Exploring the International Application of Machine Learning in Asset Pricing: An Empirical Study

CHANG QIN

Exploring the International Application of Machine Learning in Asset Pricing: An Empirical Study

Chang Qin

ABSTRACT

This thesis delves into the application of machine learning models for predicting cross-sectional returns in diverse markets. Chapter One explores the predictive abilities of XG-Boost, Random Forest, and neural network models in relation to fund performance and fund manager information characteristics. The findings indicate that fund performance characteristics prove to be more informative of future fund performance than the characteristics of fund managers. Chapter Two probes the presence of bimodality in momentum stocks and examines the profitability of deep momentum, a machine learning return prediction model, in the UK, Japan, and South Korea. The findings demonstrate that bimodality is a phenomenon linked to developed markets and can cause losses for JT strategy investors. However, the deep momentum model generates substantial profits in all markets by relieving bimodality in long-short portfolios. Chapter Three investigates the efficacy of the momentum factor in Chinese stock markets. We compare the performance of the traditional linear JT model, the XG-Boost model, the neural network model, and neural network reclassification models as developed by Han (2022). The study finds that machine learning models based on the momentum factor outperform the traditional JT linear regression model, indicating a non-linear relationship between the momentum factor and stock returns in China. Han's reclassification models perform the most strongly after reclassification of the true target distribution within high-return deciles moves from a bimodal shape to a right-skewed distribution. The study also observes a significant positive correlation between the return of the long-only portfolio developed using the momentum factor in the machine learning framework and the size and sentiment index. Overall, this thesis attests to the practicality of machine learning models for predicting cross-sectional returns in various markets, with potentially gainful implications for investors and policymakers.

Supervisors: Chulwoo Han ,Guanming He and Yawen Zheng

Exploring the International Application of Machine Learning in Asset Pricing: An Empirical Study

Chang Qin

chang.qin@durham.ac.uk

Supervisor Team

Chulwoo Han

chulwoo.han@durham.ac.uk

Guanming He

guanming.he@durham.ac.uk

Yawen Zheng

yawen.zheng@durham.ac.uk

A thesis presented for the degree of Doctor of Philosophy

Business School

Durham University

Submission year: 2023

Contents

Statement of Copyright

Acknowledgements

I am deeply indebted to the many people who have helped make this doctoral journey possible and I wish to extend my heartfelt thanks to all of you.

First and foremost, I would like to express my heartfelt gratitude to my team of supervisors, Dr Chulwoo Han, Dr Guanming He, and Dr Yawen Zheng, for their invaluable guidance, support, and encouragement throughout this research. Dr Chulwoo Han's mentorship has been instrumental in introducing me to the field of machine learning and expanding my horizons. Dr Guanming's unwavering support and encouragement helped me to overcome the more challenging moments and regain my confidence in my abilities. Dr Yawen Zheng's critical feedback and constructive suggestions were instrumental in sharpening my thesis and bringing it to a stage of completion.

I am deeply grateful for the unwavering support and encouragement of my mother, who has always been a pillar of strength and inspiration throughout my life. Her unwavering commitment, love, and belief in me have given me the courage to pursue my dreams and goals. My mother's example of strength and resilience in the face of adversity has been a guiding light and I dedicate this thesis to her.

I would also like to express my sincere appreciation to all of my friends and family members who have supported and encouraged me throughout this journey. Your unflinching faith, love, and encouragement have sustained me through the ups and downs of doctoral life and I am deeply grateful for your presence in my life.

Further, I want to take a moment to acknowledge my own journey of overcoming depression and illness during my PhD study. It was an arduous road, but the constant support and encouragement I received from my supervisory team and loved ones were the pillars that kept me standing. I hope that my personal experience can inspire others to see the light through the darkness and persuade them that they should never give up on their dreams. I remain eternally grateful for their making this journey possible and for enriching my life in countless ways.

Chapter 1

Exploring the capability of machine learning models to learn from past fund performance and managerial characteristics

ABSTRACT

This study employs three different machine learning models to predict cross-sectional returns for fund selection, investigating whether fund manager information can predict fund performance in China. The three models used in this study are XG-Boost, Random Forest, and neural networks. The input features of the models are grouped into three categories, which include fund performance characteristics, fund manager information characteristics, and mixed characteristics. The models are each trained individually, utilising three distinct feature sets, resulting in the acquisition of nine sets of findings (3*3). The findings reveal that past fund performance is indicative of future fund performance, while the predictive ability of fund manager information is poor in both model analysis and finance in our out-of-sample empirical tests. A comparison of the two suggests that fund manager characteristics do not provide more valuable insights than fund performance characteristics and even generate noise that interferes with model predictions.

Keywords asset pricing · machine learning · empirical analysis

1.1 Introduction

As most investors choose mutual funds, they typically evaluate them based on two factors, namely the fund manager and his or her past performance. In several studies, scholars have investigated whether common factors can provide a reliable forecast for future fund performance (Carhart, 1997; Li et al., 2018). On the other hand, some scholars also actively research whether the fund manager's information can predict a fund's future performance (Golec, 1996; Chevalier and Ellison, 1999; Maxam et al., 2005; Fang and Wang, 2015). However, a consensus has not been reached on these issues among scholars. As a result, predicting the performance of funds continues to be a matter of significant and enthusiastic interest among scholars.

While relatively young compared to its US equivalent, the Chinese fund market has experienced notable progress and now plays a critical role as a forum for investment. Presently, the market claims over 6,000 mutual funds as of the end of 2021. Further, mutual funds hold significant sway, as institutional investors in China seek to enhance investment structures in the Chinese security market and foster greater public awareness of the value investment concept. Given the pace of the market's growth, investigations into funds in this emerging market retain considerable interest.

According to recent studies, mutual fund performance predictability in China displays conflicting evidence. This discrepancy can be attributed to the limited availability of data and short analysis periods. Additionally, prior studies have primarily employed linear models, indicative of a failure to capture the complexity of the data. Therefore, there is an evident need for methods that can handle extensive data and learn complex data input interactions. Machine learning is a promising option because it can operate under a unified framework and capture the predictive power of a broad range of input variables (Wu et al. (2021)). Further, it offers superiority over classical statistical techniques for prediction problems, as it does not require distribution or linearity assumptions and is more robust than traditional statistical models. Several studies have demonstrated the exceptional predictive power of machine learning models in asset pricing (Ludwig and Piovoso, 2005; Gu et al., 2020; Wu et al., 2021).

This paper aims to predict the future performance of funds by framing this as a supervised machine learning classification problem. To achieve this, we employ three machine learning models: neural networks, Random Forests, and XG-Boost. To facilitate effective fund selection, we develop up to three groups of features. These groups comprise the historical performance characteristics of funds, the characteristics of fund managers, and a combination of the two. Unlike most studies that focus on the US market, we investigate whether the past performance of funds in developing markets contains information that can predict future performance. Moreover, we explore the predictive power of fund manager information from a different perspective. Firstly, we examine whether fund manager information can predict fund performance in various market environments. Secondly, we investigate how the model would be influenced by the combination of fund manager characteristics with the fund's past performance characteristics. To validate our hypothesis,

we evaluate our models from two perspectives using out-of-sample data. First, we assessed the performance of the machine learning model. Secondly, we analysed the financial performance of the portfolio constructed from the model's predicted outcomes. By adopting this approach, we aim to gain a comprehensive understanding of the effectiveness of our model in providing accurate predictions and generating profitable outcomes.

The paper makes a significant contribution by utilising machine learning frameworks to analyse the impact of both fund characteristics and fund manager characteristics on fund selection. Moreover, our study on the Chinese fund market provides a valuable addition to the current empirical evidence in developing countries. Further, our research complements established empirical evidence in applying machine learning in financial asset pricing. Upon analysis of the predictive capabilities of the selected machine learning models, we have determined that fund historical performance characteristics are effective in predicting future fund performance. Our top-class portfolio, constructed using fund characteristics, surpasses both the fund market and equally weighted indexes. However, neither the model nor financial performance provides significant evidence supporting the predictive power of fund manager characteristics. Based on the accuracy of the machine learning model, the outcome derived from fund manager characteristics appears to be a random guess. Our importance test of the mixed feature group indicates that the fund manager's educational background and general information are not crucial factors in fund selection. Furthermore, following the inclusion of fund manager information, there is a decline in the performance of our tree-based model. We, therefore, conclude that fund manager information adds more noise than predictive value.

This paper is organised as follows: Section 2 provides a comprehensive review of the existing literature pertaining to the impact of fund and fund manager characteristics on fund performance, as well as machine learning applications. In section 3, we focus on discussing the contributions of this paper and present some empirical conclusions. A detailed explanation of the data set and input features is provided in Section 4. The methodology used in this study is presented in Section 5. In Section 6, we present the out-of-sample empirical results and analyse both the model performance and financial performance independently. Finally, we conclude our study in Section 7.

1.2 Literature review

The question of whether mutual fund performance can be predicted empirically has been a subject of significant attention in the extant literature. Early research documents evidence of persistence in mutual fund performance in the US market over short-term horizons of one to three years (Hendricks et al., 1993; Wermers, 1997), as well as over longer horizons of five to ten years (Elton et al., 1996). Carhart (1997) suggests that the common factors in stock returns and investment expenses can explain persistence in equity mutual funds' mean and risk-adjusted returns.

However, not all studies can reach a consensus on the persistence of fund performance. For example, Lakonishok et al. (1992) declare that they find little consistency between the performance of a given year and the subsequent year. Their analysis reveals that consistency is only evident for a three-year period. Collaboration with previous winners is shown to result in a benefit of approximately 110 basis points annually over the following three years. However, the authors caution against relying solely on past performance to predict future success. They document that the best-performing funds still record an annual decline in their performance of approximately 4.8%, while the worst-performing funds demonstrate an 8.4% improvement in performance. According to Berk and Green (2004), the lack of persistence in active fund performance is the result of investors making rational use of information in chasing performance.

By evaluating journal papers published between 1960 and 2015 containing the word "fund" in the title or "mutual fund" in the abstract, Jones and Mo (2021) strive to compile a complete sample of mutual fund predictors. They review publications from five general economics journals and six finance journals. Their findings suggest that the majority of predictors, initially employed to anticipate the performance of equities funds, are equally useful in predicting the alphas of corporate bond funds. The question of whether academic finance research is applicable in practice primarily depends on whether its conclusions hold outside of the sample. However, their out-of-sample results for the predictive factors used in the financial literature to estimate future mutual fund alphas fall short of expectations, with predictability decreasing by at least half in the post-sample analysis.

As feature engineering continues to develop, more scholars are expanding their analysis beyond a single factor. Prather et al. (2004) employ a generalised multi-factor model to explain abnormal fund returns. Their base model employs ordinary least squares regression (OLS) with the risk-adjusted performance measure as the dependent variable. To control for the impact of investment objective and year of observation, they include the year of observation ($Year_t$), investment objective, and an interaction variable for the year ($INVOBJ_t * YEAR_t$) and investment objective ($INVOBJ_t$) to serve as control variables. Meanwhile, their full model contains a persistence variable and 25 characteristic variables from the four categories, namely fund popularity (agility), growth (risk), operating costs, and managerial characteristics. They demonstrate that fund performance is positively correlated with price ratio factors and negatively correlated with market capitalisation, expense ratio, and the number of funds under management. Although they claim that multi-collinearity does not appear to be a severe issue, it still exists. As such, traditional regression methods are constrained when dealing with a large set of predictors.

Aside from statistical linear regression models, researchers have explored other model structures to address the issue at hand. The application of machine learning to predict mutual fund performance can be traced as far back as 1996. Chiang et al. (1996) use neural networks to develop a model for predicting the end-of-year net asset values of US mutual funds. They select a branch of economic variables as their input. They posit that neural networks exhibit superior performance to the traditional regression model due to the latter being constrained by the degrees of freedom. By using

neural network models, Ray and Vina (2004) attempt to determine the association between economic variables and the returns of mutual funds in the Indian context. They prefer using economic features as inputs, which include the 91-day Treasury Bill Rate, wholesale Price Index, Money Stock, as represented by M3, and the Index of Industrial Production, etc. Indro et al. (1999) use an artificial neural networks (ANN) approach to predict the performance of equity mutual funds. Their inputs focus on operating characteristics such as turnover, price-earnings ratio, price-book ratio, and so forth. According to their findings, the predictive capabilities of neural networks (NNs) are superior to those of linear models in forecasting fund performance across all styles. Due to the early publication of these two papers (Chiang et al., 1996; Indro et al., 1999), the selected samples only contain 101 and 438 funds, respectively. It is noteworthy that machine learning algorithms necessitate a substantial quantity of data, and it should be taken into account that inadequate data may lead to imprecision.

In a recent study, Li and Rossi (2020) initiated their study by utilising mutual fund holdings, which are then integrated with a substantial number of stock characteristics (totalling 94) with the purpose of constructing mutual fund-level characteristics based on the stocks held within the fund. They transfer the target into a regression problem while applying boosted regression trees (BRTs). They suggest that machine learning models significantly outperform standard linear frameworks, especially BRTs. Based on the evidence, they suggest that the relationship between fund performance and fund characteristics is non-linear and that there are noteworthy degrees of interplay between various fund characteristics and fund performance. In addition, they suggest that the predictive importance of fund characteristics for fund performance, as well as the relationship between fund characteristics and fund performance, are subject to temporal changes. Using machine learning, Kaniel et al. (2022) demonstrates that fund features can reliably distinguish high-performing from low-performing mutual funds. They split the fund-specific and stock-specific characteristics into nine groups (specifically past returns, investment, profitability, intangibles, value, trading frictions, fund momentum, fund characteristics, and fund family characteristics) which cover 59 features. They validate and refine the study of Li and Rossi (2020) by demonstrating that holding-based stock characteristics can only forecast the systematic portion of fund returns. DeMiguel et al. (2021) combine machine learning and fund characteristics to design long-only equity fund portfolios that yield positive and significant out-of-sample alpha net costs. They suggest that the outstanding performance of these portfolios can be attributed to the model's accommodation of non-linearity and interaction between attributes and fund performance. They use three machine learning models (elastic net, Random Forests, and gradient boosting) and the share-class characteristics to capture the positive alpha of the mutual funds. One of the distinguishing characteristics of the study conducted by DeMiguel et al. (2021) is their focus on long-only mutual fund portfolios, whereas the study by Kaniel et al. (2022) emphasises the short leg of their long-short fund portfolios as the primary source of predictability in after-fee fund performance (see Figure 5b). DeMiguel et al. (2021) build their dataset of 17 share-class characteristics using readily available fund characteristics and information based on historical

returns. In comparison to prior research, our study examines the application of machine learning models to predict fund performance, while considering not only the fund's characteristics, but also those of the fund manager's characteristics.

In addition to the fund's characteristics, the fund manager's role is also indispensable in the performance forecast of active funds. Although the active managers seem to concur that it is difficult to capture a positive alpha after taking out the fee, some of the literature indicates that there exists a subset of managers who have the ability to outperform the benchmark (Fama and French, 2010; Kacperczyk et al., 2014; Berk and Van Binsbergen, 2015). There is a dearth of literature that delves into anticipating the performance of funds by analysing the characteristics of fund managers. This is in stark contrast to the ample research available on the forecast of fund performance based on the fund's characteristics. One of the early papers that shifted its focus to the manager rather than the fund is the study conducted by Golec (1996). The author studies a set of 530 funds and documents that fund managers who hold an MBA degree perform better than those who do not.

When examining the traits of fund managers, their educational backgrounds are often a topic of discussion. One commonly held belief is rooted in the principle of human capital, which posits that fund managers with better education possess greater human capital. This, in turn, implies that they should deliver better results and be compensated accordingly. The educational background of fund managers can be analysed from various angles, including their degree level and certifications, *inter alia*. In their study, Chevalier and Ellison (1999) expand upon fund manager characteristics by including factors such as the SAT score of the manager's undergraduate alma mater, their age, their tenure with the fund, and their MBA status. They challenge the findings of Golec (1996) by claiming that fund managers with MBAs do not outperform managers without MBAs after adjusting for systematic risk. Gottesman and Morey (2006) dispute the assertion of Chevalier and Ellison (1999) that managers with an MBA degree do not perform differently than managers without an MBA degree, given that Chevalier and Ellison's research only examines the period during which the market was relatively bullish (1988–1994). Additionally, Gottesman and Morey (2006) find no evidence that a manager holding a PhD has a significant impact on mutual fund performance. In contrast, research conducted on the Chinese fund market by Li et al. (2018) reveals that the educational qualifications of fund managers play a crucial role in determining fund performance. Their results indicate a noteworthy positive effect on fund performance when the fund manager possesses a master's or doctoral degree, an MBA degree, or overseas experience.

In addition, the significance of high-quality educational institutions is highlighted. Such institutions are able to offer ample resources and better support for students in building social networks with greater ease and efficacy. The research conducted by Chevalier and Ellison (1999) suggests that fund managers who received their undergraduate degrees from colleges with higher SAT scores tend to achieve greater risk-adjusted excess returns. This finding is supported by Li et al. (2011), who also find evidence of a positive relationship between SAT scores and hedge fund performance.

Meanwhile, Gottesman and Morey (2006) expand upon the existing literature on MBA program quality by investigating the relationship between fund manager performance and the GMAT score requirement of the MBA programme. Their study suggests that fund managers who completed MBA programmes with higher GMAT score requirements exhibit superior performance relative to those who do not hold an MBA degree or else complete MBA programmes with lower GMAT score requirements. The authors note that the average GMAT score of the MBA programme is positively and significantly correlated to the fund's performance. Additionally, fund managers who obtain their MBA from a business school ranked within the top 30 of the Business Week MBA programme rankings tend to experience better performance. Maxam et al. (2005) also support the notion that managers who graduate from elite American universities outperform their peers. Furthermore, they suggest that the fund manager's major does not impact the fund's performance.

