Johan Knif. "Estimating systematic risk using time varying distributions". In: *European Financial Management* 8.1 (2002), pp. 59–73.

[31] Josef Lakonishok and Alan C Shapiro. "Systematic risk, total risk and size as determinants of stock market returns". In: *Journal of Banking & Finance* 10.1 (1986), pp. 115–132.

[32] Martin Lettau and Markus Pelger. "Estimating latent asset-pricing factors". In: *Journal of Econometrics* 218.1 (2020), pp. 1–31.

[33] Moshe Levy and Richard Roll. "The market portfolio may be mean/variance efficient after all: The market portfolio". In: *The Review of Financial Studies* 23.6 (2010), pp. 2464–2491.

[34] Jonathan Lewellen. "The cross section of expected stock returns". In: *Forthcoming in Critical Finance Review, Tuck School of Business Working Paper* 2511246 (2014).

[35] Qi Lin. "Noisy prices and the Fama–French five-factor asset pricing model in China". In: *Emerging Markets Review* 31 (2017), pp. 141–163.

[36] John Lintner. "Security prices, risk, and maximal gains from diversification". In: *The journal of finance* 20.4 (1965), pp. 587–615.

[37] John Lintner. "The aggregation of investor's diverse judgments and preferences in purely competitive security markets". In: *Journal of financial and quantitative analysis* 4.4 (1969), pp. 347–400.

[38] Jianan Liu, Robert F Stambaugh, and Yu Yuan. "Size and value in China". In: *Journal of financial economics* 134.1 (2019), pp. 48–69.

[39] Marcial Messmer. "Deep learning and the cross-section of expected returns". In: *Available at SSRN 3081555* (2017).

[40] Merton H Miller and Myron Scholes. "Rates of return in relation to risk: A reexamination of some recent findings". In: *Studies in the theory of capital markets* 23 (1972), pp. 47–48.

[41] Benjamin Moritz and Tom Zimmermann. "Tree-based conditional portfolio sorts: The relation between past and future stock returns". In: *Available at SSRN 2740751* (2016).

[42] Jan Mossin. "Equilibrium in a capital asset market". In: *Econometrica: Journal of the econometric society* (1966), pp. 768–783.

[43] David Rapach and Guofu Zhou. "Forecasting stock returns". In: *Handbook of economic forecasting*. Vol. 2. Elsevier, 2013, pp. 328–383.

[44] David E Rapach et al. "Industry interdependencies and cross-industry return predictability". In: (2015).

[45] Barr Rosenberg, Kenneth Reid, and Ronald Lanstein. "Persuasive evidence of market inefficiency". In: *The Journal of Portfolio Management* 11.3 (1985), pp. 9–16.

[46] William F Sharpe. "Capital asset prices: A theory of market equilibrium under conditions of risk". In: *The journal of finance* 19.3 (1964), pp. 425–442.

[47] Dennis Stattman. "Book values and stock returns". In: *The Chicago MBA: A journal of selected papers* 4.1 (1980), pp. 25–45.

[48] James R Thompson et al. "Nobels for nonsense". In: *Journal of Post Keynesian Economics* 29.1 (2006), pp. 3–18.

[49] Seha M Tinic and Richard R West. "Risk and return: Janaury vs. the rest of the year". In: *Journal of Financial Economics* 13.4 (1984), pp. 561–574.

Appendix A

Appendix A

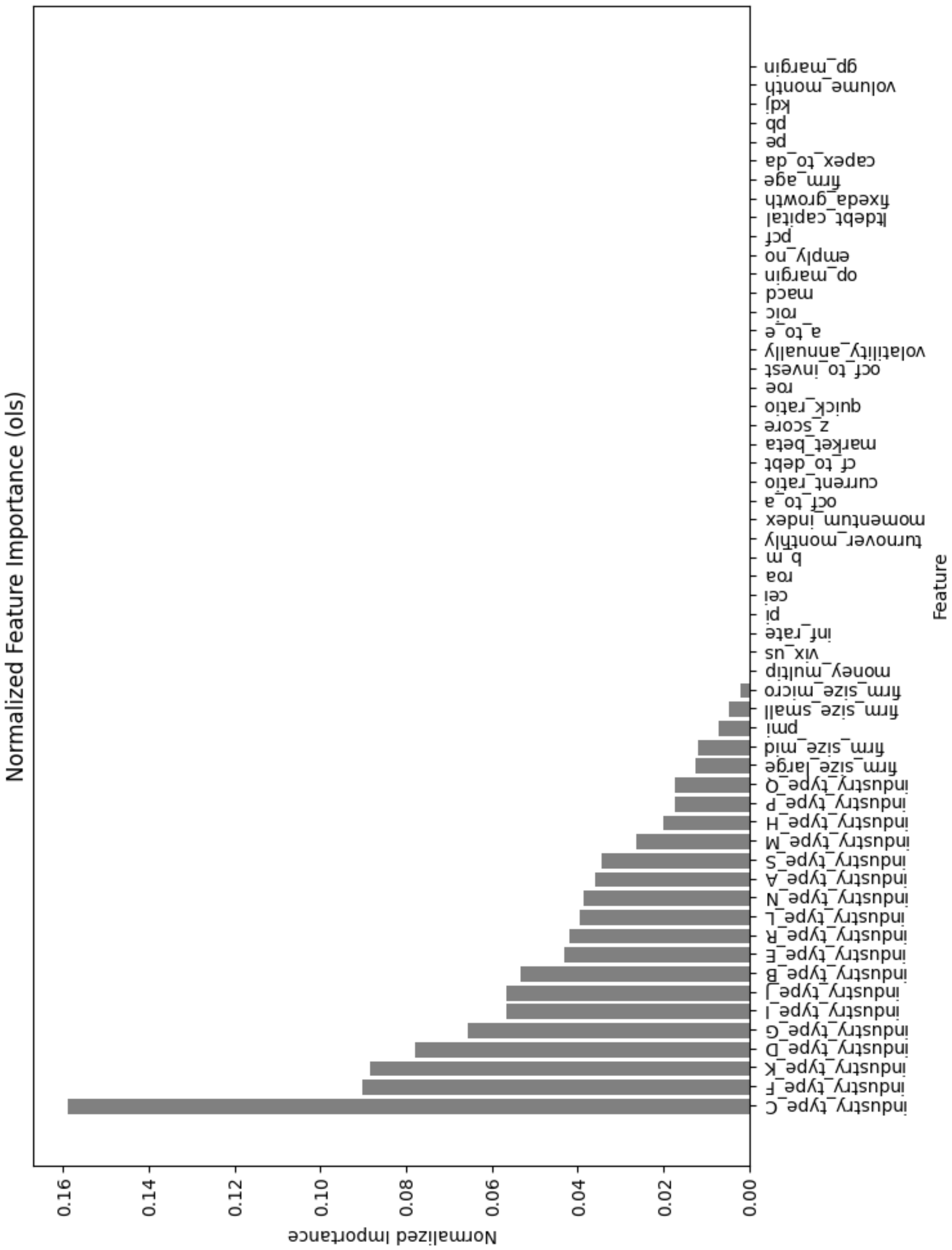


Figure A.1: Feature Importance (ordinary linear regression)

(a) In Figure A.1, we present the feature importance of all the features included in ordinary linear regression. Feature importance is used to evaluate the contribution of each feature to the predictive power of the model. By ranking features based on their importance, we can identify those with the most significant impact on model performance, gain a better understanding of the relationships between features and how they collectively influence the predicted outcomes.