With the expansion and specialisation of the industry, the proportion of fund managers who obtain various professional certifications has risen progressively. Gottesman and Morey (2006) argue that a fund manager's possession of a Chartered Financial Analyst (CFA) certification bears no impact on the performance of the mutual funds under their management. Andreu and Puetz (2017) conduct a comparison between the investment risk and style of fund managers holding a CFA certification along with an MBA degree against those who possess only one of these credentials. They demonstrate that managers holding both degrees tend to take fewer risks, adopt less extreme investment strategies, and achieve relatively moderate levels of performance. Fang and Wang (2015) conduct cross-sectional regressions to analyse the relationship arising between the fund return and several lagged fund manager characteristics in the context of the Chinese market. They claim managers with an MBA or a CFA qualification demonstrate better performance after comprehensive analysis. However, Li et al. (2018) do not find supportive evidence in China about the relationship arising between the CFA or Certified Public Accountant (CPA) qualification and fund performance by applying the data envelopment analysis (DEA) models and threshold panel models. Different data samples may lead to such conflicting results. Previous scholars have not yet reached a consensus as to whether fund performance is related to fund managers holding professional credentials.

Furthermore, a few academics investigate the personal characteristics of fund managers, including the managers' gender and age. Compared with other industries, the fund industry has low gender inclusivity (Madison Sargis (2016)). According to a 2016 report by MorningStar, Singapore boasts the highest percentage of female hedge fund managers globally, standing at a mere 30%. As per the data obtained from the China Stock Market and Accounting Research Database (CSMAR database, 2022), only 23% of the fund managers in mainland China are female. In earlier literature, Estes and Hosseini (1988) document a gender gap on Wall Street. They suggest that women are less confident about investment decisions. Powell and Ansic (1997) suggest that females are less risk-seeking than males. More recent literature also documents supportive evidence that females are more risk-averse and less overconfident than males (Beckmann and Menkhoff, 2008; Charness and Gneezy, 2012; Bernasek and Shwiff, 2001; Andreu et al., 2019).

Ahmadpour and Frömmel (2022) present divergent views when analysing hedge fund and commodity trading advisor (CTA) managers. They suggest that gender has only a minor impact on fund managers' risk-shifting behaviour, as they do not discover any conclusive evidence of risk-averse female managers. According to Bliss and Potter (2002), female managers of both U.S. and overseas equities mutual funds have a tendency to generate higher raw returns than their male counterparts. Clare (2017) shows that female fund managers generate significantly lower benchmark-adjusted returns than male fund managers. However, according to the research conducted by Atkinson et al. (2003) and Babalos et al. (2015), there is no significant difference arising in the performance of male and female fund managers in the U.S. market. Fang and Wang (2015) also claim that gender does not significantly affect the excess return of funds in the Chinese fund market. While it is recognised that there may be discrepancies in investment styles and risk-taking behaviours arising between genders, there remains a lack of consensus among previous studies regarding the impact of gender on fund returns.

1.3 Contributions and a preview of empirical findings

Our paper contributes to several streams of the financial literature. Firstly, it distinguishes itself from the vast body of existing literature based on the U.S. fund market, as our data is derived from China's fund market. As one of the largest and fastest-growing markets in the world, China's fund market is desirable to researchers. Over 16 years, the number of mutual funds in China has skyrocketed from 46 in 2001 to 4,395 in 2017. Furthermore, the total net assets of these funds have seen a remarkable increase of 58.9 billion yuan, reaching an impressive 12.9 trillion yuan(Amstad et al., 2020). There are significant disparities between the Chinese and U.S. fund markets, making the study of the Chinese fund market highly appealing. Just as with other developments in China's financial industry, the birth of mutual funds has been both spurred and facilitated by the government's deregulation. Thus, the Chinese government exerts more influence over the mutual fund industry. Additionally, the Chinese financial market imposes restrictions on short positions. These impact not only the construction of fund strategies but also the fund's performance. Consequently, we concur with DeMiguel et al. (2021) that our research will exclusively focus on the performance of the long-side position. Doing so makes our research on China's fund market more applicable and relevant.

Moreover, it is important to note that the history of mutual funds in China is significantly shorter than that of the U.S., resulting in limited data availability for research purposes. Since linear regression has strict requirements for data, previous studies on China's fund market have relied on a limited sample size of funds and a specific time period. For example, Su et al. (2012) analyse Chinese mutual funds and test performance persistence with only a tiny sample of 42 funds over the period from 2002 to 2009. Zhao and Wang (2007) collate data from only three years (2003-2005) – an incredibly short time window. In our study, we extensively analyse the complete dataset of stock funds and hybrid funds highly invested in stocks in China up to 2022. Our research approach involves the utilisation of machine learning

algorithms to construct models. Machine learning models have exhibited superior performance as compared to linear regression models, particularly in those situations where the data demonstrates non-linearity, contains noise, or else exhibits a high-dimensional nature. As a result, there is considerable interest in conducting research on the Chinese fund market, employing machine learning algorithms for model development.

In addition, the characteristics of Chinese fund managers also diverge from those of American fund managers. Fund managers' educational backgrounds vary extensively due to differences arising in educational policies between countries. Our statistics suggest that the percentage of Chinese fund managers holding MBA degrees is much lower as compared to their American counterparts. Hence, we design a local feature set for Chinese fund managers in our study.

Our paper adds to the literature on mutual fund performance attribution in the Chinese market. We conduct a comparative analysis by separately studying three groups of characteristics, namely fund characteristics, fund manager characteristics, and mixed characteristics. In our study, we find supportive evidence that past fund characteristics can provide predictive information for machine learning models. Using the model to select the top 20% of funds each month can achieve a maximum accuracy of 24%, which is a significant improvement when compared to the accuracy rate of randomly assigning five groups. While our study includes 14 fund manager characteristics relating to education background, general information, and team information, we are unable to find substantial evidence supporting the notion that these characteristics can effectively predict fund performance. In addition, our research demonstrates that characteristics obtained from the past performance of funds have a more significant influence when compared to characteristics pertaining to the fund manager. This finding is confirmed by the results of our importance score tests which are based on a combination of different feature groups. This finding suggests that there may be hidden or latent management characteristics that are otherwise unaccounted for in the selected characteristics of fund managers, yet are reflected in their past performance. Alternatively, the results may indicate that fund characteristics include factors that are unrelated to the manager's ability to govern the fund.

Furthermore, we provide supporting evidence that the predictive information contained in the selected features decays over time in China. When using the same features to make predictions, the model predicts the following month's performance more accurately than for the next six months. The paper also highlights the importance of a fund's market rank in predicting its future performance. Furthermore, our model results imply that the best-performing and worst-performing funds in the Chinese market share similar fund characteristics. These similar characteristics mislead our machine learning models, causing our models to label some funds belonging to the highest return group as the lowest return group funds and vice versa. The presence of such a bimodal distribution is undesirable, as it indicates a likelihood for some of the funds purchased based on high return predictions to produce low returns in the following month. This result is similar to previous findings from Han (2022) in the U.S. stock market.

Secondly, our paper also diverges from previous studies in the construction of the model. Comparing previous machine learning regression applications (Li and Rossi, 2020; Wu et al., 2021) in fund performance prediction studies, we use supervised machine learning models while transferring the problem into a classification problem. Ludwig and Piovoso (2005) apply NNs, decision trees, and Naïve Bayes to select money managers. They uncover that all three machine learning methods have performed well on financial classification problems. We investigate three distinct machine learning models: deep neural networks (DNNs), Random Forests, and XG-Boost. Besides, our input features differ from previous studies in that we include both fund manager and fund characteristics. Although we try to expand the data range as far as possible, the existing data for the Chinese market is still far smaller than for the U.S. market. Thus, we do not accept a large number of input features. In our study, in addition to using the importance score to display the difference among features, we also train and compare different feature groups separately. We venture further to uncover the black box in machine learning models.

We add to the literature on the empirical application of machine learning in finance. Insofar as we know, we are the first researchers to conduct a detailed analysis of fund and fund manager characteristics under the framework of machine learning models. We document that the fund characteristics convey more information than the fund manager characteristics under nonlinear models. Although performance can be improved by combining all characteristics, the improvement is not considerable. Moreover, we find that market-wide performance-related features play a more significant influence than the fund's own past returns. This implies a difference in feature engineering between machine learning algorithms and statistical modelling. Future scholars who wish to investigate the application of machine learning models in the financial sector will find our proposal about feature construction to be both novel and instructive.

1.4 Data and feature selection

1.4.1 Data collection

This study employs data obtained from the China Stock Market and Accounting Research (CSMAR) from July 2006 to December 2022. As a comprehensive research-oriented database, CSMAR collects data on dead funds while they are still alive, meaning that our data is free of any survivorship bias. The complete data range begins in 2006, immediately after the reform of non-tradeable shares on the Chinese stock market in 2004. Only a limited number of stocks are listed in the market initially, while the mutual fund market is also less developed (Chen et al., 2015; Jiang et al., 2018). Rao et al. (2018) investigate the ten-year post-reform performance of the equity funds in China (2004-2014) and claim that the past performance of equity funds is not predictive of future fund performance. However, a large set of data is needed for machine learning models. Therefore, unlike previous research, our chosen research period spans almost the entire database.

Our study specifically focuses on actively managed, open-ended mutual funds. To ensure consistency in our dataset, we have excluded several fund types that may significantly impact our results, such as exchange-traded funds (ETF), listed open-ended funds (LOF), qualified domestic institutional investor funds (QDII), umbrella funds, innovative funds, and index funds. To align with our fund selection, we chose the Shanghai Securities Open-end Fund Index as our benchmark.

As per the "Securities Investment Fund Operation and Management Measures" in China, funds are categorised into four groups, namely equity funds, bond funds, money market funds, and hybrid funds. Equity funds refer to those funds that invest more than 80% of their assets in stocks; bond funds invest more than 80% of their holdings in bonds; money market funds strictly invest in money market instruments, whereas hybrid funds invest in multiple asset classes and thus do not meet the aforementioned classification criteria. The current study focuses primarily on equity and hybrid funds, as these categories comprise more actively managed funds. Such funds present greater opportunities for fund managers to showcase their skills. For consistency in data, only hybrid funds with above 50% equity holdings are considered.

The fund return is one of the essential features of this study. In contrast to stocks, mutual funds disclose their performance via monthly reports of their net asset value (NAV) rather than directly reporting returns. The NAV represents the net value of a given mutual fund's assets after deducting its liabilities, and then dividing by the total number of outstanding shares. The monthly growth rate of NAV with dividends and the split is calculated as follows:

$$R_{monthly} = (NAV_t / NAV_{t-1}) - 1 \tag{1.1}$$

where:

$R_{monthly}$: Growth Rate of NAV with dividends and split
NAV_t: Accumulative unit NAV at the end of the current month
NAV_{t-1}: Accumulative unit NAV at the end of the previous month

As part of our data-cleaning process, we implemented measures to minimise the effects of outliers and errors in the database. Specifically, we excluded those funds with monthly returns exceeding 30% in absolute terms and retained only those with historical records spanning at least 18 months. Our analysis encompasses a sample of 2,263 distinct mutual funds from 2006 to 2022.

In 2013, China promulgated the "Securities Investment Fund Law" to regulate mutual funds (Amstad et al. (2020)). The introduction of this law put the mutual fund business in China on the path to rapid expansion. Figure 1.1 shows the

change in the number of funds that meet our filter criteria. From our analysis, we observe that the number of funds that fulfil the requirements of our filter drastically increases over time.

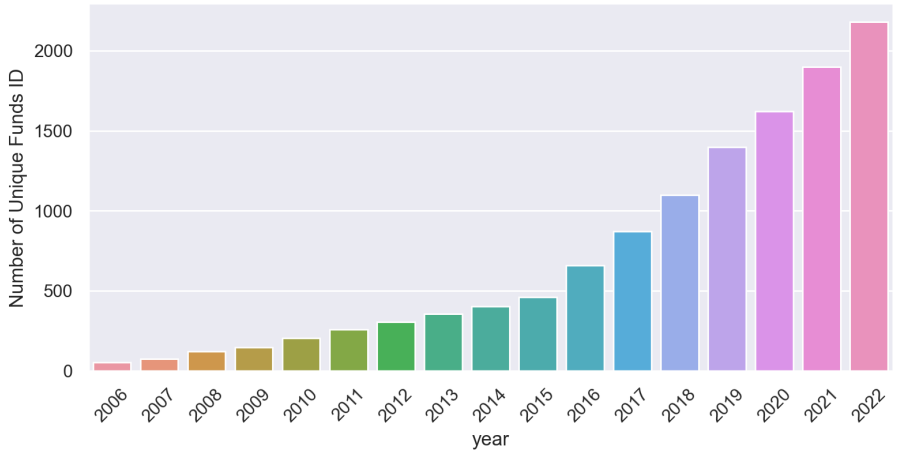


Figure 1.1: Annual count of selected funds.

The figure presents the annual count of selected funds from 2006 to 2022, each possessing a unique fund identifier. These funds were screened based on our pre-defined criteria.

We present the descriptive statistics on the monthly returns of our selected funds in Table 1.1. The mean returns of the selected funds range from -0.027 to 0.049, indicating significant variability in return performance over time. The mean return of the Shanghai Securities Open-end Fund Index provides a useful benchmark for comparison. The correlation between the mean returns of a selected fund data set and the market index is 0.86. Overall, the fund returns fluctuated over the period with wide deviations from the mean.

1.4.2 Data process and feature selection for machine learning models

As a machine learning-based exercise, our methodology diverges from the approach of traditional academic articles. To ensure the accuracy of our results, we divide our dataset into three subsets, specifically the training set, validation set, and test set. We used data from 2006 to 2018 as our training set, which we utilise to train our machine learning model. We ensure the accuracy of the model by including as much data as possible that matches the filter criteria. Additionally, 10% of the data is reserved for the validation set, which serves the purpose of impartially assessing the performance of a model fit on the training dataset. Finally, we select data from January 2019 to July 2022 as the test set to examine the predictability of our model. It is worth noting that the test data is completely isolated from the training and validation sets.

For this study, we have redefined our research problem as a supervised classification problem. To achieve this, we require both input features and output labels when utilising supervised machine learning models. Our methodology for setting up the output labels involves ranking the fund returns in descending order for each month and dividing the funds

Table 1.1: Summary statistics of returns

Year	mkt mean	mean	std	min	25%	50%	75%	max
2006	0.085	0.049	0.064	-0.090	0.013	0.040	0.098	0.203
2007	0.079	0.043	0.063	-0.181	0.010	0.042	0.079	0.281
2008	-0.079	-0.027	0.051	-0.239	-0.055	-0.020	0.008	0.129
2009	0.033	0.027	0.058	-0.255	0.011	0.027	0.053	0.202
2010	-0.015	0.004	0.040	-0.124	-0.021	0.005	0.025	0.155
2011	-0.020	-0.015	0.037	-0.182	-0.035	-0.013	0.008	0.094
2012	0.039	0.004	0.048	-0.214	-0.023	-0.002	0.025	0.269
2013	0.001	0.010	0.046	-0.180	-0.014	0.010	0.035	0.181
2014	0.029	0.016	0.041	-0.161	-0.005	0.010	0.033	0.298
2015	0.016	0.030	0.101	-0.270	-0.023	0.034	0.094	0.299
2016	-0.003	-0.006	0.067	-0.299	-0.018	0.001	0.018	0.238
2017	0.008	0.010	0.030	-0.136	-0.007	0.009	0.026	0.158
2018	-0.015	-0.018	0.039	-0.203	-0.042	-0.016	0.006	0.251
2019	0.019	0.027	0.047	-0.293	0.000	0.020	0.051	0.297
2020	0.019	0.037	0.065	-0.242	-0.004	0.031	0.076	0.294
2021	0.001	0.009	0.053	-0.233	-0.024	0.009	0.040	0.297
2022	-0.013	-0.011	0.080	-0.213	-0.079	-0.010	0.046	0.280

Table 1.1. presents a statistical summary of the monthly returns of our chosen funds from 2006 to 2022. In addition, it also includes the mean return of the index (mkt mean) as a benchmark for comparison. The statistical display is structured to show the mean, standard deviation, minimum value, 25th percentile, 50th percentile, 75th percentile, and maximum value of the fund returns, enabling the reader to understand the distribution of returns over the years.

into five groups (0,1,2,3,4). The top group, labelled as group 0, comprises funds that have returns in the top 20% of the overall sample in the forecast month. By correctly predicting the class, our model reveals the future position of the return of funds in the market.

Meanwhile, the input features are those attributes or characteristics of a dataset that are used to make predictions about the output. We considered two substantial input feature selection standards. First, we intended to use features that contain predictive information. The more informative the features are, the better our model can learn. Secondly, the features are both public and accessible to investors to make the forecast model practical. In this chapter, we use three groups of features. In the fund characteristics group, we contain those features related to the fund's past performance. Meanwhile, in the fund manager characteristics group, we select those features related to the fund manager information. The mixed characteristics group comprises those features derived from both fund and fund manager characteristics.

Through the above input feature group construction method, we can independently assess each feature group's predictive capability and make comparisons. If fund characteristics, as input features, play a dominant role in the model training process, it may suggest the presence of hidden or latent managerial attributes that are not otherwise contained within our selected fund manager characteristics but are otherwise reflected in the past performance characteristics of the fund. For example, certain fund managers may possess superior abilities in market timing, stock selection, or risk management that are not directly observable, yet can impact their consistent outperformance over time. On the other

hand, it may suggest that fund's past performance reflects hidden information that is not directly related to fund manager characteristics such as pure luck or investor objectives directed by the fund.