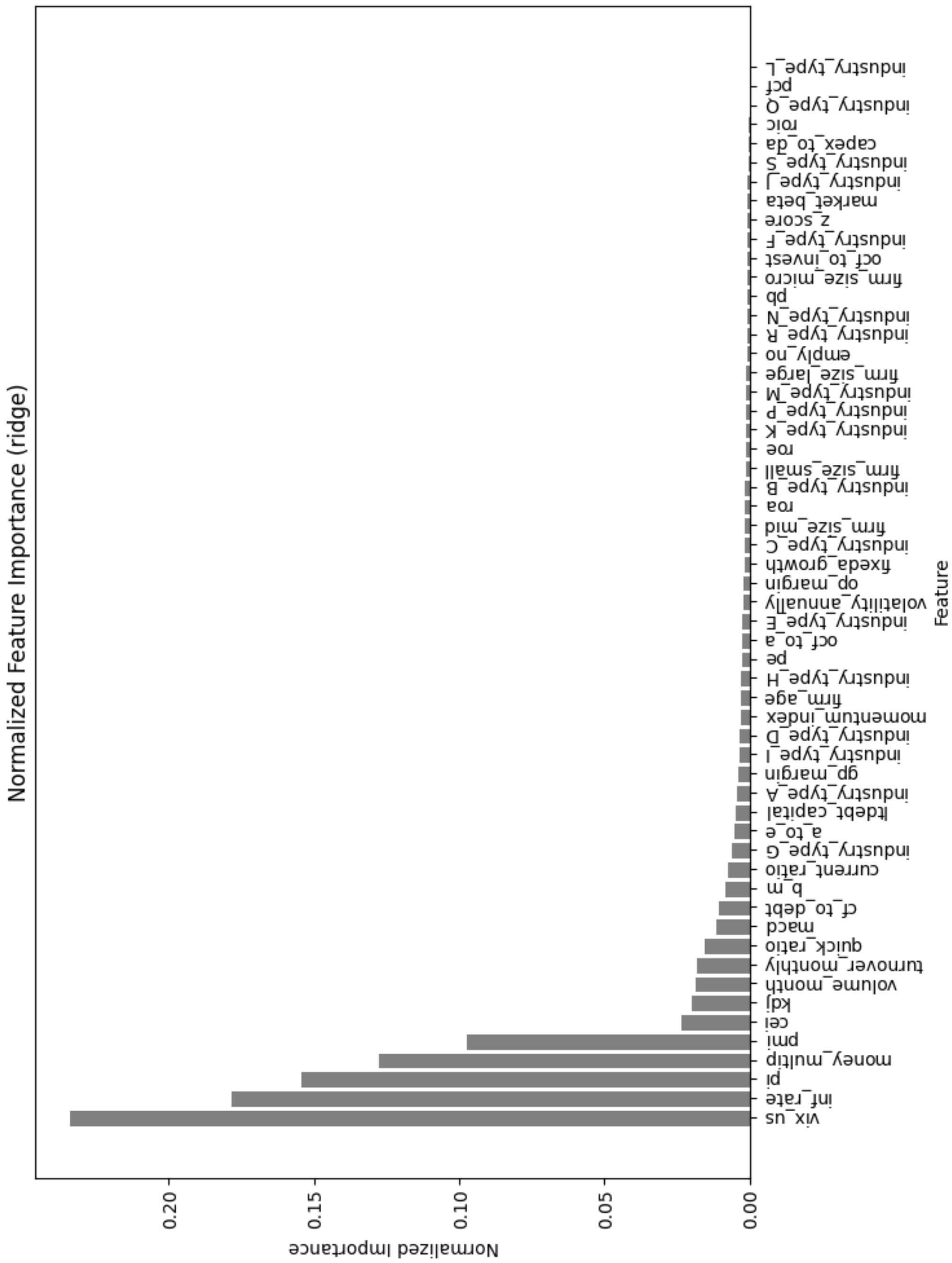


Figure A.2: Feature Importance (ridge)

(a) In Figure A.2, we present the feature importance of all the features included in ridge regression. Feature importance is used to evaluate the contribution of each feature to the predictive power of the model. By ranking features based on their importance, we can identify those with the most significant impact on model performance, and gain a better understanding of the relationships between features and how they collectively influence the predicted outcomes.

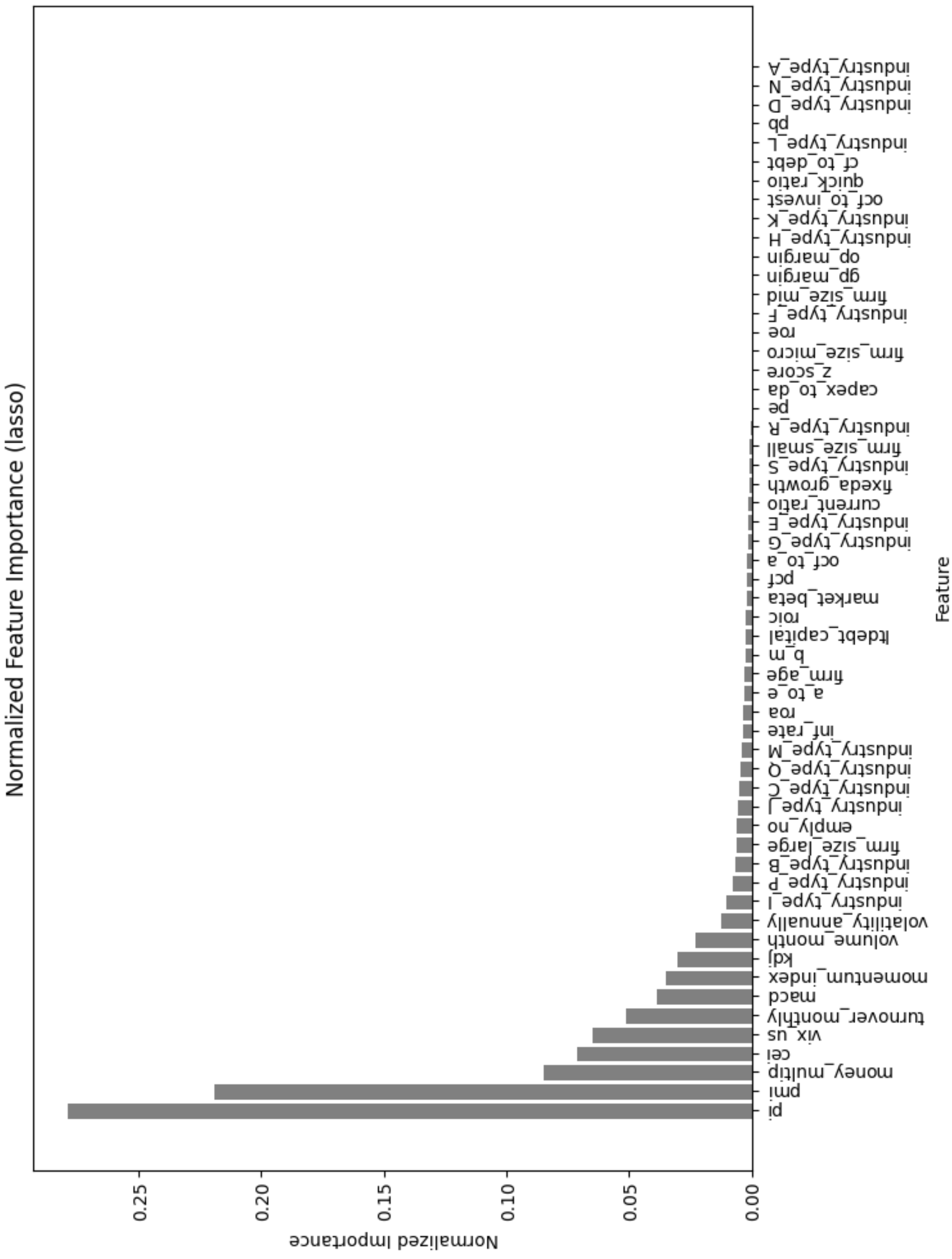


Figure A.3: Feature Importance (lasso)

(a) In Figure A.3, we present the feature importance of all the features included in LASSO. Feature importance is used to evaluate the contribution of each feature to the predictive power of the model. By ranking features based on their importance, we can identify those with the most significant impact on model performance, and gain a better understanding of the relationships between features and how they collectively influence the predicted outcomes.

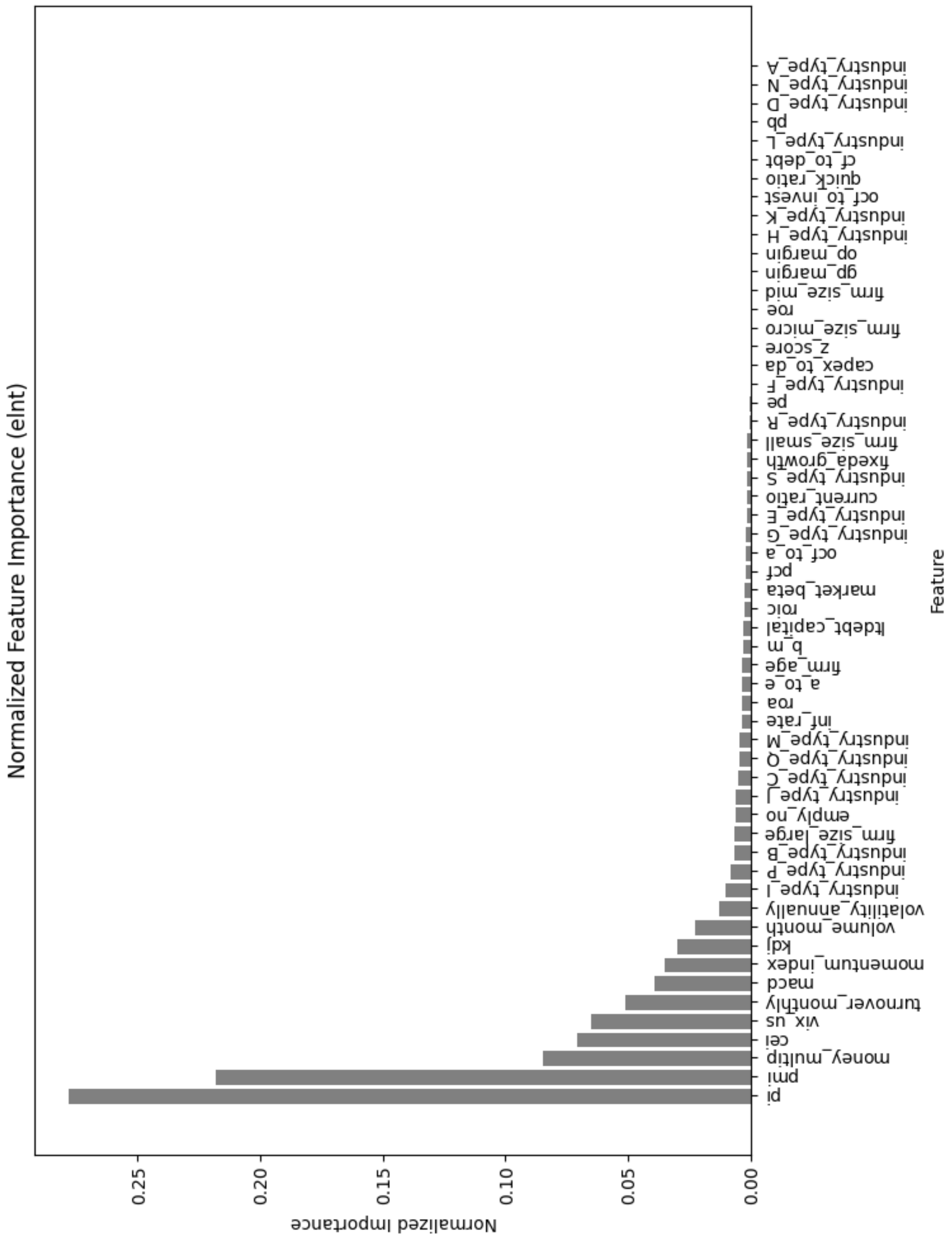


Figure A.4: Feature Importance (elastic net)

(a) In Figure A.4, we present the feature importance of all the features included in Elastic Net. Feature importance is used to evaluate the contribution of each feature to the predictive power of the model. By ranking features based on their importance, we can identify those with the most significant impact on model performance, and gain a better understanding of the relationships between features and how they collectively influence the predicted outcomes.