In order to achieve better model performance, we apply some special processing of the input features. Data scaling is a step in the preparation of numerical characteristics. Many machine learning algorithms require data scaling for optimal performance. We utilise the MinMaxScaler method to normalise input features on a monthly basis. MinMaxScaler adjusts the values to a specified range [0,1] without altering the original distribution's structure. It is worth noting that we only apply MinMaxScaler to numerical features and not the other one-hot encoded features. Please refer to the following formula for the calculation method:

$$Feature_{scaled} = \frac{Feature - Feature_{min}}{Feature_{max} - Feature_{min}} \tag{1.2}$$

Furthermore, we apply one-hot encoding to categorical features. This technique transforms categorical features into numerical arrays that are suitable for machine learning algorithms. It assigns binary values (0 or 1) to signify the existence or non-existence of that value in each observation. One-hot encoding allows the conversion of discrete data into continuous data, which can improve both the performance and interpretability of our models. In section 1.4.2.1 and section 1.4.2.2, we describe our chosen input features in more detail.

1.4.2.1 Fund Characteristics

No.	Feature descriptions	Symbol	Data type
1	Fund type (1: Equity fund, 0: Hybrid fund)	'EH'	One-Hot Encoded
2	The return of the fund in the past 1, 3, 6, 9, 12 months	'pre1m', 'pre3m','pre6m','pre9m','pre12m'	Numerical
3	Mean of past 3,6,12 monthly returns	'month3', 'month6', 'month12'	Numerical
4	Standard deviation of past 12-monthly returns	'stdmonth12'	Numerical
5	Sharpe ratio of past 12-monthly returns	'sharp12'	Numerical
6	Time lag 1,2,3 autocorrelation of the past 12 month's returns	'lag1','lag2', 'lag3',	Numerical
7	Downside returns	'downret'	Numerical
8	Drawdown and Maximum drawdown	'DD','MDD'	Numerical
9	Month star: the fund performance position in the past 1,3,6,9,12 months	'Month_star1', 'Month_star3', 'Month_star6', 'Month_star9', 'Month_star12'	Numerical
10	Gap feature in the past 1,3,6,9,12 months	'gap1', 'gap3', 'gap6', 'gap9', 'gap12'	Numerical
11	Features constructed based on past 12-month fund regression results (alpha, beta, p-value of alpha, p-value of beta)	'alpha', 'beta','p_beta', 'p_alpha'	Numerical
12	Returns of the market index in the past 1, 2, 3, 6, 9, 12 months	'Pre1mew', 'Pre2mew', 'Pre3mew', 'Pre6mew', 'Pre9mew', 'Pre12mew'	Numerical

Table 1.2: Fund Characteristics

Table 1.2 displays the features we employed in the fund characteristics group. Within the selected set of fund characteristics, we have employed 12 distinct feature groups. Feature 1 shows the type of fund, whether a stock fund

or a hybrid fund. We have labelled different fund categories using one-hot encoding, which converts the category information into binary '1' and '0'. Features 2 and 3 contain the basic past information of fund returns. The past returns from different months can represent the fund performance over different periods. If the market is inefficient, past returns are expected to provide a certain degree of predictive power. It is interesting to know whether the predicted power exists in the past trend of returns.

Feature 4 is the standard deviation of the returns over the past 12 months, whereas feature 5 shows the fund risk-adjusted return performance. Return fluctuations over time may reveal a trend or pattern, which, if it exists, can be useful for forecasting future returns. Getmansky et al. (2004) contend that serial correlations detected in a hedge fund's return series may be erroneous and deceptive regarding the predictability of fund returns since such serial correlations are mostly the product of the hedge fund's illiquidity exposure and return smoothing. Feature 6, we quote the auto-correlations of lag 1, lag 2 and lag 3 in the fund's returns series over the past 12 months, which is applied in Wu et al. (2021). By including these features, we attempt to remove the effect of false serial correlations from the original return series.

Sun et al. (2014) present unique evidence that hedge fund performance persists following poor market conditions but is not ongoing following robust markets. They construct two performance measures, DownsideReturns and UpsideReturns, conditioned on the level of overall fund sector returns. They suggest that Downside Returns can be a consistency indicator in predicting the future returns of the funds. In our research, we also calculate the downside returns, which are the average returns of the past 12 months when the market experiences a loss (Feature 7).

Our research is a classification problem. Our models attempt to predict the fund's future position in the market. Based on the historical returns, we calculate the fund's previous ranking in the market (Feature 9). To calculate Feature 9, similar to the output label, we sort the fund performance of the previous specified month in descending order and divide it into five groups.

Accordingly, we calculate a special indicator, 'Gap' (Feature 10). We try to use this feature to reflect the distance of each fund under different market performances. It is calculated as follows:

$$Gap_i = r_{i-1} - rm_{i-1} \tag{1.3}$$

where i is the number of months ago. For example, when $i = 1$, $Gap1$ represents the difference between the fund's return and the market's return in the previous month.

Feature 11 is the linear regression results with respect to the market returns over the past 12 months and the corresponding t-statistic. Finally, we include the previous market information (Feature 12) and try to connect this information with the fund's individual performance.

1.4.2.2 Fund manager characteristics

No.	Feature descriptions	Symbol	Data type
Education Background			
1	Whether information related to the educational background of the fund manager exists: Yes:1, No:0	'IsEducation'	One-Hot Encoded
2	Whether the fund manager has graduated from an internationally renowned institution/985 institution: Yes:1, No:0	Isinternational985'	One-Hot Encoded
3	Whether the fund manager has graduated from 211 institution: Yes:1, No:0	'Is211'	One-Hot Encoded
4	Whether the fund manager holds a PhD: Yes:1, No:0	IsPhd'	One-Hot Encoded
5	Whether the fund manager holds a master's degree: Yes:1, No:0	'IsMaster'	One-Hot Encoded
6	Whether information related to the major background of the fund manager exists: Yes:1, No:0	'IsMajor'	One-Hot Encoded
7	Whether the fund manager has graduated from a business-related major (finance, economics, accounting, management): Yes:1, No:0	'Isbusiness'	One-Hot Encoded
8	Whether the fund manager holds a certification (CFA, CPA, FRM): Yes:1, No:0	IsCertificate'	One-Hot Encoded
9	A composite of educational scores, including features 1, 2, 4, 5, 8	'EducationScore'	Numerical
General information			
10	The gender of the fund manager: Male: 1, Female: 0	'IsMan'	One-Hot Encoded
11	Nationality of Fund Manager: Mainland China: 1, others: 0	IsMainlandChina'	One-Hot Encoded
12	Whether the fund manager has overseas background: Yes:1, No:0	IsOveseaBack'	One-Hot Encoded
Team information			
13	Whether the fund manager independently manages the fund: Yes: 0, No: 1	teamworks'	One-Hot Encoded
14	Whether the team members of non-independently managed funds are different from the previous month: Yes: 1. No:0	Teamchange'	One-Hot Encoded

Table 1.3: Fund Manager Characteristics

Until December 2022, the CSMAR database contained records of 5,400 fund managers and their corresponding information. Whether fund manager characteristics are related to a fund's performance is one of the critical issues discussed in this paper. Specifically, we explore the impact of fund manager characteristics from three perspectives, namely education, general, and team information. To describe the fund managers comprehensively, we collected 14 characteristics from existing materials (Table 1.3). As the majority of fund manager characteristics are categorical features, we transformed them into binary variables using the one-hot encoding method. Under different market conditions, fund managers may make different decisions, and we aim to capture the relation between fund managers' characteristics and fund performance under specific market conditions. To achieve this goal, we trained a model that includes market returns from (Table 1.2 Feature 12).

In the education background section, we contain 9 features, which cover the fund manager's educational background, major information, certification status, and a composite score. We calculate the composite score to represent the fund manager's overall educational level of attainment. To obtain a higher degree, people need to spend more time studying.

A higher degree may represent a better knowledge base (Chevalier and Ellison, 1999). In the database, around 11% of fund managers are PhD degree holders. Most fund managers are master's degree holders, accounting for 95% of the sample. Unlike the U.S. market, few fund managers hold MBA/EMBA diplomas in the Chinese market. We checked the entire database and only about 1.5% of registered fund managers hold an MBA or EMBA degree.

Previous studies consider the SAT and GMAT scores as one of their inputs. In China, we cannot apply the above factors. As a comparison, we should use the college entrance examination score as a factor. However, since the educational background of fund managers is no longer limited to Chinese universities, we do not use college entrance examination results as our standard. Instead, we link admissions criteria to university rankings and notabilities. When considering the quality of the fund manager's educational background, we separate the university into two groups. According to the statistics for 2022, there are currently 3,013 universities in China, of which 39 are classified as Project 985 universities and 112 are classified as Project 211 universities. In the first group, the managers graduated either from famous international universities (Top QS 200) or project 985 universities. Project 985 was a terminated project first announced by the General Secretary of the Chinese Communist Party in 1998. The project's goal was to promote the development and reputation of the Chinese higher education system by founding world-class universities. Members of Project 985 are doctoral universities with very high research activities among all Chinese universities – Tier 1 universities among more than 3,000 higher education institutions in China, regarded as some of the most prestigious universities among all Chinese universities and which are consistently ranked among the best in the world. In the second group, the universities belong to Project 211, slightly behind Project 985. However, only 54% of fund managers' university information is contained in the database. Therefore, we also construct Feature 1 to indicate whether university information actually exists.

Besides considering the university, we also consider the subject background of the fund manager as one of the factors. Information as to the major is also incomplete, with approximately 28% of the data missing. We use Feature 6 to mark whether the major background of the fund manager exists in the database. Different majors have different course requirements. Business school students may have easier access to investment-related training and resources than others. The original major information in the database was reorganised to examine whether fund managers possess relevant professional backgrounds from business schools. Feature 7 is employed to establish this distinction. However, only 55% of fund managers graduated with a business-related degree. This may be due to the development of the fund's strategy. For example, with the increase in quantitative strategies, fund managers with a computer and quantitative background have become popular in the industry. A STEM degree pertains to a degree in the disciplines of science, technology, engineering, or mathematics. In the database, roughly 10% of fund managers possess a STEM degree.

Previous literature also considers the certificates that fund managers hold, e.g., Shukla and Singh (1994). In China, people must pass the fund practice qualification exam to become fund managers. This test is a national examination that

tries to ensure that fund managers attain basic knowledge and professional skills. To register as a fund manager, people must hold this certificate. Therefore, we employ the criterion as to whether the manager holds an international financial certificate as one of the indicators to measure the ability of fund managers rather than use this national certificate. Since these international certificates are not Chinese local examinations, only 6% of fund managers in the total sample have them. This result is much lower than the proportion of U.S. fund managers with certificates. Meanwhile, among fund managers with overseas certificates, 18% hold multiple certificates. These fund managers are more likely to concur with the valuation of the certificate.

In the general information section, we include the fund manager's gender, nationality, and overseas life experience. Moreover, based on current data, the fund market in China can be considered a male-dominated market. In the entire database, 77% of fund managers are male (until 2022, CSMAR).

In this article, we use Features 13 and 14 to describe the team management status. We use Feature 13 to mark whether the fund manager independently manages the fund. If a fund manager manages the fund independently, his or her influence over the fund should be greater than if he or she were a member of a team. At the same time, we use Feature 14 to monitor changes in the fund's management team.

1.5 Methodology

This paper stands out from traditional asset pricing research by incorporating machine learning models as a fundamental component. As described by Gu et al. (2020), machine learning in empirical asset pricing refers to a varied range of high-dimensional models designed for statistical prediction. The models are complemented by regularisation techniques to facilitate model selection and prevent overfitting. Additionally, efficient algorithms are used to sift through numerous potential model specifications. To accommodate readers without a background in machine learning, this section provides a brief introduction to these methods.

The field of machine learning encompasses two primary approaches, specifically supervised and unsupervised learning. In the former, outputs are labelled, and algorithms are trained to accurately classify or predict outcomes. The latter, however, requires no labelled outcomes and is instead utilised to uncover underlying data structures and relationships. In this paper, we address the challenge of forecasting mutual fund performance through a supervised learning framework. Our approach involves categorising funds into five groups based on their forecast month returns and using these group labels as the classification target variable. Specifically, funds with the highest monthly returns are placed in Class 0, while those with the lowest returns are placed in Class 4.

Previous empirical asset pricing studies have predominantly focused on utilising machine learning techniques to address three primary issues, namely prediction, variable selection, and functional form. To fit our main objective, we mainly

focus on variable selection. Therefore, we design three comparing groups to test our assumptions. We split the fund characteristics and fund manager characteristics into different groups as input features. We train the model separately with different features so that we can more intuitively compare the impact of distinctive features on the model.

This study employed three distinct machine learning model algorithms: neural network (deep learning), Random Forests (RF), and XG-Boost. Using NNs and RF models has garnered significant interest in recent research (DeMiguel et al., 2021; Kaniel et al., 2022). Both NNs and RF models are well-suited for multiclass classification problems due to their capability to generate multiple probabilities for each class. The NN model comprises interconnected neurons, whereas the RF model is based on a tree structure. Given the fundamental disparities in the construction of NN and RF models, in this research, we aim to compare the predictive power of these two models for fund performance. The XG-Boost model, proposed by Friedman in 2001, has received less empirical attention in asset pricing than the NN and RF models. There is a lack of relevant published studies utilising the XG-Boost model to predict future fund performance. Consequently, our research aims to address this gap and contribute to the existing empirical evidence on this model's application in asset pricing. By including the XG-Boost model in our study, we hope to provide valuable insights and ideas for future researchers. In subsequent sections, further details will be provided on the NN (deep learning), RF, and XG-Boost models.

1.5.1 Deep learning

NNs are widely considered to be one of the most powerful modelling tools in machine learning. They have emerged as the method of choice for tackling complex machine learning problems such as computer vision, natural language processing (NLP), recommendations, and fraud detection. Given that NNs are suitable for prediction tasks, they offer an ideal solution for the problem of fund performance prediction. Moreover, many researchers have acknowledged that NNs exhibit superior performance as compared to other machine learning methods, as they allow for non-linear feature interactions that may otherwise be overlooked by other approaches.

The basic structure of a NN comprises three layers:- an input layer, a hidden layer, and an output layer. In this research, we have utilised supervised deep learning (DL) as our fundamental framework. We employ 6 layers in our NN model. The term "deep" denotes the presence of more than one hidden layer in the network. This structure of adaptive layers facilitates the extraction of nonlinear features from the input data and enables their combination to describe the desired output variables.

During the training process, a DNN trains the data by using a given target. In the figure, the features represented by $X = x_1, x_2...$ are the different input factors intended to provide explanatory information. It is important to note that the DNN does not select the features itself, but rather assigns various weights to them during the training process, based on the results.

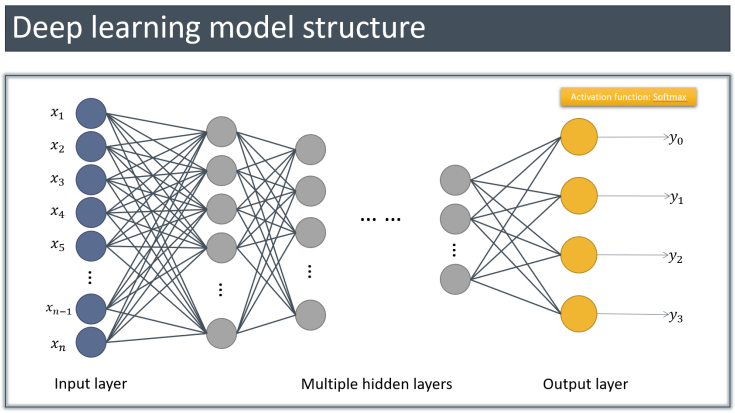


Figure 1.2: DNN structure

The target of training the DNN model is to minimise the cost function. A cost function is trying to evaluate how incorrect the model is in terms of its ability to predict the output. There are different types of cost functions. We have more than two label classes. Therefore, we chose the cross-entropy loss function as our cost function (see Equation 1.4). Before training the model, we transferred our target label into a one-hot representation.

$$-\sum_{c=1}^{M} y_{o,c} log(p_{o,c}) \tag{1.4}$$

where:

M = number of classes

y = binary indicator (0 or 1) if class label c is the correct classification for observation o

p = predicted probability observation o is of class c

The activation function in DL converts an input signal of a node to an output signal, which introduces nonlinearity into the network. Between the last hidden layer and the output layer, we chose Softmax as our activation function. The Softmax function is employed to transform a vector z of K real numbers into a probability distribution. The elements of the output vector are in the range (0, 1) and sum to 1.

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{1.5}$$

for i = 1, ..K and $z = (z_1, ..., z_K) \in R$.

1.5.2 Tree ensemble model

The RF and XG-Boost algorithms rely on an ensemble method as their basic premise. This technique enhances model robustness by combining the predictions of multiple base estimators built with the given learning algorithm. For tree

ensembles, the final prediction is the sum of predictions from each tree. Ensemble methods typically fall into two categories:- average methods and boosting methods. RFs belong to the average method category, wherein several estimators are built independently and their predictions are averaged. In contrast, XG-Boost belongs to the boosting method category, wherein base estimators are built sequentially, and the goal is to lessen the bias of the combined estimator.

1.5.3 Random Forests

RFs are proposed by Breiman (2001), which is a classification algorithm consisting of more than one decision tree. The main improvement of RFs is adding an additional layer of randomness to bagging (Breiman, 1996). Compared with NNs, the RF model requires less parameter tuning and is less computationally expensive when considering classification problems. Moreover, since the RF is a tree-based model, it does not require feature scaling.

Typically, individual decision trees exhibit high variance and tend to overfit. However, under the RF structure, each node is split using the best among a subset of predictors randomly chosen at that node, which can reduce overfitting. Although sometimes, the bias may slightly increase after combining diverse trees, the variance reduction is often significant. Meanwhile, the RF uses a rules-based approach instead of distance calculation so that no feature scaling is required for the data preparation. The basic principle of the RF model is to obtain all decision trees through parallel training of the sample subsets. The RF model randomly re-sampled the training samples, randomly selected features, and finally averaged the results of each tree. The process is shown in Figure 1.3.

There are four main steps involved in a RF algorithm:

Step 1: Assuming that the RF model receives a training sample S_0, the sample subsets S_i are randomly selected from S_0 with replacement. This procedure is known as bootstrap aggregation or "bagging". In this way, the RF model builds K decision trees, making them grow from S_i. Most of the data will be used multiple times in different subsets, S_i.

Step 2: RF randomly extracts features to train each decision tree (similar to "dropout"), thereby ensuring the independence and diversity of the trees. Suppose that the number of features is m. The number becomes 2m when the node splits for each decision tree. The Gini coefficient and the maximum information gain principles are used to obtain the variable importance. The RF model estimates the relative importance of each feature in the model using the out-of-bag (OOB) subset.

Step 3: OOB samples provide an unbiased estimation of the generalisation error. As the prediction performance of only one decision tree is unstable due to randomness, its generalisation error may be significant. However, as the number of trees increases, the generalisation error gradually decreases.

Step 4: The outputs of all K decision trees are averaged to obtain a final estimation.

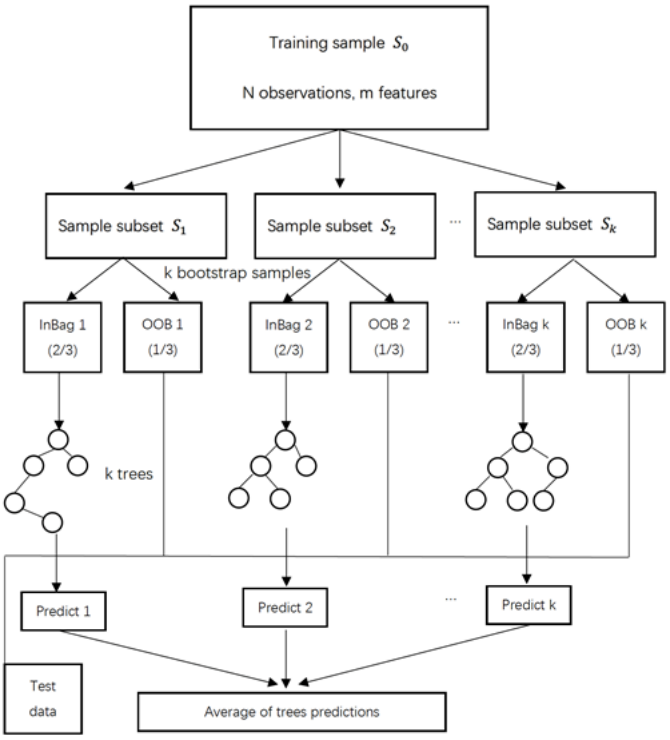


Figure 1.3: Random Forests

1.5.4 XG-Boost

Extreme Gradient Boosting (XG-Bboost) is a scalable machine learning system for tree boosting that focuses on computational speed and model performance. The term "gradient boosting" is first introduced by Friedman (2001). XG-Boost evolves from a single CLI program to a package. The design goal is to make the best use of available resources to train the model. As a highly sophisticated algorithm, XG-Boost exhibits an ability to deal with irregularities arising within the data. Compared with other algorithms, XG-Boost includes three critical features:

1. Compared with the traditional decision tree, XG-Boost can solve the missing data. Usually, the best split in the tree is the split that maximises the computed score, but the Sparsity aware feature in XG-Boost attaches a default direction to the split, which is a novel distributed weighted quantile sketch algorithm.

2. XG-Boost uses a block structure that stores the data in in-memory units. Data is sorted in the compressed column (CSC) format, with each column sorted by the corresponding feature value. This feature supports the parallel of tree construction.

3. XG-Boost supports continued training, which means we can further boost an already fitted model with new data.

XG-Boost also converts on the basis of the tree model when dealing with multi-classification problems. XG-Boost is trained by minimising the loss of an objective function against a dataset. Our objective function is specified as the "Softmax", while our loss function is specified as the "Multi-class log-loss". The log-loss is defined as the negative log-likelihood of a logistic model that returns y_{pred} probabilities for its training data y_{true}. The true labels for a set of samples can be represented by encoding them as a 1-of-K binary indicator matrix Y. For instance, $y_{i,k}$ is assigned a value of 1 to indicate that sample i is associated with label k, which is selected from a pre-determined set of K labels. To quantify the likelihood of these labels, a matrix P is constructed to hold probability estimates, where $p_{i,k}$ represents the probability, denoted as $Pr(y_{i,k} = 1)$. Consequently, the overall log loss for the entire set can be defined as follows:

$$L_{log}(Y, P) = -\frac{1}{N} \sum_{i=0}^{N-1} \sum_{k=0}^{K-1} y_{i,k} log(p_{i,k}) \tag{1.6}$$

1.5.5 Model evaluation methodologies

1.5.5.1 Accuracy scores

The accuracy score is a widely used evaluation metric in machine learning, one that measures the percentage of correctly classified instances by a model. In this research, we transfer the problem into a multi-class classification problem. Each month-end, all funds are sorted by return in descending order and divided into five classes $Y \in \{0, 1, 2, 3, 4\}$. The equation 1.7 displays how the accuracy score is calculated for the multi-class classification problem. Calculating the accuracy score allows us to assess the performance of our machine learning model. The accuracy score is directly determined based on the confusion matrix. Since our model is a multi-label classification, the function calculates subset accuracy. The funds' class labels, as predicted, must exactly match the corresponding set of tags in Y. A high accuracy score indicates a model that is effective and efficient, while a low score suggests that the model is not performing well in correctly predicting the class labels for the different categories.

$n_{samples}$ is the number of samples or observations in the data set. If $\hat{yi}$ denotes the predicted value for the i-th sample and yi represents the corresponding true value, the fraction of correct predictions over $n_{samples}$ can be defined as follows:

$$Accuracy(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{yi} = yi) \tag{1.7}$$

1.5.5.2 Importance analysis

The importance scores provide insight into the relative contribution of each feature to the predictive accuracy of the model. The importance is calculated based on the frequency of features selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees (Hastie et al., 2009). The greater the value, the greater the importance of the feature.

In this research, we follow the mean decrease in impurity. If we assume only two child nodes, the importance of node j can be calculated as follows:[1]

$$inode_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \tag{1.8}$$

$inode_j$: the importance of node j

$wsub(j)$: weighted number of samples reaching node j

$Csub(j)$: the impurity value of node j

$left(j)$: child node from left split on node j

$right(j)$: child node from right split on node j

The importance of each characteristic i on a decision tree is then computed as follows:

$$if_i = \frac{\sum_{j:nodej\ splits\ on\ feature_i} inode_j}{\sum_{k \in all\ nodes} inode_k} \tag{1.9}$$

The importance of feature i, calculated from all trees, is:

$$final\ score\ if_i = \frac{\sum_{j \in all\ trees} \frac{if_i}{\sum_{j \in all\ features} if_i}}{Total\ number\ of\ trees} \tag{1.10}$$

The more a feature reduces the impurity, the more important it is.

1.6 Empirical analysis

This chapter investigates how different sets of features affect the performance of machine learning models in forecasting mutual fund returns. We examine three sets of features with different emphases, namely fund, fund manager, and mixed characteristics. We train XG-Boost, RF, and DNN models using each set of features, and attempt to determine the

[1] Inspired by Stacey Ronaghan published in Towards Data Science. Please find more details from Scikit-learn documents

different impacts of these features on each model. The empirical results are primarily analysed from two perspectives, namely model performance and financial performance. In the model performance section, we focus more on the machine learning models themselves. We evaluate our model performance from two aspects:- the accuracy scores and the feature importance. Meanwhile, in the financial performance analysis section, we use conventional fund performance evaluation indicators to evaluate our portfolio performance.

1.6.1 Machine learning model performance

1.6.1.1 Accuracy scores

Our aim is to employ the accuracy score evaluation matrix to address two issues. First, we appraise the performance of the model and identify the presence of any overfitting in the models. Second, we want to compare the performance of models trained with different features. Table 1.4 presents the accuracy score of the training, validation and test sets of different models.

Table 1.4: Accuracy scores for each model

This table provides information concerning the performance of three models, namely the XG-Boost, the Random Forests, and neural networks, in predicting the fund performance group. The accuracy scores for the three models are presented for the training, validation, and test sets. The data used for the training set covers the period from July 2006 to December 2018, while the validation set consists of 10% of the data from the training set. The test set, on the other hand, covers the period from January 2019 to July 2022.

Model Type	Features	Train set	Validation set	Test set
	ALL	0.369	0.323	0.241
XG-Boost	*F*	0.372	0.326	0.243
	FM	0.251	0.196	0.203
	ALL	0.268	0.256	0.244
Random forest	*F*	0.269	0.258	0.243
	FM	0.210	0.183	0.199
	ALL	0.329	0.284	0.231
Neural networks	*F*	0.355	0.326	0.225
	FM	0.220	0.203	0.203

*ALL: Training with All feature
*F: Training only with fund characteristics
*FM: Training only with fund manager characteristics

One of the main problems in machine learning is overfitting. The model may be too complex and learn the 'noise' when using a large set of features. To address the overfitting issue, we apply distinct approaches for different models (please see methodology section for specifics). To evaluate the model performance, we split our data into training, validation, and test sets to test our model performance. If overfitting arises, the model will fit the training data with high accuracy although against the unseen data, it will achieve a low degree of accuracy.

As the table shows, the differences in accuracy scores crossing the three sets are slim. For all three models, the average gap between the training set and the validation set is less than 0.05. Overall, the XG-Boost and RF models display a slightly higher accuracy score than the NN models. Meanwhile, the accuracy gap of the XG-Boost model is slightly

larger than for the other two type models, with a maximum accuracy gap of 0.055. Nevertheless, this result falls within the acceptable range. The small accuracy score gap arising between the training set and the unseen data set indicates that our model does not suffer from an obvious overfitting issue.

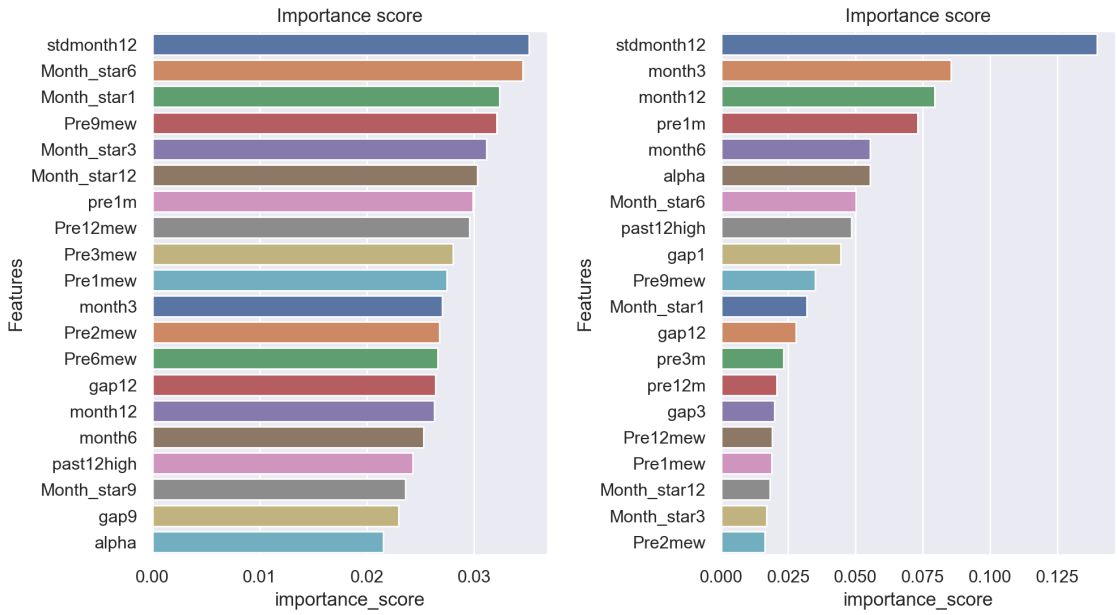
Meanwhile, the accuracy performance of the three feature groups is examined separately. Models trained with the fund characteristics exhibit a relatively high degree of accuracy in both validation and test sets. This indicates that these models have a certain ability to predict the correct fund classes with out-sample information. Meanwhile, these results also suggest that the fund characteristics are informative. In contrast, the managerial characteristics feature group has the least accuracy. From Table 1.4, we can see that the accuracy of the test set for feature group FM is approximately 0.2, which is close to the probability of a random guess. This means that only 20% of the test instances are correctly classified by our model. The result indicates that, when using fund manager characteristics, our model performs poorly and fails to generalise well with unseen data. The reasons for such low accuracy may vary, including the selected features not being informative enough, insufficient data, or the model not being complex enough to capture any underlying patterns in the data. Since we apply the same data set and the model structure in our research, the low accuracy score of the fund manager characteristics group implies that managerial characteristics provide less information to our supervised model than other feature groups.

Interestingly, although the total number of features increases after combining the fund characteristics with managerial characteristics, the accuracy has still not improved significantly. The accuracy score of the training set even slightly decreased after combining the fund manager characteristics. For example, the training set accuracy in the NN trained with fund characteristics is 0.355, while with all characteristics it is 0.329. We do not find any significant supportive evidence that fund manager characteristics have predictive power as regards fund performance from the accuracy score test results. In addition, the results suggest that the fund manager-related features may even add noise to the model-building process.

1.6.1.2 Feature importance

This section shows the feature importance calculated by the XG-Boost and the RF model. The importance provides a score that indicates how informative or valuable each feature is within the construction of the boosted decision trees in a given model. We rank factors from high to low according to their importance score and display them in the following figures. The importance scores permit us to understand our features more intuitively.

In Figure 1.4, the 20 features with the highest importance scores of both XG-Boost and RF models are presented. Interestingly, after combining fund and fund manager characteristics, the latter does not appear to be the most informative feature. This suggests that fund manager characteristics may not have a significant impact on mutual fund performance. Additionally, our findings indicate that neither the fund manager's educational background nor general information



(a) XG-Boost **(b)** Random Forests

Figure 1.4: Top 20 important features

The group of all features involves 54 distinct features. Based on the features' respective importance scores, the top 20 features with the highest scores are presented in the figure. We display the outcomes stemming from two distinct models:- XG-Boost and Random Forests.

provides more information than the fund characteristics. In other words, the contribution of fund manager characteristics in predicting fund performance is negligible. This is further supported by the low accuracy score of our model as trained with fund manager characteristics. These results suggest that there is no clear evidence from the model performance that the group of fund manager characteristics predicts fund performance.

Figure 1.5 (a) depicts the findings of the XG-Boost model, wherein the importance score of each feature is observed to be consistently low and, moreover, similar in magnitude. The range of importance scores is within the narrow spectrum of 0.04 to 0.06, which is indicative of the model's inability to discern and weigh the contributions of individual features effectively. Moreover, this result underscores the features' limited ability to capture the underlying patterns in the data and facilitate accurate predictions of the target variable. In essence, the findings suggest that the model's predictive performance is not significantly influenced by those fund manager characteristics under consideration.

The findings presented in 1.5 demonstrate notable disparities in results between RF and XG-Boost models. Specifically, 1.5 (b) displays the result of the RF model, which exhibits feature importance scores that are more dispersed in comparison to the XG-Boost model when trained on identical data and features. It should be noted that, while the RF and XG-Boost are similar models, their underlying algorithms differ in that RF utilises a bagging ensemble model, while XG-Boost employs a boosting ensemble model. Consequently, discrepancies in results may arise. The analysis revealed that feature 'IsPhd' obtained the highest importance score of 0.1, while feature 'IsMaster' obtained the lowest

(a) XG-Boost **(b)** Random Forests

Figure 1.5: Importance scores of fund manager characteristics.

The above figures illustrate the importance score of thirteen distinct characteristics associated with fund managers. The importance score is based on the model trained solely on the allocated group of fund manager characteristics. We present the outcomes stemming from two distinct models, XG-Boost and Random Forests.

score of only 0.01. Notably, the magnitude of the difference between these two features was found to be ten-fold. These findings suggest that 'IsPhd' may be crucial in forecasting fund performance, while 'IsMaster' may have a limited impact in comparison. In the subsequent discussion, we undertake an extensive analysis of the result in question, incorporating additional information to provide a more comprehensive understanding.

1.6.2 Out-of-sample financial performance analysis for one-month target

This section examines the out-of-sample financial performance of the machine learning methods in forecasting fund returns. The test is based on our test set, which is separate from training data. The empirical results cover the period from the end of January 2019 to the end of July 2021. We construct equally-weighted fund portfolios based on the fund classes label as predicted by the machine learning models. Our top-class portfolios comprise those funds predicted by the machine learning models into Class 0. These portfolios are rebalanced each month. Table 1.5 displays the out-of-sample financial performance produced by our portfolios. To provide a more concise table description, we code the portfolios using abbreviations for model type and input features. We select two benchmarks as being comparable: the markets index and the EW index. The market index selected is the Shanghai Securities Open-end Fund Index, while the EW index is the equally weighted performance of the funds in our test set.