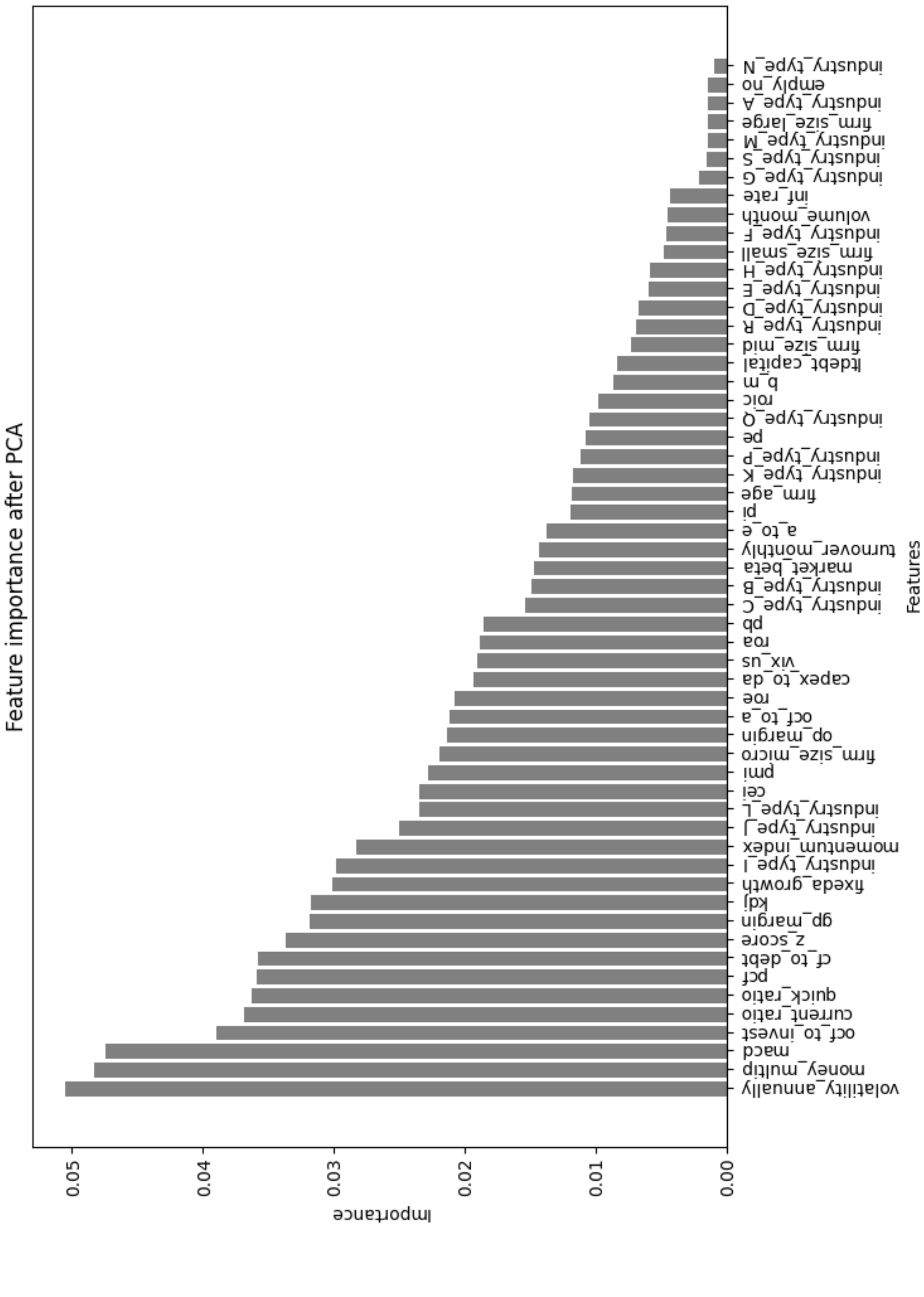


Figure A.5: Feature Importance (pcr)

(a) In the Figure A.5, we present the feature importance of all the features included in Pricinpal Components Regression. Feature importance is used to evaluate the contribution of each feature to the predictive power of the model. By ranking features based on their importance, we can identify those with the most significant impact on model performance, gain a better understanding of the relationships between features and how they collectively influence the predicted outcomes.