It should be noted that the top-class portfolio obtained from training the RF model based on fund manager characteristics (RF-FM) is not presented in the table. This is because our RF model does not perform well in distinguishing and

classifying the top-class portfolio when trained on fund manager characteristics. Let us revisit the findings from Section 1.6.1.1, where the RF model, as trained with fund manager characteristics, yielded a test accuracy score of only 0.199, which was inferior to the accuracy of random guessing. Taking into account our previous discoveries, we hypothesise that the insufficient predictive performance of the RF model in distinguishing funds belonging to the top-class group during training with fund manager characteristics may be attributed to the inadequacy of this feature in providing significant information for predicting fund performance.

It is evident that, with the exception of the portfolio based on the NW-FM, all our other top-class (Class 0) portfolios demonstrate higher annualised returns than the benchmark we have chosen. The top-class portfolios constructed using the models trained with fund manager characteristics exhibit significantly lower returns as compared to other portfolios. Upon conducting a comparative analysis, it was observed that the inclusion of fund manager characteristics in the input features of both XG-Boost and RF models resulted in a decrease in the annual returns. This phenomenon is particularly evident in the case of the XG-Boost model, where the annual return for the portfolio constructed solely using fund characteristics training is 0.293, while the inclusion of fund manager characteristics resulted in a decrease in annual returns by 0.008. In theory, a greater number of effective characteristics can provide more information and improve the accuracy of predictive results. However, our empirical findings suggest that, rather than providing effective predictive information, fund manager characteristics introduce more noise to the tree-based models. On the contrary, even the worst-performing top-class portfolio constructed solely based on fund characteristics has an annualised return that is 0.021 above the EW index. This result confirms that the models based on fund information have the ability to select funds that outperform the market.

Although the results of annualised return in Table 2.3 display the power of our models, such raw returns could be due to compensation for risk-taking. For investors, it is more interesting to study the relationship arising between risks and returns. We select the Sharpe and Sortino ratios to present the risk-adjusted performance (Sharpe, 1966, 1998). The Sharpe ratio is the average return earned over the risk-free rate divided by the volatility. A higher Sharpe ratio represents a higher reward to variability. The Sharpe ratio considers both the upside and downside volatility, which may penalise the high positive returns. Therefore, we calculate the Sortino ratios (Sortino and Price, 1994) to measure the performance of the investment relative to the downward deviation.

We observe from the table that the risk-adjusted performance of our top-class portfolios is better than the market index. The highest Sharpe ratio of our top-class portfolio is 1.43 (*XGB-F*) which is more than twice as high as the market index. The Sharpe ratio of the XG-Boost and RF models drops when paired with fund management features. Unfortunately, with the exception of portfolios constructed by XG-Boost, the Sharpe ratios for our other portfolios do not outperform the equally weighted index.

Table 1.5: Financial performance

The following table depicts the financial performance of our portfolios. We train the models with the training set (July 2006 – Dec 2018) and use these models to predict the performance of funds in the test set (Jan 2019 – July 2022).

	Annual returns	std	Sharpe ratio	Sortino	Cumulative return	MaxDD
Market index	0.110	0.141	0.670	1.157	1.417	-0.187
EW index	0.228	0.169	1.263	2.361	2.102	-0.214

(a) Beachmark

Model	Class	AR	std	Sharpe ratio	Sortino	Cumulative return	MaxDD	Turnover
XG-Boost (XGB-F)	***0***	0.293	0.195	1.43	2.72	2.58	-0.226	52.6%
	1	0.231	0.153	1.41	2.78	2.14	-0.188	71.4%
	2	0.208	0.163	1.18	2.22	1.97	-0.205	54.2%
	3	0.167	0.156	0.97	1.75	1.71	-0.199	74.1%
	4	0.229	0.186	1.15	2.05	2.09	-0.256	50.0%
Neural Networks (NW-F)	***0***	0.249	0.209	1.12	1.91	2.20	-0.298	69.8%
	1	0.188	0.140	1.24	2.28	1.86	-0.195	82.5%
	2	0.210	0.162	1.20	2.24	1.99	-0.203	34.5%
	3	0.228	0.173	1.23	2.34	2.10	-0.203	82.7%
	4	0.278	0.207	1.27	2.41	2.44	-0.261	58.0%
Random Forest (RF-F)	***0***	0.289	0.218	1.26	2.22	2.51	-0.281	37.5%
	1	0.256	0.176	1.37	2.58	2.31	-0.222	85.9%
	2	0.209	0.159	1.22	2.31	1.98	-0.194	23.0%
	3	0.154	0.166	0.83	1.45	1.63	-0.244	89.2%
	4	0.227	0.202	1.05	1.89	2.05	-0.283	61.1%

(b) Train with the fund characteristics

Model	Class	AR	std	Sharpe ratio	Sortino	Cumulative return	MaxDD	Turnover
XG-Boost (XGB-FM)	***0***	0.238	0.167	1.34	2.63	2.14	-0.203	68.2%
	1	0.227	0.170	1.24	2.34	2.09	-0.217	35.0%
	2	0.223	0.164	1.27	2.39	2.07	-0.205	40.0%
	3	0.215	0.162	1.24	2.27	2.02	-0.204	46.1%
	4	0.247	0.175	1.33	2.49	2.24	-0.221	34.3%
Neural Networks (NW-FM)	***0***	0.188	0.154	1.12	2.15	1.84	-0.208	30.3%
	1	0.192	0.158	1.12	2.02	1.87	-0.217	15.2%
	2	0.228	0.167	1.28	2.39	2.11	-0.210	3.6%
	3	0.234	0.174	1.26	2.36	2.14	-0.220	3.4%
	4	0.223	0.166	1.25	2.32	2.07	-0.217	5.7%
Random Forest (RF-FM)	***1***	0.205	0.174	1.09	2.05	1.94	-0.258	28.1%
	2	0.229	0.168	1.27	2.37	2.11	-0.213	1.2%
	3	0.199	0.167	1.10	2.11	1.90	-0.213	17.6%
	4	0.220	0.176	1.17	2.19	2.04	-0.229	20.1%

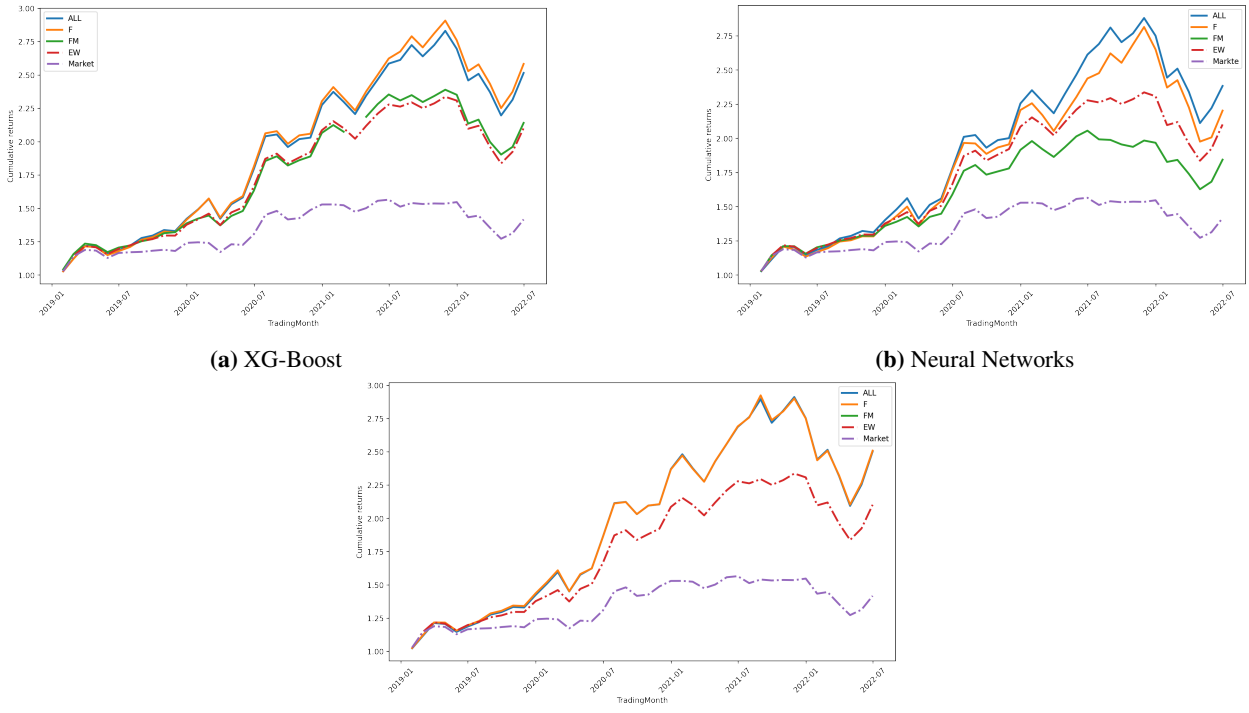
(c) Train with the fund manager's characteristics

Model	Class	AR	std	Sharpe ratio	Sortino	Cumulative return	MaxDD	Turnover
XG-Boost (XGB-ALL)	***0***	0.285	0.194	1.39	2.63	2.52	-0.224	52.7%
	1	0.230	0.154	1.40	2.81	2.14	-0.174	73.5%
	2	0.212	0.162	1.22	2.33	2.00	-0.200	52.4%
	3	0.171	0.162	0.97	1.76	1.74	-0.206	75.9%
	4	0.230	0.183	1.17	2.09	2.10	-0.253	50.0%
Neural Networks (NW-ALL)	***0***	0.271	0.204	1.26	2.20	2.38	-0.267	63.4%
	1	0.230	0.165	1.30	2.50	2.12	-0.208	69.1%
	2	0.208	0.160	1.21	2.30	1.97	-0.194	49.6%
	3	0.182	0.160	1.04	1.87	1.81	-0.219	74.3%
	4	0.270	0.200	1.27	2.34	2.38	-0.258	66.3%
Random Forest (RF-ALL)	***0***	0.288	0.218	1.25	2.22	2.50	-0.281	33.3%
	1	0.260	0.176	1.39	2.62	2.34	-0.218	85.6%
	2	0.211	0.160	1.23	2.35	2.00	-0.195	23.3%
	3	0.153	0.166	0.83	1.47	1.63	-0.239	86.4%
	4	0.220	0.201	1.02	1.83	2.01	-0.274	63.1%

(d) Trained with all characteristics

The standard deviation of top-class portfolios based on fund and all characteristics is greater than 0.19, exceeding that of both the market index and the EW index. Conversely, the portfolios based on fund managers' characteristics exhibit a low standard deviation, while those constructed from the XG-Boost and NN models show less standard deviation than the EW index. It should be noted, however, that while the portfolios based on fund managers' characteristics display lower volatility, their Sharpe ratio remains inferior to that of other portfolios. This finding emphasises that the profitability of portfolios based on fund managers' characteristics is lower as compared to other portfolios.

The Sortino ratios of the XG-Boost and RF models exhibit the same trend as the Sharpe ratios. The Sortino ratios of *XGB-ALL* and *RF-ALL* portfolios are lower than those of the *XBG-F* and *RF-F* portfolios. It is noteworthy that the performance of the NN model differs significantly from that of the other two models. After incorporating fund manager characteristics, both the Sharpe and Sortino ratios of the NW-ALL portfolio exhibit a notable increase. It is hypothesised that the noise generated by fund manager characteristics may have a greater impact on tree-based models than on NNs.



(a) XG-Boost

(b) Neural Networks

(c) Random Forests

Figure 1.6: Cumulative returns and maximum draw-down.

The cumulative returns show the total change in the investment of the portfolio over the test period. By monitoring the cumulative return, we can observe the performance of the top-class portfolios more intuitively. At the period end, almost all our top-class portfolios outperformed the benchmark index. Portfolios based solely on fund characteristics evidently outperformed the EW index by at least 0.10. In each combination, we can observe that portfolios only rely on

the fund manager's information and perform relatively poorly. The cumulative return of the *NW-FM* portfolio is only 1.845 which is lower than the EW index. Meanwhile, as Figures 1.6 show, after adding the fund manager characteristics, the performance of the NN models increases from 2.20 *NW-F* to 2.38 *NW-ALL*. However, the cumulative returns of the XG-Boost and RF models decrease. The cumulative return of the *XGB-ALL* portfolio is 0.06 lower than that of the *XGB-F*. This phenomenon casts doubt on the predictive power of fund managers' information as it implies that fund manager information is likely to introduce noise to top-class portfolio construction.

Maximum drawdown (MDD) represents the maximum observed loss from a peak to a trough of a portfolio. MDD measures the size of the largest loss. In comparison to the benchmarks, the majority of our portfolios exhibit a higher MDD. Among them, the largest MDD can reach -0.298 (*NW-F*). Interestingly, the MDD of fund manager information portfolios is smaller than the other portfolios, with the lowest being -0.203 (*XGB-FM*). After combining the previous results, although a low maximum drawdown is preferred, the overall financial performance of this portfolio does not outperform the others. This issue arises because MDD does not measure the loss frequency or the size of gains.

Table 2.3 also presents the average turnover rate for each portfolio. To obtain the average turnover rate, we follow a three-step process. First, a portfolio is constructed based on the classification results for each respective month. Second, the turnover rate is calculated by determining the percentage of funds in which their classification label changes in the subsequent month. The monthly turnover rate represents the percentage of our portfolio's holdings that changed over the month. Thirdly, this process is repeated monthly, and the average turnover rate is obtained by calculating the mean of the monthly turnover rates.

Table 1.6: Average turnover rate of true label.

	0	**1**	**2**	**3**	**4**
True label	72.5%	77.7%	76.5%	78.3%	74.2%

For comparison purposes, the average turnover rate of the true labels is displayed in Table 1.6. The testing period reveals that the average turnover rate of the actual label portfolios is high, with an average turnover rate of each class being above 70%. The location of most funds in the market changes in the subsequent (adjacent) month, indicating a lack of consistency in the Chinese fund's performance during the testing period.

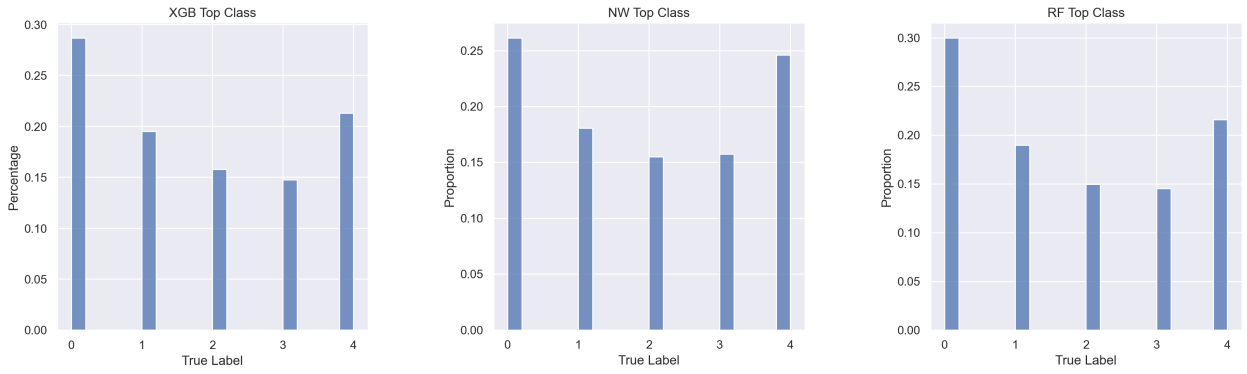
It is intriguing to note that, when trained with fund characteristics, the average turnover rate displays an 'M' shape. The average turnover rate of groups 1 and 3 is significantly higher than that of groups 0, 2, and 4. This phenomenon is observed in XG-Boost, NN, and RF models. It indicates that funds predicted to belong to groups 1 and 3 have a high probability of being predicted to belong to other groups in the next month. The highest average turnover rate reaches 89.2%, with almost 90% of the funds in the Class 4 portfolio moving to a different group in the subsequent month. After combining both fund and fund manager characteristics, the average turnover rates slightly decrease for the

NN and RF models. Though fund manager characteristics can, to some extent, enhance the portfolio's stability, the improvement is not significant.

It is worth noting that the features in the fund manager characteristics group change less as compared to those within the fund characteristics group. The lower average turnover rate observed in the NN and RF models may indicate insensitivity to modifications in the input feature values. Consequently, these models may be incapable of predicting changes in future fund performance based on fund manager characteristics. This assertion is bolstered by two pieces of evidence. First, the average turnover rate for classes 2, 3 and 4 in NNs ranges from 3.4% to 5.7%, which is markedly lower than for the other classes. Second, the RF models are unable to identify the fund that should belong to Class 0. These outcomes can provide further proof that fund manager characteristics do not offer sufficient information to predict future fund performance.

1.6.3 The evidence of bimodality

In order to better observe the financial performance of the models constructed by fund characteristics, we calculate the annualised returns for each forecast group separately. According to an ideal situation, the lower the predicted group, the greater the annualised return present. However, from Table 2.3 we can see that this trend is not obvious. The return distribution of investment portfolios constructed through groupings based on model predictions exhibits a closer proximity to a U-shape. This shape bears a striking resemblance to the morphology described in Han (2022). Han (2022) documents the bimodality of momentum stock returns wherein both high- and low-momentum stocks have non-trivial probabilities for both high and low returns.

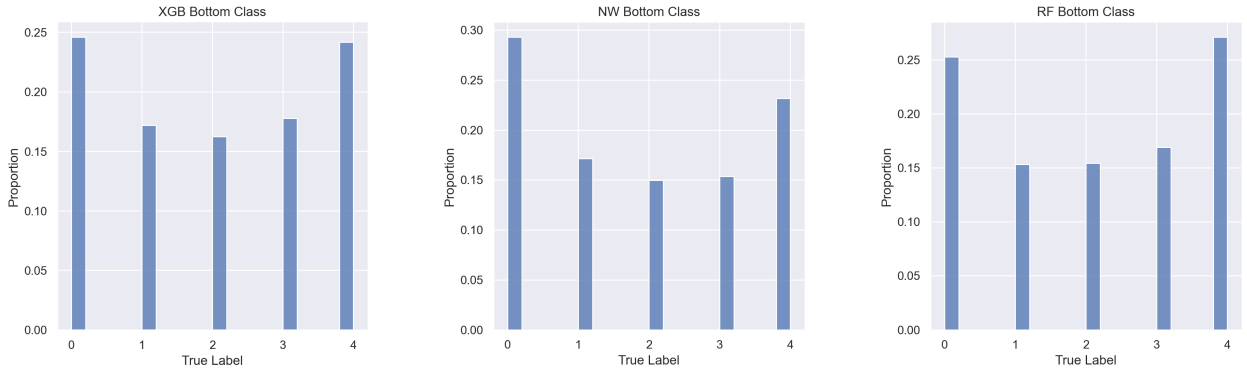


(a) XG-Boost **(b)** Neural Networks **(c)** Random Forests

Figure 1.7: True labels in top class

The figure shows the true group label of funds predicted to be classified in group 0 by the XG-Boost, neural network, and Random Forest models, respectively.

To explore the reason for this problem, we split the test set by the true group label of funds. Then, we calculate the number of funds based on the set of labels as predicted by the model. Figure 1.7 displays the funds are predicted to be



(a) XG-Boost **(b)** Neural networks **(c)** Random Forests

Figure 1.8: True labels in bottom class

The figure shows the true group label of funds predicted to be classified in group 4 by the XG-Boost, neural networks, and Random Forests models, respectively.

classified into Group 0 (TOP 20%), while Figure 1.8 displays the funds predicted to be classified into Group 4 (Bottom 20%). By visualising, we can see how the funds are labelled in more detail. From the figure, we can see an interesting phenomenon. Many funds that belong to the real Group 4 are incorrectly classified into Group 0. This means that many poorly performing funds are added to our top-class portfolios, adversely affecting the performance of our portfolios. This unfavourable error also occurs in those funds belonging to Group 4. This phenomenon poses a significant risk to investors. The presence of such bimodality implies that, when investors choose to invest in funds that are forecast to have high returns based on the model, they are also highly likely to experience financial losses in the future.

The models erroneously predict that some excellent funds are poor. This explains why fund portfolios constructed with funds in forecast Group 4 have a relatively high annualised return. These results imply that such funds perform exceptionally well and those that perform exceptionally poorly may have similar fund characteristics. These features can mislead machine learning models, causing the models to produce incorrect predictions. One possible influencing factor is the standard deviation of the fund's returns over the past year. The analysis in Figure 1.4 reveals that fund feature 4, which represents the standard deviation of the fund's past 12-month returns, exhibits the highest importance score in both the XG-Boost and RF models. Moreover, when examining the out-of-sample financial performance (refer to Table 1.5 (b)), it is evident that the standard deviation of Class 0 and Class 4 is significantly higher when compared to the other Classes. This pattern is consistent across all three models evaluated. To enhance the predictive ability of the fund's future returns and mitigate the occurrence of such bimodality, future researchers can delve deeper into the historical standard deviation of fund returns. By disassembling the input features in more detail, it is possible to enhance the model's accuracy in forecasting a fund's future performance.

1.6.4 Performance analysis with different training targets

Previous model performance tests have demonstrated the predictive ability of fund characteristics. This section aims to further extend the analysis of machine learning models constructed solely based on these fund characteristics. To obtain more information for analysis, new targets are created to train the models. The newly established multi-class classification labels are based on cumulative returns over periods of three and six months.

Table 1.7: Accuracy score of models trained with fund characteristics.

The accuracy scores of three models, namely XG-Boost, Random Forests and neural networks, as trained with different output labels, are presented in the table. The models employ the group of fund characteristics as input features to train with three distinct output targets. The three output targets were generated from the future one-month returns, future three-month cumulative returns, and future three-month cumulative returns, respectively. The data used for the training set covers a period from July 2006 to December 2018, while the validation set consists of 10% of the data from the training set. The test set, on the other hand, covers the period from January 2019 to July 2022.

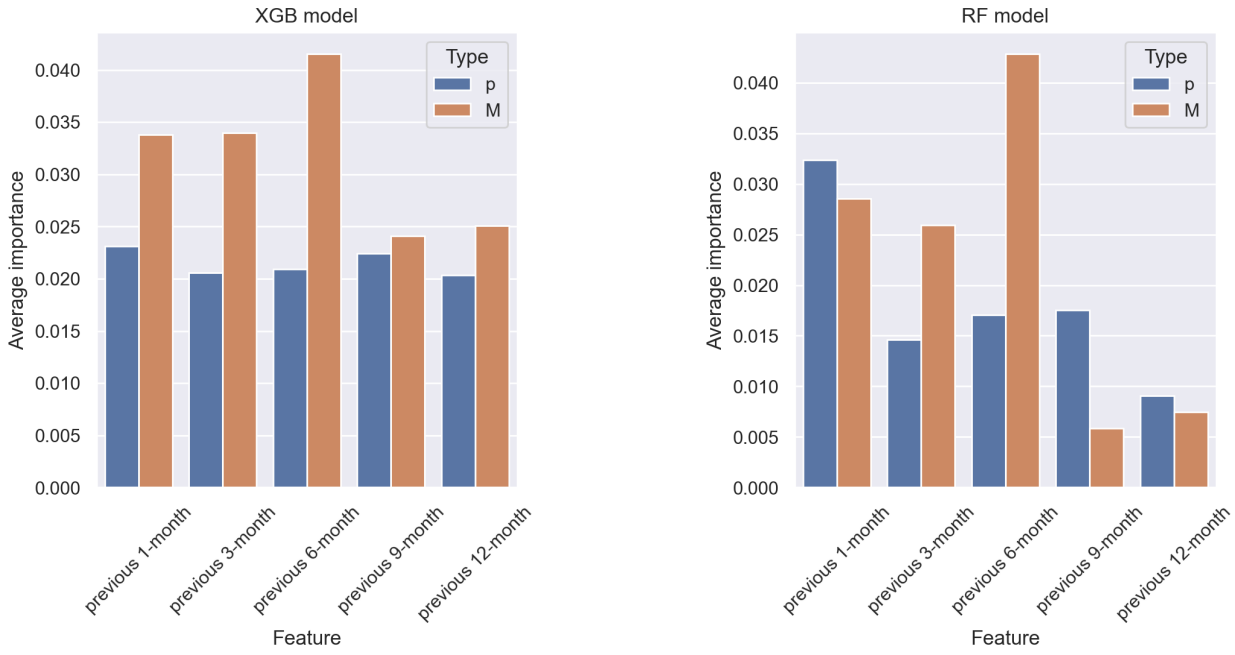
	Train set Accuracy	Validation set Accuracy	Test set Accuracy	Target
XG-Booost	0.372	0.326	0.243	*1-month*
	0.381	0.327	0.235	*3-month*
	0.366	0.328	0.233	*6-month*
Neural networks	0.355	0.326	0.225	*1-month*
	0.366	0.336	0.233	*3-month*
	0.345	0.320	0.223	*6-month*
Random forest	0.269	0.258	0.243	*1-month*
	0.288	0.283	0.234	*3-month*
	0.286	0.277	0.238	*6-month*

One of the main objectives of this inquiry is to investigate whether predictive models, developed solely based on the allocated group of fund characteristics, can accurately forecast fund performance over time. The accuracy score of the models is demonstrated in Table 1.7, which shows that the highest accuracy score of test set performance is 0.243, while the lowest accuracy score is 0.223. It is noteworthy that all models outperform random classification, even after the training target changes. Therefore, the results suggest that our models trained with fund characteristics possess a certain degree of predictability regarding the future performance of funds. The models are solely based on fund characteristics, and the test results provide supportive evidence for the predictive power of fund information. Consequently, the findings of this study highlight the significance of fund characteristics in predicting future fund performance classes.

The second main objective of the analysis is to investigate whether the predictive power of the models varies with the forecast period. The results show that the XG-Boost model trained with the one-month target exhibits the highest test set accuracy of 0.243. However, as the predicted period increases, the accuracy score decreases, in line with the logical expectation that the predictive power of the input factors declines when the time horizon becomes more extensive. Similar findings are observed for the NN and RF models. In contrast, the results of NN and RF models are not as fully expected as those of the XG-Boost model. However, the NN model trained with the six-month target performed most poorly, with an accuracy of 0.223. Meanwhile, in the RF model group, the 1-month target model displays the highest

accuracy score of 0.243. These results also can provide some evidence that the prediction power of fund characteristics may decline when predicting targets in the far future. Therefore, it is crucial for investors and financial analysts alike to consider the appropriate forecast horizon when using machine learning models for fund performance classification.

Meanwhile, we demonstrate some interesting findings in Figure 1.9. Before we begin, let us review what the features in group *pre*(Table 1.2 Feature 1) and *Month star* (Table 1.2 Feature 9) represent. The features in the *pre*-group represent the previous month's returns of the fund. These focus only on the fund's historical performance. The features in the *Month star* contain the fund ranking information derived from past fund returns.



(a) XG-Boost **(b)** Random Forests

Figure 1.9: The importance scores of *pre* and *Month star*

This figure displays the average importance scores of *pre* and *Month star* after training on different targets, using both XG-Boost and Random Forest models. The X-axis of the figure indicates the time period during which the features were generated. The legend on the figure presents p and M as representatives of feature *pre* and feature *Month star*, respectively.

The present test displays the relative importance of different features in predicting the future position of funds in the market. In Figure 1.9, it is discernible that the importance scores of features in group *Month star* are generally higher as compared to those in group *pre*. This trend is particularly pronounced in the XG-Boost model. The training objective of our models is to predict where the fund's future returns will rank in the market. The test results formally establish that factors which combine the performance of other funds in the market are more important than the fund's own return information. This result implies that the emphasis placed on previous fund ranking in the market can provide predictive information for the future ranking of the fund in the market.

Furthermore, we note that the importance of the feature *Month star* gradually decreases after a period of six months. This observation suggests the existence of information decay, where the value of *Month star* features in predicting fund future performance deteriorates over time. The earlier the information a given feature contains, the less information it can therefore provide. Our results provide supporting evidence of short-term persistence in mutual fund performance (Nicolas, 2004. Overall, in this part of the discussion, we emphasise the importance of cross-sectional performance characteristics in predicting future fund performance.

Table 1.8 presents the financial performance of the portfolios, with the annualised and cumulative returns displaying a significant decline as the training target period is extended. Of particular note is the pronounced phenomenon observed in the XGB model. When the subsequent month's return rate is used to construct the label for training the model, the annual return of the top-class portfolio is 0.293 with a cumulative return rate of 2.58. However, when the label is constructed using a three-month cumulative return, the annual return of the portfolio drops to 0.239 with a cumulative return rate of 2.24. Further, when using a six-month label training model, the annual return of the portfolio declines even further to 0.213 with a cumulative return rate of 2.07.

The observed increase in the Sharpe and Sotino ratios can be attributed to the decrease in standard deviation. This decrease is a result of converting the cumulative return rate into a monthly return while constructing the portfolio performance to enable comparison. These values, however, are not suitable for cross-table comparisons. Moreover, expanding the target displays a more prominent bimodality in returns. Under the 3-month NW model, the return indicators and risk-adjusted ratios for the bottom portfolio exhibit only a slight difference from the top portfolio. The difference in annualised return is a mere 0.001.

Table 1.8: Financial performance

The following table depicts the financial performance of our portfolios. We train the models with the training set (July 2006 – Dec 2018) and use these models to predict the performance of funds in the test set (Jan 2019 – July 2022).

Model	Class	AR	std	Sharpe ratio	Sortino	Cumulative return	MaxDD
XG-Boost (XGB-F)	***0***	0.293	0.195	1.43	2.72	2.58	-0.226
	1	0.231	0.153	1.41	2.78	2.14	-0.188
	2	0.208	0.163	1.18	2.22	1.97	-0.205
	3	0.167	0.156	0.97	1.75	1.71	-0.199
	4	0.229	0.186	1.15	2.05	2.09	-0.256
Neural Networks (NW-F)	***0***	0.249	0.209	1.12	1.91	2.20	-0.298
	1	0.188	0.140	1.24	2.28	1.86	-0.195
	2	0.210	0.162	1.20	2.24	1.99	-0.203
	3	0.228	0.173	1.23	2.34	2.10	-0.203
	4	0.278	0.207	1.27	2.41	2.44	-0.261
Random Forest (RF-F)	***0***	0.289	0.218	1.26	2.22	2.51	-0.281
	1	0.256	0.176	1.37	2.58	2.31	-0.222
	2	0.209	0.159	1.22	2.31	1.98	-0.194
	3	0.154	0.166	0.83	1.45	1.63	-0.244
	4	0.227	0.202	1.05	1.89	2.05	-0.283

(a) Train with the 1-month target

Model	Class	AR	std	Sharpe ratio	Sortino	Cumulative return	MaxDD
XG-Boost (XGB-FM)	***0***	0.239	0.120	1.86	4.85	2.24	-0.167
	1	0.189	0.105	1.66	4.23	1.89	-0.148
	2	0.162	0.098	1.51	3.35	1.73	-0.154
	3	0.178	0.096	1.71	3.74	1.83	-0.153
	4	0.231	0.127	1.70	3.75	2.17	-0.206
Neural Networks (NW-FM)	***0***	0.251	0.129	1.83	4.83	2.32	-0.175
	1	0.190	0.099	1.76	4.77	1.90	-0.132
	2	0.176	0.102	1.59	3.67	1.81	-0.157
	3	0.170	0.104	1.49	3.11	1.77	-0.176
	4	0.250	0.131	1.80	4.08	2.31	-0.207
Random Forest (RF-FM)	***0***	0.245	0.128	1.80	4.23	2.27	-0.195
	1	0.199	0.106	1.74	4.33	1.96	-0.156
	2	0.161	0.098	1.50	3.49	1.73	-0.149
	3	0.142	0.091	1.40	3.00	1.62	-0.144
	4	0.236	0.129	1.71	3.83	2.20	-0.212

(b) Train with the 3-month target

Model	Class	AR	std	Sharpe ratio	Sortino	Cumulative return	MaxDD
XG-Boost (XGB-ALL)	***0***	0.213	0.084	2.34	5.73	2.07	-0.145
	1	0.162	0.073	2.03	5.65	1.74	-0.122
	2	0.137	0.068	1.78	4.25	1.60	-0.133
	3	0.121	0.059	1.80	3.83	1.51	-0.126
	4	0.202	0.085	2.20	4.90	1.99	-0.162
Neural Networks (NW-ALL)	***0***	0.221	0.086	2.39	5.78	2.12	-0.144
	1	0.147	0.067	1.97	5.59	1.65	-0.105
	2	0.161	0.075	1.95	4.74	1.74	-0.142
	3	0.160	0.075	1.95	4.38	1.73	-0.144
	4	0.213	0.086	2.30	4.98	2.07	-0.169
Random Forest (RF-ALL)	***0***	0.224	0.089	2.35	5.37	2.14	-0.157
	1	0.162	0.081	1.82	4.49	1.74	-0.145
	2	0.134	0.066	1.80	4.44	1.58	-0.125
	3	0.036	0.043	0.50	1.00	1.13	-0.105
	4	0.199	0.088	2.09	4.86	1.97	-0.164

(c) Train with the 6-month target

1.7 Conclusion

This paper applies three machine-learning approaches, specifically NNs, RFs, and XG-Boost. We transfer our research problem into the supervised classification problem. The equity and hybrid funds are examined separately.

One of the aims of the current study is to evaluate the predicted power of the fund characteristics in China. Our empirical results display that the past performance of funds can provide useful information in predicting future performance. In the out-of-sample tests, all of our fund characteristic-based top-class portfolios outperform the benchmarks. Our findings in the Chinese fund market complement the established empirical evidence for the U.S. of Wu et al. (2021). This result proves that the Chinese market is not fully efficient.

In this research, we further explore the influence of fund manager characteristics. The findings reported here suggest that fund manager characteristics do not play an essential role in predicting fund performance. The model accuracy of the fund manager features group is close to that of a random guess. Meanwhile, the financial performance does not significantly improve after adding in the fund manager information. Neither the fund managers' education, nor general information can provide sufficient information for a machine learning model. The model and financial performance suggest the fund manager's information adds noise rather than accuracy. The fund manager features we selected were unable to provide additional information, which may be caused by the fact that the data related to the fund manager is insufficient. Moving forward, researchers could potentially investigate additional characteristics of fund managers. One possible avenue includes utilising questionnaires to better comprehend the influence of fund managers' personality traits on decision-making, with the aim of predicting the performance of the fund in varying market scenarios.

Due to the limited extent of the data, the test period of this study is not extensive enough to cover a complete economic cycle. The performance of the machine learning model may further improve with the accumulation of additional data. The scope of the study also can be broadened by enhancing the feature engineering and selection process. Moreover, our methods represent only one aspect of the implementation of machine learning models in finance, as numerous other methods require exploration and development.