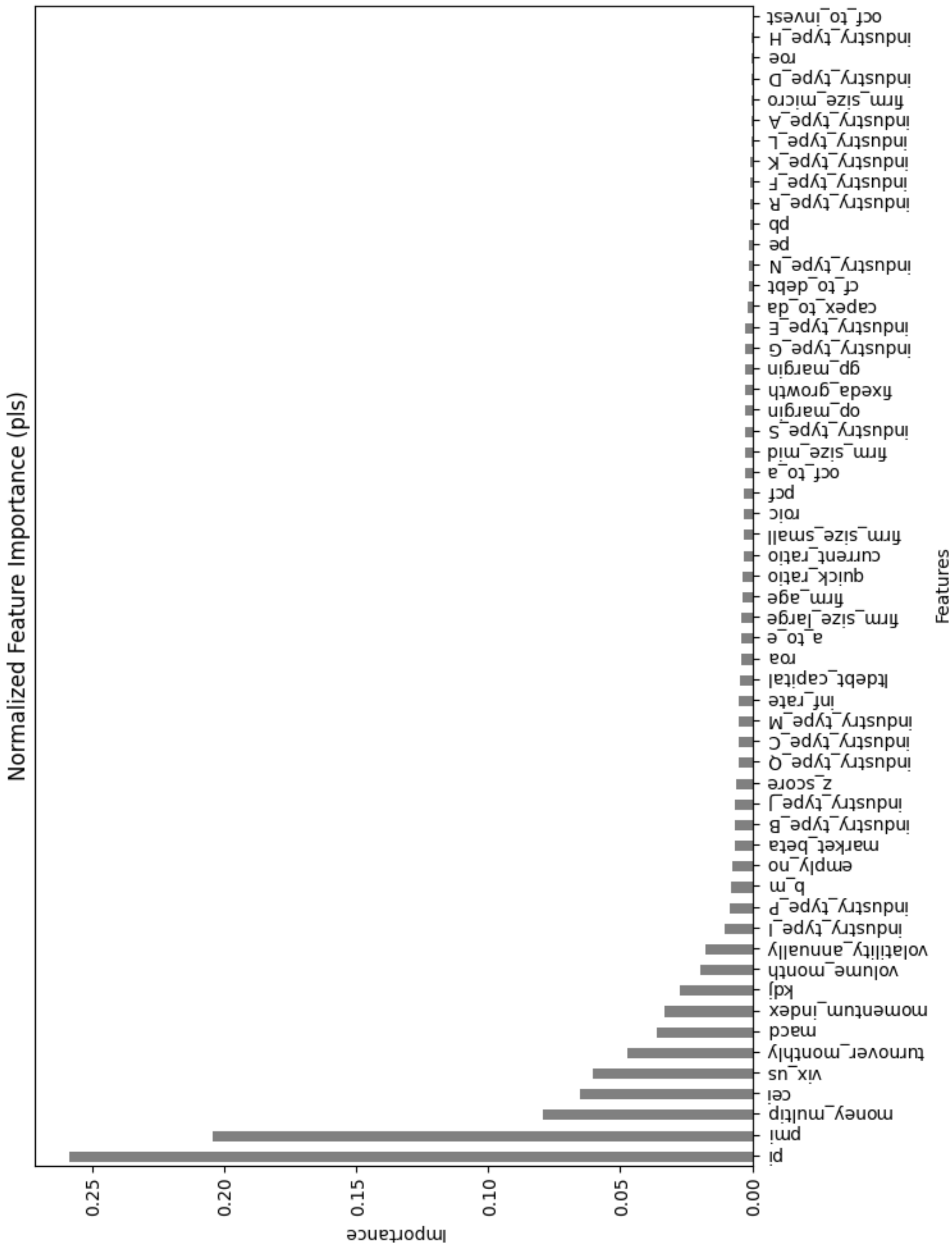


Figure A.6: Feature Importance (pls)

(a) In the Figure A.6, we present the feature importance of all the features included in Partial Least Squares. Feature importance is used to evaluate the contribution of each feature to the predictive power of the model. By ranking features based on their importance, we can identify those with the most significant impact on model performance, gain a better understanding of the relationships between features and how they collectively influence the predicted outcomes.

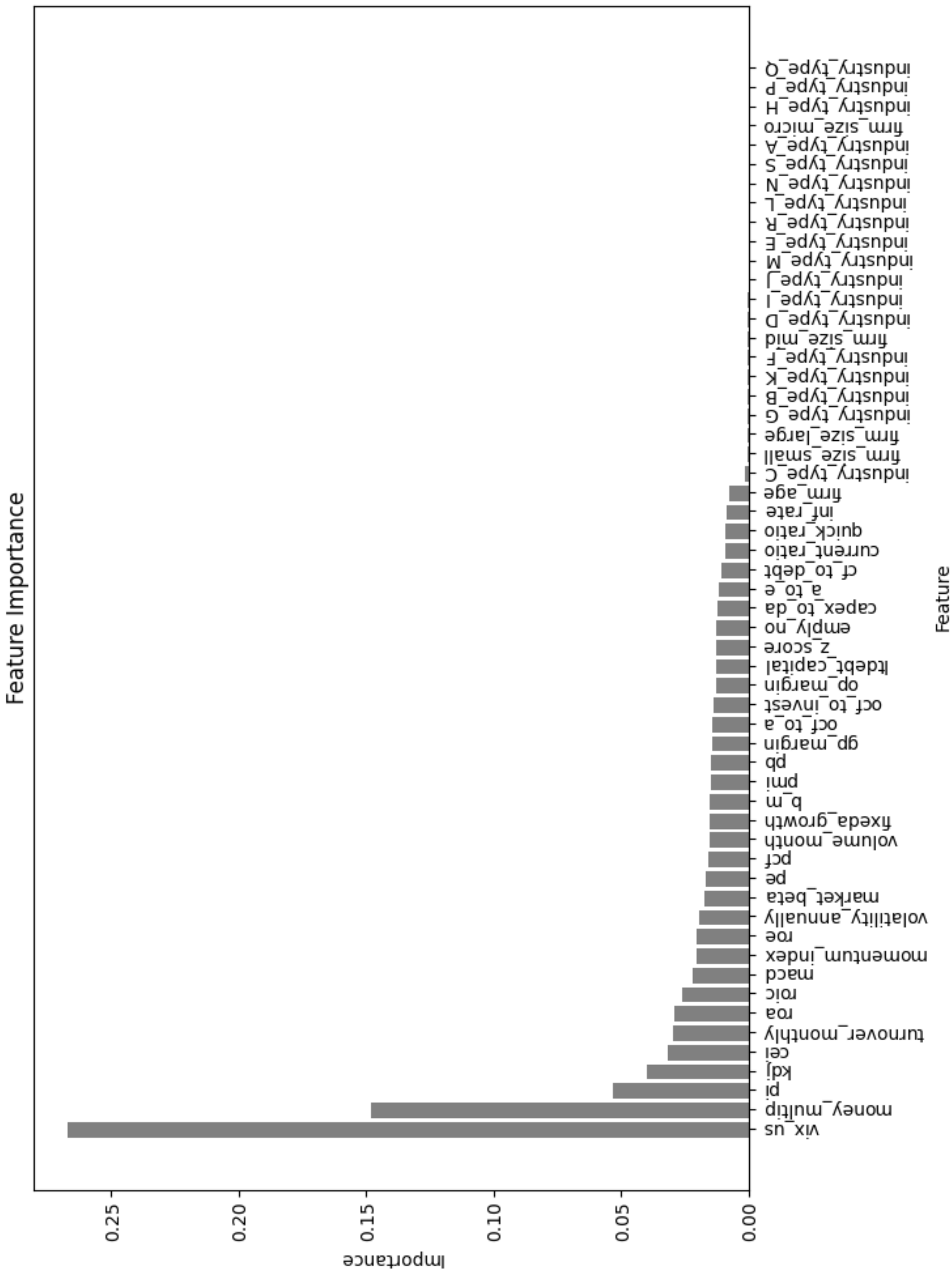


Figure A.7: Feature Importance (random forest)

(a) In the Figure A.7, we present the feature importance of all the features included in Random Forest. Feature importance is used to evaluate the contribution of each feature to the predictive power of the model. By ranking features based on their importance, we can identify those with the most significant impact on model performance, gain a better understanding of the relationships between features and how they collectively influence the predicted outcomes.

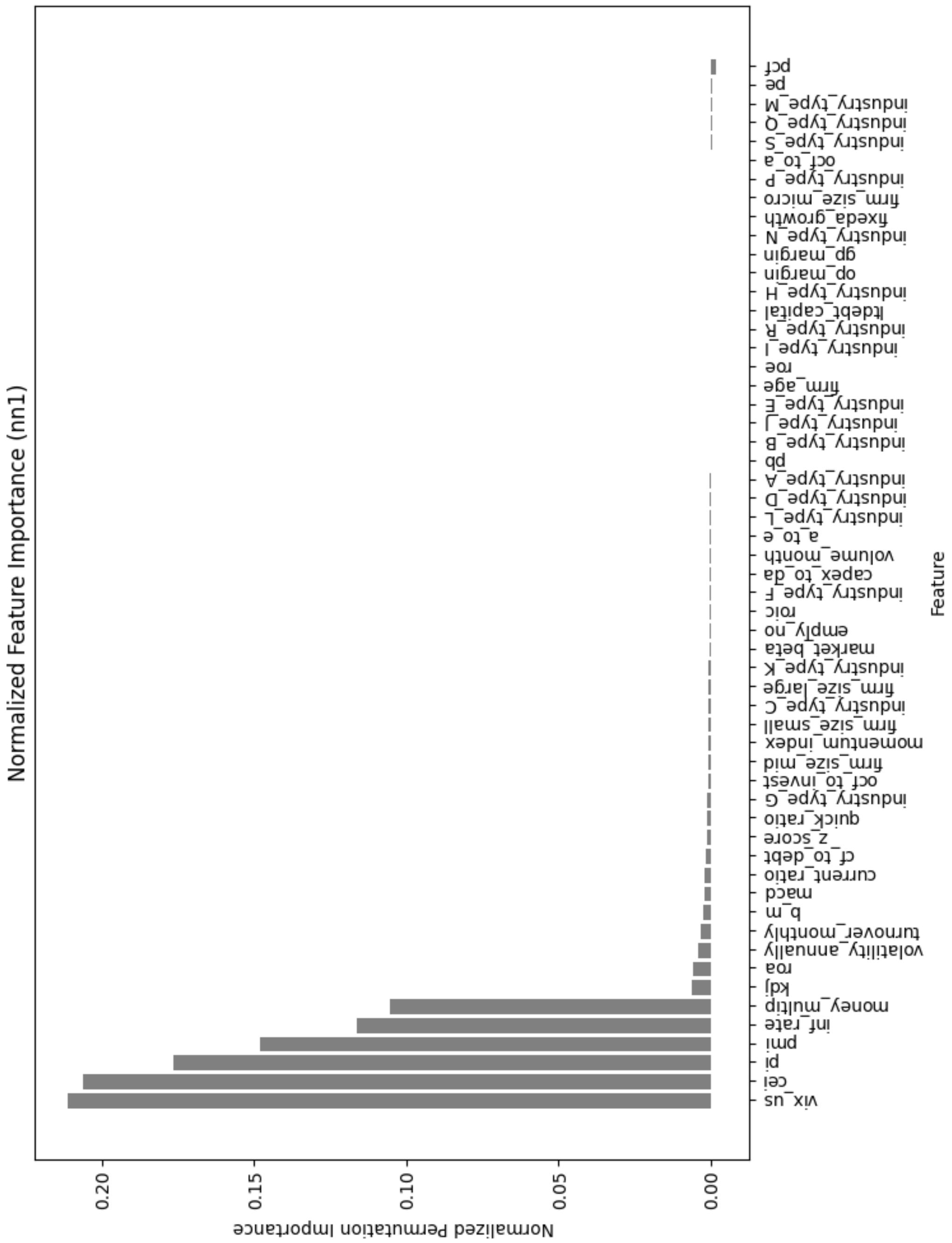


Figure A.8: Feature Importance (neural network with one hidden layer)

(a) In the Figure A.8, we present the feature importance of all the features included in Neural Network with one hidden layer. Feature importance is used to evaluate the contribution of each feature to the predictive power of the model. By ranking features based on their importance, we can identify those with the most significant impact on model performance, gain a better understanding of the relationships between features and how they collectively influence the predicted outcomes.

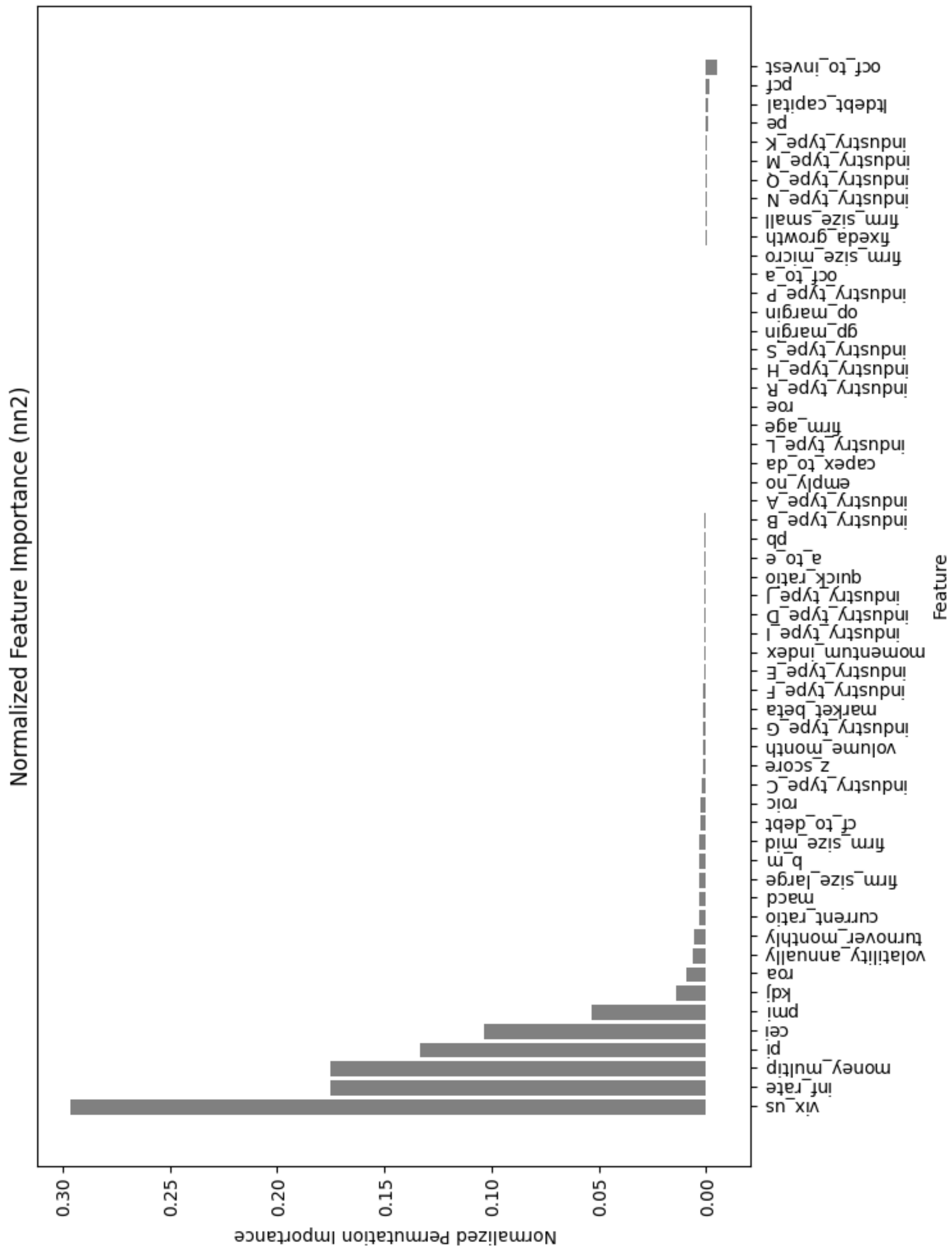


Figure A.9: Feature Importance (neural network with two hidden layer)

(a) In the Figure A.9, we present the feature importance of all the features included in Neural Network with two hidden layer. Feature importance is used to evaluate the contribution of each feature to the predictive power of the model. By ranking features based on their importance, we can identify those with the most significant impact on model performance, gain a better understanding of the relationships between features and how they collectively influence the predicted outcomes.