Chapter 2

Deep Momentum: Evidence from the developed countries UK, Japan and South Korea

ABSTRACT

Momentum stocks in the US market exhibit a bimodal distribution, wherein stocks that are initially 'winners' often become 'losers', and vice versa. This phenomenon is not unique to the United States, as we have also observed it in the stock markets of the UK, Japan, and South Korea. Furthermore, our findings provide support for the notion that the linear momentum model lacks significant predictive capabilities in Asian markets. In contrast, the machine learning-based deep momentum (DM) model effectively alleviates the bimodal nature of momentum and is able to generate substantial profits in the countries we studied. In fact, our research demonstrates that the momentum-based DM models significantly outperform both the traditional JT momentum strategy and a naïve machine learning model.

2.1 Introduction

Jegadeesh and Titman (1993) document the persistence and pervasiveness of momentum in the US market. They construct a long-short portfolio that buys stocks with the highest previous returns (winners) and sells stocks with the lowest previous returns (losers) and find that the portfolio can generate significant abnormal returns for the ensuing few months. Since its discovery, momentum has been one of the most persistent and puzzling anomalies in the financial

literature, over which researchers are not yet able to reach a consensus. The predictive ability of momentum has challenged the efficient market hypothesis and has been investigated in a wide range of countries (Fama and French, 2012; Rouwenhorst, 1998, 1999). It has also been explored in a diverse array of financial assets (Moskowitz et al., 2012; Gorton et al., 2013; Asness et al., 2013). Momentum remains significant in studies that examine the factor zoo. For instance, Green et al. (2017) examine 94 firm characteristics in the US market and find that price momentum is still significant. It is also found to play an important role in recent machine learning-based asset pricing studies (Madge and Bhatt, 2015; Peachavanish, 2016; Gu et al., 2020; Han, 2022).

The performance of momentum, however, is not stable, and this instability often occurs in times of market stress. Chordia and Shivakumar (2002) document that, while momentum profit is statistically significantly positive during economic expansions, it is insignificantly negative in economic recessions, suggesting that the momentum premium is pro-cyclical. Blitz et al. (2011) find similar evidence in that the return of a momentum strategy exhibits a substantial return during expansionary periods, while it suffers a negative return during recessionary periods. They also find that the losses during recessions are concentrated within the second half of the period when the market tends to revert. Daniel and Moskowitz (2016) document two major momentum crash periods in the US market — June 1932 to December 1939 and March 2009 to March 2013. During a crash, past losers gain more than past winners, and the gap can be as large as 200%. Since momentum crashes introduce huge risks and losses to investors, how to control such risk has posed considerable challenges for researchers. Daniel and Moskowitz (2016) propose an optimal dynamic momentum strategy based on the forecast of momentum's mean and variance, whereas Barroso and Santa-Clara (2015) suggest a constant volatility strategy, which controls the risk level of the momentum strategy.

Han (2022) documents the risk of the momentum strategy from a cross-sectional perspective and develops a machine learning-based return prediction model that addresses momentum crashes and significantly outperforms the traditional momentum strategy. He finds that momentum stocks have a U-shaped bimodal relative return distribution, which implies that, although past winners are most likely to belong to the highest-return decile in the future, their probability of belonging to the lowest-return decile is also significant. A similar observation can be made for past losers, which makes the momentum strategy fundamentally risky. He also finds that a naïve machine learning model cannot address bimodality and thus proposes a novel prediction model that can alleviate the risk of the bimodal distribution. He employs machine learning to estimate the relative return distribution and reclassify stocks based on their predicted financial performance such as the expected return, or Sharpe ratio. The empirical results show that the proposed model, deep momentum (DM), significantly outperforms the traditional momentum strategy as well as a naïve machine learning model in the US stock market, in which a value-weighted long-short portfolio can earn an annualised mean return of 35% and a Sharpe ratio of 1.66.

Inspired by the work of Han (2022), we investigate the existence of bimodality and test the DM strategy in three developed countries, specifically the UK, Japan, and South Korea. Given the time-consuming nature of machine learning models compared to linear regression, we have limited our research scope to only three countries in this article. Japan and South Korea have been selected as they are representative of the Asian market in which the traditional momentum strategy tends to underperform. The United Kingdom has been chosen as a comparable country, given that it has one of the largest stock markets and the traditional momentum strategy has historically exhibited a favourable performance in this market.

The aim of this paper is to provide answers to the following questions. First, we examine the existence of bimodality in other developed countries. Examining other countries can help us better understand this phenomenon. If the bimodal return distribution of momentum stocks also exists in other countries, it will imply that the fundamental risk of the traditional momentum strategy is not exclusively limited to the US stock market. Second, we aim to investigate the profitability of the DM strategy in other countries. By alleviating bimodality, DM exhibits significant predictive power and profitability in the US stock market. If bimodality is also present in other countries, then DM is likely to be able to alleviate it and also generate profits in those countries. Does DM perform well only in the US stock market, or can it be applied more broadly?

Our research makes several contributions in the following areas. Firstly, we document evidence indicating the existence of bimodality, not only in the United States but also in the three countries we investigate herein. However, this phenomenon has notably different characteristics in the three countries. Past winners exhibit bimodality in Japan and South Korea, while past losers exhibit bimodality in all three countries. One possible explanation for the underperformance of traditional momentum strategies in developed Asian nations could be attributed to this factor. Secondly, our research findings provide evidence that the DM model is successful in alleviating the bimodality of the high-return decile in Japan and South Korea. Further, our study demonstrates the exceptional performance of the DM model when applied to out-of-sample portfolio construction. The traditional momentum strategy yields annualised Sharpe ratios of 0.57, 0.00, and 0.15, respectively, in the UK, Japan, and South Korea. A naïve machine learning model outperforms the momentum strategy and yields Sharpe ratios of 0.94, 0.92, 0.68, and 0.98 in these countries. In contrast, a DM strategy that reclassifies stocks based on expected return further improves the performance and yields Sharpe ratios of 1.66, 1.11, and 1.64. When size dummies are added to the predictors, the Sharpe ratio increases to 1.97, 1.75, and 2.76. The return of the DM strategy cannot be explained by the Fama-French three factors or the momentum factor.

This chapter is organised as follows. Section 2 reviews the literature on the international evidence of momentum; Section 3 describes the methodology of Han (2022); Section 4 carries out empirical analyses; while Section 5 concludes.

2.2 International evidence of momentum

This section reviews the literature on momentum in the three countries we examine.

2.2.1 The United Kingdom

The UK stock market is one of the world's oldest stock markets. The London Stock Exchange (LSE) is the largest in Europe, ahead of Euronext, which has more than 2,000 companies currently listed. Rouwenhorst (1998) examines the momentum strategy using stocks from twelve European countries, including 494 stocks from the UK. Using the sample from 1980 to 1994, he finds that the momentum strategy yields a significantly positive return in the UK. Lui et al. (1999) find significant momentum profits in the UK during the period from 1977 to 1998, which again cannot be explained by the Fama-French three factors. Hon and Tonks (2003) also document evidence that supports this result. Agyei-Ampomah (2007) posit that the loser portfolio is heavily weighted toward small and illiquid stocks. when transaction costs are taken into account, momentum portfolios are still profitable for a longer horizon, but they can still incur losses for a short horizon. Siganos (2010) construct a momentum strategy that consists of extreme winners and losers and document statistically significant momentum profits from 1988 to 2006.

2.2.2 Japan

The evidence of momentum in Japan is mixed. (Asness, 2011) find it difficult to capture momentum premium in Japan. Chou et al. (2007) suggest that the Japanese market exhibits a contrarian effect rather than a momentum effect. Chui et al. (2010) are also unable to find a significant momentum premium for the period from 1981 to June 2003, attributing it to cultural differences, whereupon the Japanese are described as collectivists rather than individualists, lacking in confidence, and thus prices react to information more slowly. Fama and French (2012) test size, value, and momentum premiums in the international stock markets from 1989 to 2011 and do not find significant evidence for momentum in the Japanese market. Hanauer (2014) notes that momentum returns are higher, when the market is stable, than when it transitions to another state. The higher frequency of transitions in Japan leads to low momentum profits. Barroso and Santa-Clara (2015) demonstrate that the performance of momentum can be improved by managing its risk, although the Sharpe ratio remains low at 0.24.

2.2.3 South Korea

The literature on momentum in South Korea is scarce. Most studies do not observe a significant momentum profit. Chui et al. (2000) claim that momentum does not exist in South Korea during the period from 1997 to 2000. Although the portfolio acquires a positive momentum return, it is not statistically significant. They suggest this result may be

caused by the volatility of the momentum portfolio during the Asian Financial Crisis. Further, Chui et al. (2000) report a non-statistically significant negative momentum profit in South Korea from 1978 to 1997. Chae and Eom (2009) also find negative momentum portfolio returns from 1980 to 2005 by expanding on the decomposition method used by Lo and MacKinlay (1990). They show that the sum of the autocovariance is negative or zero, and that the sum of cross-serial covariances is positive. the observed negative momentum profit in the market can potentially be attributed to investors' under-reaction to market-wide information. Based on the sample period from 1990 to 2010, Park and Kim (2014) find that the stock momentum profit in South Korea is -0.40%.

Pyo and Yong (2013) suggest a positive relationship between momentum returns and idiosyncratic volatility. By strictly selecting the data, they acquired a positive momentum profit.

2.3 Methodology

We adopt Han (2022)'s DM return prediction model for our empirical analysis. DM is a classification model that predicts a stock's future return decile. DM consists of two steps. The first estimates the cross-sectional return distribution via a NN and the second reclassifies stocks based on their predicted financial performance. For the completeness of the paper, the DM model is described below.

2.3.1 Cross-sectional return distribution estimation

In the study conducted by Han (2022), two classifiers are utilised to estimate the cross-sectional return distribution, specifically a nominal classifier and an ordinal classifier. The nominal classifier is a conventional multi-class classifier based on NNs. It is worth noting that the nominal classifier is more commonly employed and has simpler and more standardised methods for activation, loss, and evaluation as compared to the ordinal classifier for NNs. Given these advantages, the nominal classifier is selected for the empirical analysis in this research.

Let $\boldsymbol{x} = \{x_1, \ldots, x_M\}$ denote M input features (explanatory variables), and c and $\boldsymbol{y}$ the target variable (output class) and its one-hot encoding, respectively. If the possible outcomes are $1, \ldots, K$, and the true class is k, then $c = k$ and $\boldsymbol{y}$ is a K-dimensional vector with the k-th element equal to 1 and the other elements equal to 0.

Given a dataset of N observations, $\{\boldsymbol{x}^i, \boldsymbol{y}^i\}_{i=1}^N$, a NN is trained so that the following cost function is minimised:

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} L\left(\boldsymbol{y}^i, \hat{\boldsymbol{y}}^i\left(\boldsymbol{x}^i; \boldsymbol{\theta}\right)\right), \tag{2.1}$$

where $\boldsymbol{\theta}$ is a set of model parameters, and $\hat{\boldsymbol{y}}(\boldsymbol{x};\boldsymbol{\theta})$ is the output of the neural network, whose k-th element is given by the softmax function:

$$\hat{y}_k(\boldsymbol{x};\boldsymbol{\theta}) = \frac{e^{z_k(\boldsymbol{x};\boldsymbol{\theta})}}{\sum_{k=1}^{K} e^{z_k(\boldsymbol{x};\boldsymbol{\theta})}}, \quad k = 1, \ldots, K, \tag{2.2}$$

for some function $z_k(\boldsymbol{x};\boldsymbol{\theta})$. The exact form of $z_k(\boldsymbol{x};\boldsymbol{\theta})$ is determined by the network architecture. Note that $\hat{y}_k(\boldsymbol{x};\boldsymbol{\theta})$ can be interpreted as the probability of a sample with features $\boldsymbol{x}$ belonging to class k, $P(c = k|\boldsymbol{x})$. The loss function is defined as the cross-entropy

$$L\left(\boldsymbol{y}, \hat{\boldsymbol{y}}(\boldsymbol{x};\boldsymbol{\theta})\right) = -\sum_{k=1}^{K} y_k \log \hat{y}_k(\boldsymbol{x};\boldsymbol{\theta}). \tag{2.3}$$

2.3.2 Reclassification

The standard method to predict the class of a stock is to choose the class with the maximum probability, *i.e.*, $\hat{c} = \mathrm{argmax}_k\, \hat{y}_k$. However, the class with the highest probability does not represent the stock's expected return when the distribution is bimodal, and the stocks classified into the top (bottom) class are unlikely to be those with the highest (lowest) expected returns. To address this issue, Han (2022) proposes five reclassification methods and finds two methods that reclassify stocks based on the predicted expected return and one using the Sharpe ratio which perform particularly well. We adopt these three methods.

Given the expected returns of the classes, $\mu_k,\ k = 1, \ldots, K$, the expected return of a stock can be obtained from the law of total expectation:

$$\mu = E[r] = E\left[E[r|c]\right] = \sum_{k=1}^{K} P(c = k)\mu_k. \tag{2.4}$$

The expected return can then be used to reclassify stocks. Since the mean returns of the classes are unknown, they need to be estimated. The first two reclassification methods take different approaches to estimate them.

2.3.2.0.1 Reclassification on Probability Difference (PrDf)

The first reclassification method assumes that the mean returns of the classes decrease linearly from top to bottom. Under this assumption, it can be shown that the mean return of a stock is proportional to:

$$PrDf = \sum_{k=1}^{K/2} (\hat{y}_k - \hat{y}_{K+1-k}) \left(\frac{K}{2} + 1 - k\right). \tag{2.5}$$

PrDf reclassifies stocks using this value.

2.3.2.0.2 Reclassification on Mean Return (Return)

The second method estimates the mean returns of the classes using the sample analogue over the past five years. Denoting the sample mean of class k $\hat{\mu}_k$, the estimate of a stock's mean return can be obtained from Equation (2.4):

$$\hat{\mu} = \sum_{k=1}^{K} \hat{y}_k \hat{\mu}_k. \tag{2.6}$$

2.3.2.0.3 Reclassification on Sharpe Ratio (Sharpe)

The last method reclassifies stocks based on the Sharpe ratio. Using the law of total variance, the variance of stock return is given by:

$$\sigma^2 = V[r] = E\left[V[r|c]\right] + V\left[E[r|c]\right] = \sum_{k=1}^{K} P(k)\left(\sigma_k^2 + \mu_k^2\right) - \mu^2, \tag{2.7}$$

where σ_k^2 is the variance of class k's return. Substituting μ_k and σ_k^2 with their sample analogues $\hat{\mu}_k$ and $\hat{\sigma}_k^2$, the variance of a stock's return can be estimated using the equation

$$\hat{\sigma}^2 = \sum_{k=1}^{K} \hat{y}_k \left(\hat{\sigma}_k^2 + \hat{\mu}_k^2\right) - \hat{\mu}^2. \tag{2.8}$$

The Sharpe ratio of the stock is then estimated by $\hat{SR} = \hat{\mu}/\hat{\sigma}$.

2.3.3 Input Features

Following Han (2022), we choose five momentum features as the input features of the NN. The m-month price momentum, $MOM_{m,i}$, is defined as the cumulative return for the past m months for each stock i in the sample, except for $m = 1$, which is the previous one-month return:

$$MOM_{m,i} = \prod_{j=t-m}^{t-2} (r_{j,i} + 1) - 1, \quad m = 3, 6, 9, 12, \tag{2.9}$$

$$MOM_{1,i} = r_{t-1,i}, \quad m = 1, \tag{2.10}$$

where $r_{j,i}$ denotes the return of stock i in month j. To eliminate the market trend, these are standardised cross-sectionally for each month using the formula

$$nMOM_{m,i} = \frac{MOM_{m,i} - M_{MOM_m}}{S_{MOM_m}}, \tag{2.11}$$

where M_{MOM_m} and S_{MOM_m} respectively denote the cross-sectional mean and standard deviation of MOM_m. The final feature set consists of the standardised momentum features, $nMOM_m$, and the cross-sectional means, M_{MOM_m}. The latter is included to take the macroeconomic status into account.

Han (2022) also includes a size feature and finds that the size feature can significantly improve the performance of the strategy. The size feature is defined as follows: In each month, stocks are divided into deciles based on the market capitalisation at the end of the previous month and assigned one of ten size dummies, $D_s,\ s = 1, \ldots, 10$. We use only momentum features for our primary empirical analysis as our objective is to compare the DM model with the traditional momentum strategy. The results that include size dummies are discussed after the primary results.

2.3.4 Neural network architecture and hyperparameter tuning

We test a different number of layers (between 3 and 10), with a different number of neurons (12, 24, 36, 64, 128) in each, and choose the architecture via cross-validation using data from before 2000. The optimal architecture turns out to be one with 3 or 4 layers and 36 or 64 neurons per layer, depending on the sample. Since the prediction performance is similar among these choices, we select an architecture with 3 layers and 64 neurons for all countries analysed. We use the rectified linear unit (ReLu) as the activation function of a hidden layer and the Adam optimiser. The number of epochs is determined via early stopping.

2.3.5 Model Evaluation

The model is evaluated using conventional classification performance metrics. For the overall classification performance, the average loss and accuracy are employed, as follows:

$$\text{Average Loss} = \frac{1}{N}\sum_{i=1}^{N} L\left(\boldsymbol{y}^i, \hat{\boldsymbol{y}}^i\right), \tag{2.12}$$

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{N}. \tag{2.13}$$

To assess the performance in each class, the following metrics are employed:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{2.14}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{2.15}$$

$$\text{F1-score} = \frac{2}{1/\text{Precision} + 1/\text{Recall}}, \tag{2.16}$$

where TP, FP, and FN respectively denote true positive, false positive, and false negative, respectively. Precision measures how many of the samples that were predicted to be true are actually true and are deemed to be more relevant to return prediction than recall. The omission of some stocks that are going to have a high or low return is not harmful, but misclassifying a low-return stock into a high-return class or vice versa can be disastrous.