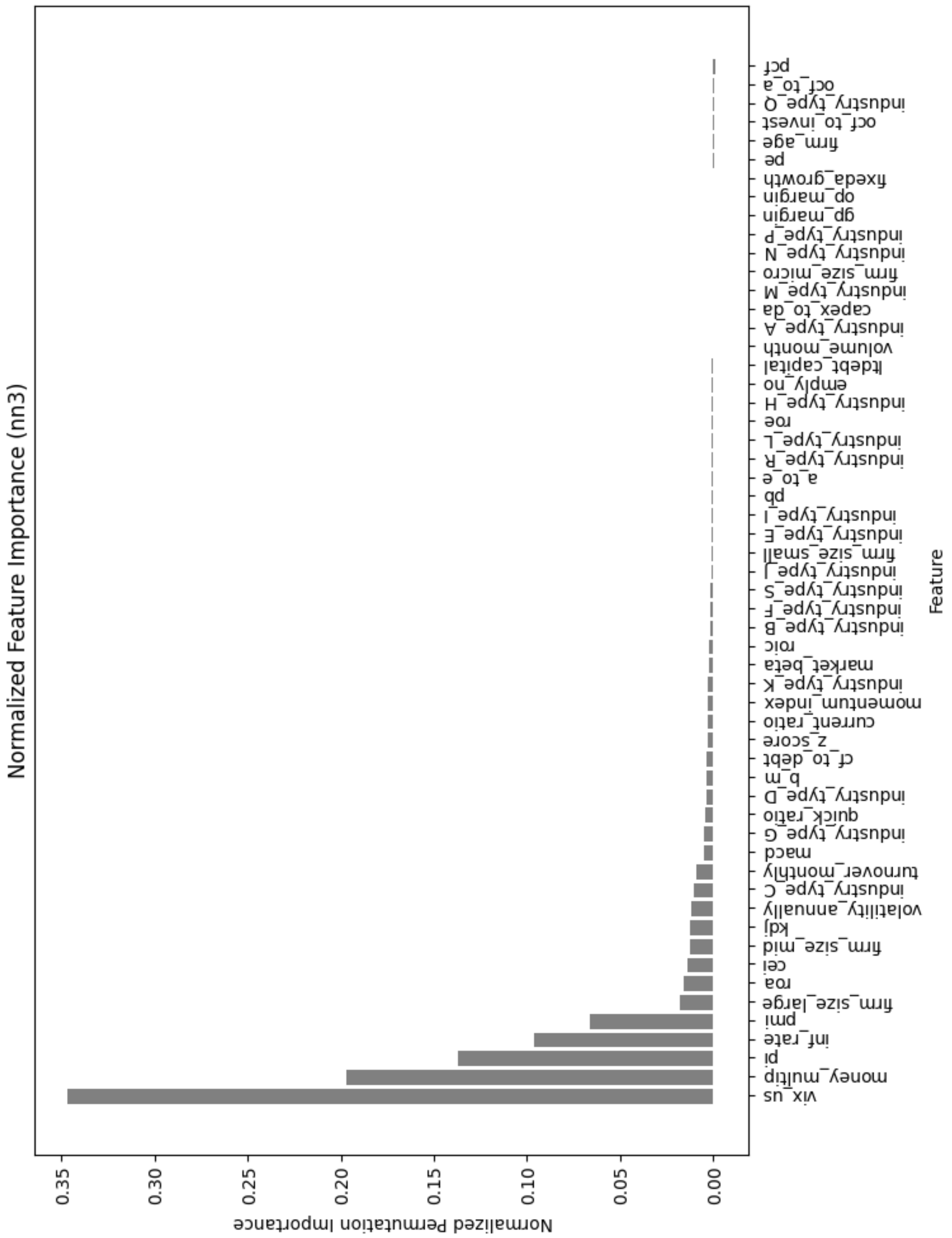


Figure A.10: Feature Importance (neural network with three hidden layer)

(a) In the Figure A.10, we present the feature importance of all the features included in Neural Network with three hidden layer. Feature importance is used to evaluate the contribution of each feature to the predictive power of the model. By ranking features based on their importance, we can identify those with the most significant impact on model performance, gain a better understanding of the relationships between features and how they collectively influence the predicted outcomes.

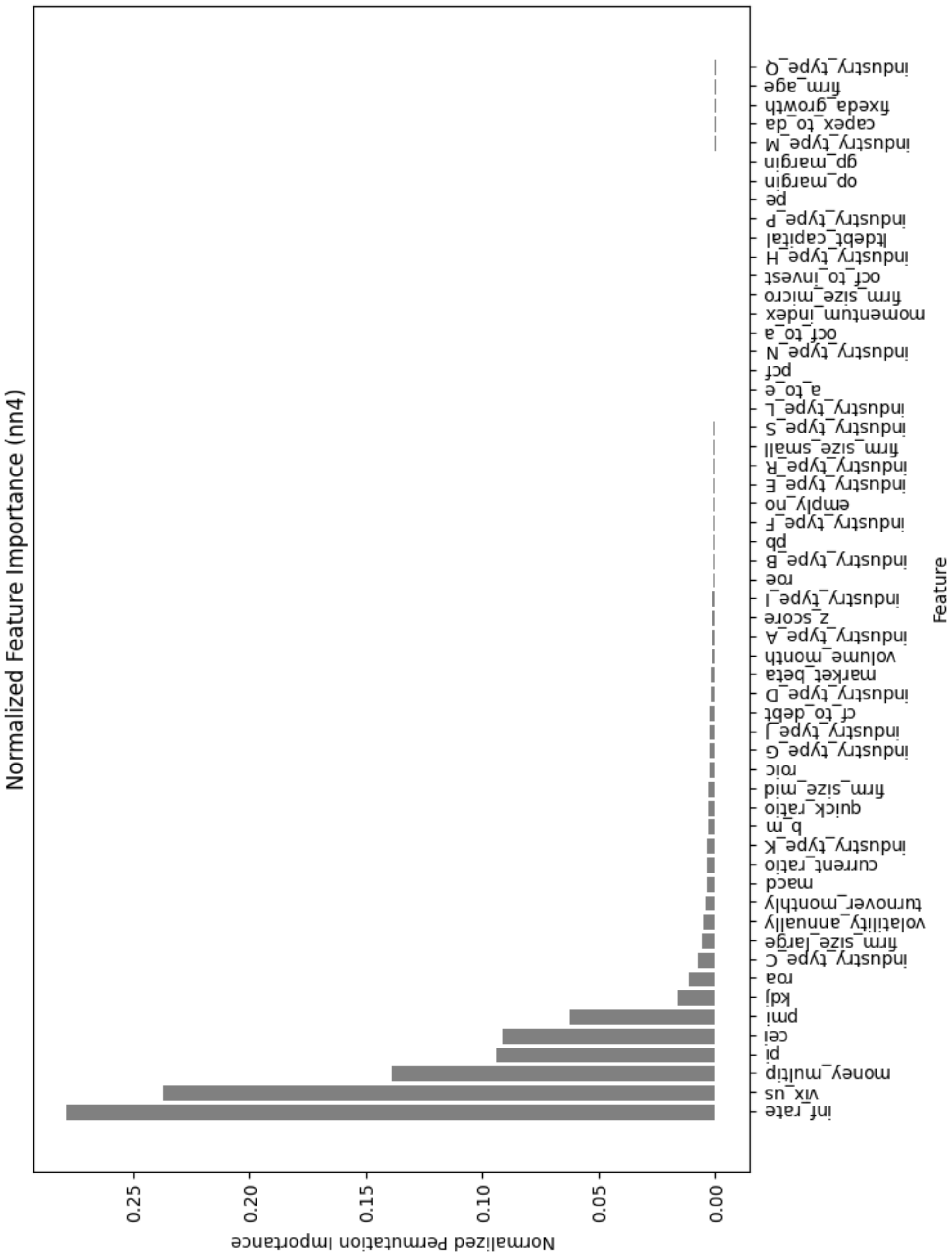


Figure A.11: Feature Importance (neural network with four hidden layer)

(a) In the Figure A.11, we present the feature importance of all the features included in Neural Network with four hidden layer. Feature importance is used to evaluate the contribution of each feature to the predictive power of the model. By ranking features based on their importance, we can identify those with the most significant impact on model performance, gain a better understanding of the relationships between features and how they collectively influence the predicted outcomes.

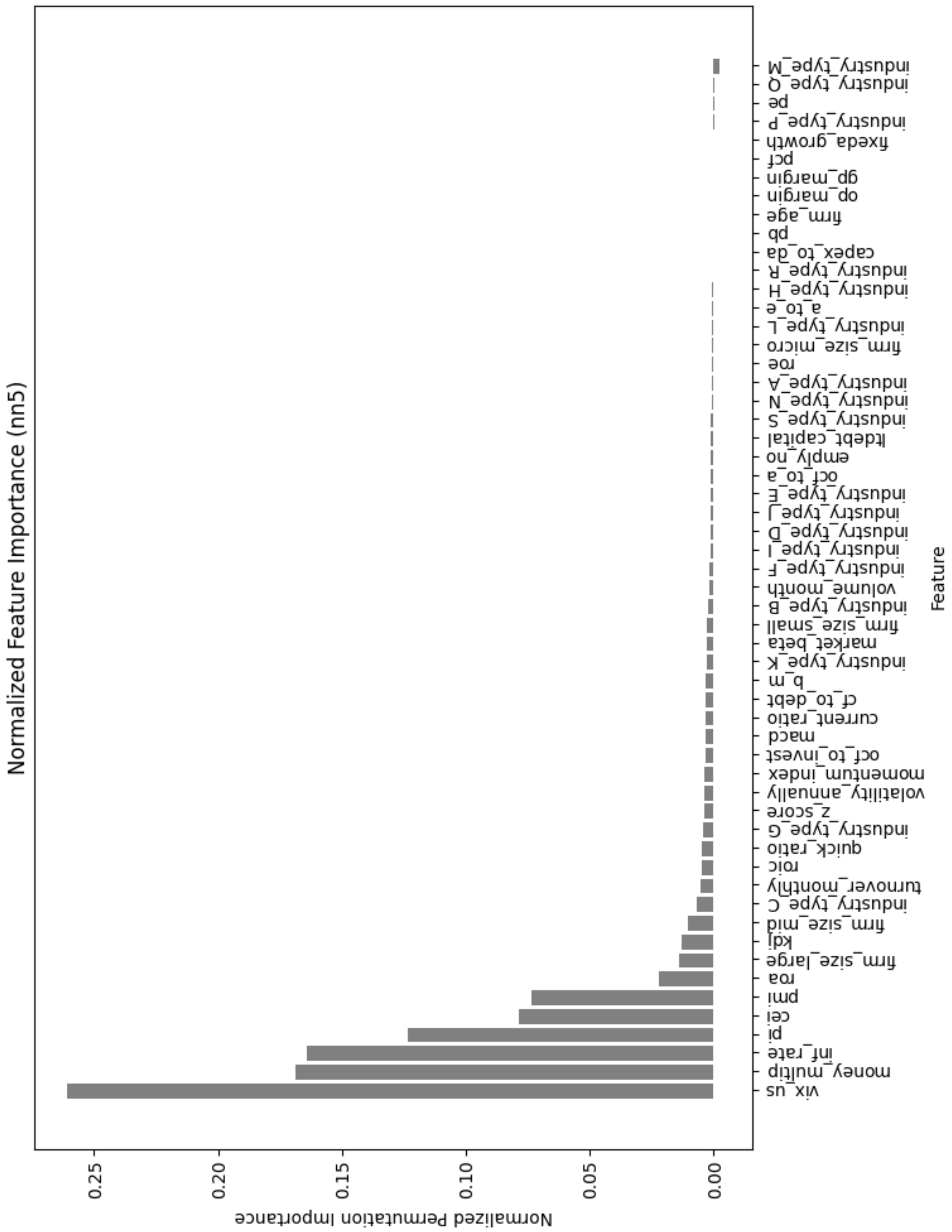


Figure A.12: Feature Importance (neural network with five hidden layer)

(a) In the Figure A.12, we present the feature importance of all the features included in Neural Network with five hidden layer. Feature importance is used to evaluate the contribution of each feature to the predictive power of the model. By ranking features based on their importance, we can identify those with the most significant impact on model performance, gain a better understanding of the relationships between features and how they collectively influence the predicted outcomes.