The following error measures designed to account for ordinality are also employed:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(c^i - k)^2\hat{y}_k^i, \tag{2.17}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}|c^i - k|\hat{y}_k^i. \tag{2.18}$$

Both the mean squared error (MSE) and mean absolute error (MAE) have a minimum value of 0 when the predicted probability for the true class is 1, and thereafter increase as the probabilities for distant classes increase. These measures can also be defined at the class level: they are respectively denoted by the mean squared error of prediction (MSEP) and mean absolute error of prediction (MAEP) when stocks are grouped based on predicted classes, and by the mean squared error of residuals (MSER) and mean absolute error of residuals (MAER) when grouped based on true classes. MSEP and MAEP are similar to precision in the sense that they measure the prediction error within a predicted class, whereas MSER and MAER measure the prediction error within a true class and are similar to recall. Like precision, MSEP and MAEP are deemed to be more relevant to stock classification.

2.4 Empirical Analysis

2.4.1 Data

We collect from Datastream the monthly returns, prices, volumes, and market capitalisation of the stocks listed in the UK, Japan, and South Korea. The UK sample comprises all common shares listed on the LSE during the sample period from January 1965 to December 2019. The Japanese sample consists of all common shares listed on the Tokyo Stock Exchange and the Osaka Securities Exchange for the sample period from January 1973 to December 2019. Finally, the South Korean sample comprises all common shares listed on the South Korean Stock Exchange for the sample period from July 1996 to December 2019.[1] We include both listed and delisted stocks to avoid survivorship bias.

We set the test period to be from January 2000 to December 2019 for all countries and re-trained the model every year during the test period. For the UK and Japanese samples, we used the past 240 months of data to train the model in January 2000 and expanded the sample every year. For the South Korean samples, which do not have 20 years of historical data as of January 2000, we used all available historical data on the training date. The training sample is randomly split into a training set and a validation set in a ratio of 7:3. The validation set is a dataset consisting of examples used to assess the performance of a model. It is important to note that the validation set serves a different purpose from the test set. The latter is solely used to evaluate the final performance of the model after it has been

[1] During the composition of this article, it came to our attention that the Datastream database did not possess data pertaining to South Korea prior to 1995. To compensate for this limitation, previous researchers, such as Chui et al. (2000), resorted to gathering data from the PACAP database prior to the aforementioned year.

trained and tuned. Using a validation set can prevent overfitting the model to the training data. Overfitting occurs when the model performs exceptionally well on the training data but fails to generalise to unseen data. By incorporating a validation set into the training process, we can select the optimal model by minimising the validation error. This error serves as an approximation of the test error, providing an estimate of how the model will perform on unseen data.

2.4.2 Classification Performance

This section evaluates the classification performance of the DM model. Table 2.1 presents the overall classification performance in each market. The accuracy of the test set and validation set are comparable to the accuracy in the training set for all countries, which suggests that the model does not suffer from overfitting. The accuracy in the test is above 14% in all countries. As a reference, the test set accuracy in the US market reported in Han (2022) is 14.8%. These values are significantly higher than the accuracy of a random guess at 10%. The model performs particularly well in the UK, where the test set accuracy is 14.84 and the MSE is 14.63. As discussed later, the traditional momentum strategy also performs well in the UK, which implies that the momentum features have more predictive power relative to the other countries.

Table 2.1: Classification performance
This table reports the classification performance of DM. The evaluation metrics are defined in Section 2.3.5. The training set (Train) and validation set (Valid) metrics are the averages of the metrics obtained from yearly retraining, and the test set (Test) metrics are obtained from the test period, 2000.01 to 2019.12.

	Train				Valid				Test			
	Loss	Acc	MSE	MAE	Loss	Acc	MSE	MAE	Loss	Acc	MSE	MAE
UK	2.24	14.86	14.99	3.10	2.25	13.99	15.10	3.11	2.22	14.84	14.63	3.05
Japan	2.26	14.35	16.01	3.21	2.26	13.97	16.09	3.22	2.26	14.12	16.00	3.21
S. Korea	2.24	16.04	15.79	3.18	2.26	14.08	15.99	3.21	2.26	14.04	16.07	3.21

Table 2.2 compares the classification performances in the highest-return decile (H) and the lowest-return decile (L) before and after reclassification. Before reclassification (Org), the precision of both H and L is higher than the average (avg) in all countries, except for South Korea, where the precision of H is slightly lower than the average. The precision of L is particularly high, exceeding 20% in most countries. The table also shows that the model tends to classify more stocks into L and fewer into H. These results imply that predicting losers is easier, consistent with the empirical fact that the profit of the momentum strategy mainly results from the loser decile.

When stocks are reclassified, the precision of H tends to decrease, whereas the precision of L tends to increase. This result is partly because more (fewer) stocks are classified into H (L) after reclassification. In contrast, MSEP and MAEP, the error metrics that take ordinality into account, generally decrease in both classes after reclassification and the error reduction is particularly evident in H. For instance, the MSEP of H in the UK decreases from 21.25 to 11.00 after reclassification with the Sharpe ratio, which is in stark contrast to the precision that decreases from 16.63 to 5.45. The

reduction of both precision and MSEP implies that some stocks, which are correctly classified into H, are missed after reclassification, while many low-return stocks misclassified into H are replaced by higher-return stocks. From this result, we can expect that the financial performance of H is likely to improve more significantly than that of L after reclassification. Reclassification can effectively eliminate some stocks with opposite returns and group stocks with more similar returns. The opposite changes of precision and MSEP reveal that conventional evaluation metrics, such as precision, can be misleading and, therefore, making an investment decision based on them can result in unexpectedly poor financial performance.

Table 2.2: Reclassification performance

This table reports the class-level test set classification performance of DM before and after reclassification. 'Org' denotes DM before reclassification, and 'PrDf', 'Return', and 'Sharpe' denote reclassification based on probability difference; reclassification based on mean return; and reclassification based on the Sharpe ratio, respectively.

		Precision	Recall	f1-score	MSEP	MSER	MAEP	MAER	Support	Predicted	(%)
						UK					
Org	H	16.63	10.23	12.67	21.25	29.08	3.70	4.51	16456	10121	5.15
	L	20.41	45.12	28.11	22.17	28.45	3.75	4.34	14740	32578	16.58
	avg	14.11	15.42	13.67	14.93	15.51	3.09	3.13	196508	196508	100.00
PrDf	H	7.01	8.41	7.64	12.35	29.08	2.81	4.51	16456	19752	10.05
	L	20.59	27.28	23.47	21.63	28.45	3.69	4.34	14740	19532	9.94
	avg	12.50	12.86	12.61	14.64	15.51	3.05	3.13	196508	196508	100.00
Return	H	10.50	12.60	11.46	15.09	29.08	3.10	4.51	16456	19752	10.05
	L	19.37	25.67	22.08	20.94	28.45	3.64	4.34	14740	19532	9.94
	avg	12.84	13.21	12.95	14.64	15.51	3.05	3.13	196508	196508	100.00
Sharpe	H	5.45	6.54	5.95	11.00	29.08	2.66	4.51	16456	19752	10.05
	L	18.85	24.98	21.49	20.54	28.45	3.61	4.34	14740	19532	9.94
	avg	11.88	12.17	11.96	14.64	15.51	3.05	3.13	196508	196508	100.00
						Japan					
Org	H	15.63	15.38	15.51	19.75	29.28	3.59	4.52	63812	62805	9.78
	L	21.32	39.86	27.78	21.88	28.42	3.73	4.36	62354	116579	18.15
	avg	13.14	14.22	12.75	15.60	16.05	3.18	3.21	642235	642235	100.00
PrDf	H	9.50	9.58	9.54	14.41	29.28	3.05	4.52	63812	64330	10.02
	L	23.90	24.58	24.24	22.03	28.42	3.72	4.36	62354	64109	9.98
	avg	12.04	12.08	12.06	16.01	16.05	3.21	3.21	642235	642235	100.00
Return	H	12.12	12.22	12.17	16.71	29.28	3.28	4.52	63812	64330	10.02
	L	22.20	22.82	22.50	21.18	28.42	3.65	4.36	62354	64109	9.98
	avg	12.29	12.32	12.31	16.01	16.05	3.21	3.21	642235	642235	100.00
Sharpe	H	9.11	9.18	9.15	14.07	29.28	3.01	4.52	63812	64330	10.02
	L	21.31	21.91	21.61	20.75	28.42	3.61	4.36	62354	64109	9.98
	avg	11.65	11.68	11.66	16.01	16.05	3.21	3.21	642235	642235	100.00
						South Korea					
Org	H	12.74	10.44	11.48	18.66	29.58	3.49	4.56	17370	14239	8.11
	L	20.83	46.17	28.70	21.44	27.67	3.69	4.24	16865	37384	21.30
	avg	12.92	14.17	12.78	15.61	16.15	3.18	3.22	175480	175480	100.00
PrDf	H	7.98	8.11	8.04	12.85	29.58	2.88	4.56	17370	17659	10.06
	L	25.48	26.34	25.90	22.25	27.67	3.72	4.24	16865	17439	9.94
	avg	12.04	12.09	12.06	16.08	16.15	3.22	3.22	175480	175480	100.00
Return	H	11.66	11.85	11.76	16.79	29.58	3.30	4.56	17370	17659	10.06
	L	24.14	24.96	24.54	21.58	27.67	3.67	4.24	16865	17439	9.94
	avg	12.63	12.68	12.65	16.07	16.15	3.21	3.22	175480	175480	100.00
Sharpe	H	7.98	8.11	8.04	12.74	29.58	2.87	4.56	17370	17659	10.06
	L	24.45	25.28	24.86	21.69	27.67	3.67	4.24	16865	17439	9.94
	avg	11.93	11.97	11.95	16.08	16.15	3.22	3.22	175480	175480	100.00

To further examine the effect of reclassification, we draw the predicted probabilities and realised distributions of the stocks in H and L in Figure 2.1. The orange bars represent predicted probabilities while the blue bars represent the actual distributions of the stocks. For instance, a bar chart labelled 'Org (H)' presents the average predicted probability

of the stocks that are classified into H for each class and their actual distribution. From the results before reclassification, it is evident that the stocks in H and L exhibit bimodal distributions. The stocks classified into H (L) have a high probability for L (H), and many of these stocks actually generate low (high) returns and end up in L (H). The predicted and actual distributions are surprisingly similar, which suggests that stocks can be better classified from a financial perspective by exploiting the distribution rather than choosing the class with the highest probability.

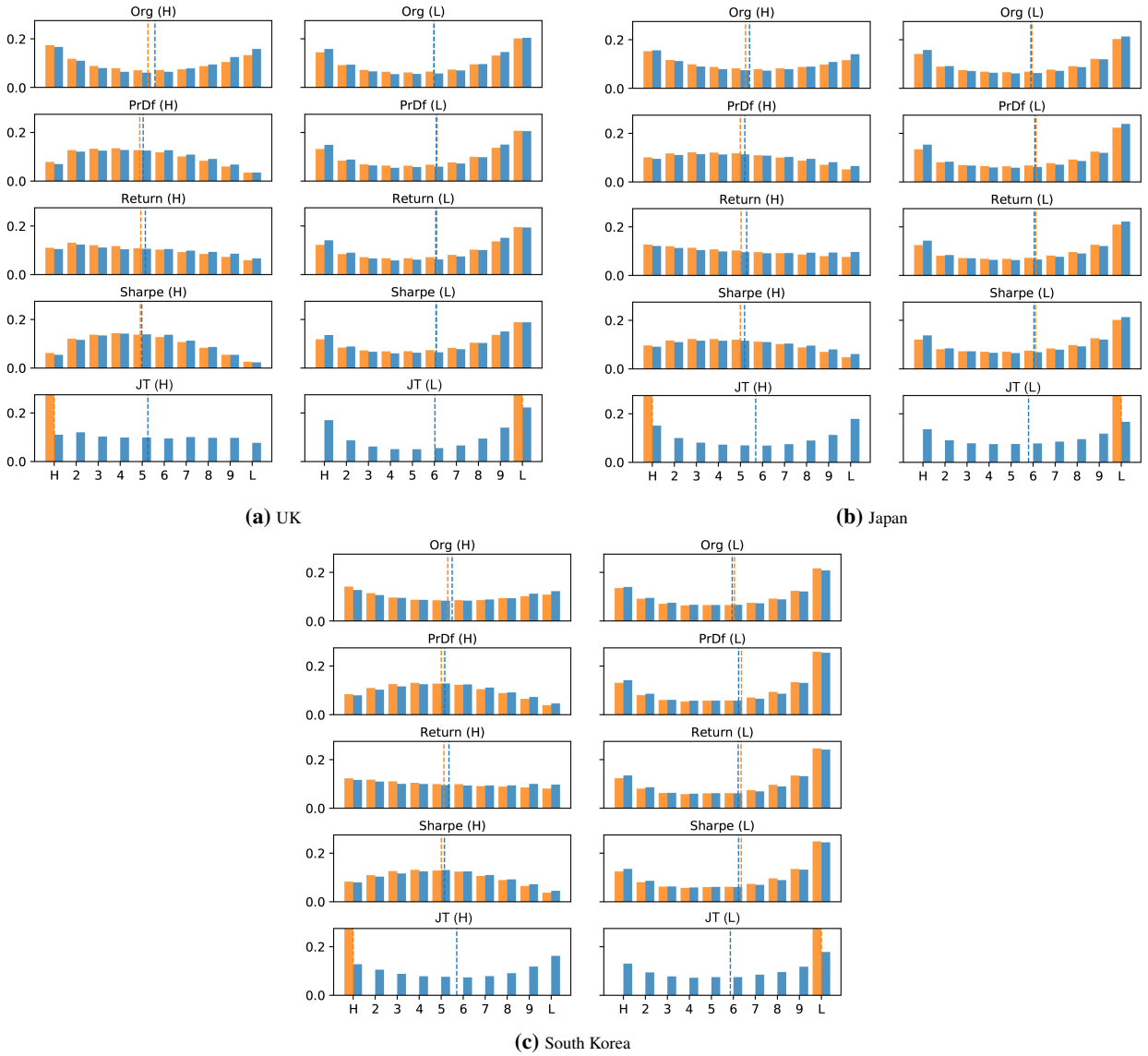
After stocks are reclassified, we no longer observe bimodality in H. Some correctly classified stocks are missed, although, at the same time, many stocks with a high probability for low return are eliminated. As a result, both predicted and actual distributions have a right-skewed bell shape or else decrease monotonically towards low-return classes. Reclassification can also alleviate the bimodality observed in L, yet cannot eliminate it completely. Overall, these results suggest that stocks with a high probability of crashing also have a high probability of jumping, and it is difficult to distinguish crashing stocks from jumping stocks. These findings are similar to those from the US as reported by Han (2022).

The distributions of the traditional momentum strategy (JT) reveal that many winners turn out to yield the lowest returns, while many losers yield the highest returns. In fact, in Japan and South Korea, more stocks in H end up in L than H. The only exception is the UK, where the stocks in H do not exhibit bimodality.

2.4.3 Financial Performance

Table 2.3 reports the mean returns and the Sharpe ratios of equal-weighted and value-weighted portfolios. Overall, the DM model significantly outperforms the traditional momentum strategy in all countries, even before reclassification. When stocks are equally weighted, the mean return increases by as much as 28%, while the Sharpe ratio increases to 0.92. Reclassification further improves the performance by adding as much as 20% to the mean return and 1.21 to the Sharpe ratio. When stocks are value-weighted, the improvements are less pronounced, though still significant. Reclassifying stocks on Return doubles the mean return and the Sharpe ratio in some countries.

In the UK, the momentum strategy performs well with an annualised mean return of 18% and a Sharpe ratio of 0.57. The long portfolio performs particularly well and renders an annualised mean return of 16% and a Sharpe ratio of 0.84. This result has been anticipated from Figure 2.1, which shows that the winners do not have a bimodal distribution. Nevertheless, DM outperforms JT significantly - even before reclassification - yielding an annualised mean return of 20% and a Sharpe ratio of 0.94. The greater increase in the Sharpe ratio suggests that the long-short portfolio constructed by DM is less volatile. Reclassifying stocks on Return further improves the performance and increases both the mean return and the Sharpe ratio to 36% and 1.66, respectively. The other reclassification methods also improve the performance to a similar level. The value-weight portfolio however performs worse. DM yields a mean return of 18% and a Sharpe ratio of 0.53 when stocks are reclassified based on Return. In contrast, JT is not affected much by the



(a) UK

(b) Japan

(c) South Korea

Figure 2.1: Cross-sectional distribution

This figure presents the cross-sectional relative return distributions in the high-return (H) and low-return (L) deciles. The orange bars represent the average predicted probabilities of the stocks and the blue bars represent their realised frequencies. The dotted lines represent the average return (*for each?*) class. The title of each graph refers to the reclassification method and the horizontal axis denotes the class of the one-month-ahead return.

weighting scheme and yields a mean return of 19% and a Sharpe ratio of 0.43. The relatively poor performance of the value-weighted portfolio is consistent with findings in the US (Han, 2022) and can be partially attributed to the training process for the NN, in which each stock receives the same weighting in terms of the cost function, regardless of its size. As discussed in the next section, adding size dummies significantly improves the performance of both equal-weight and value-weighted portfolios.

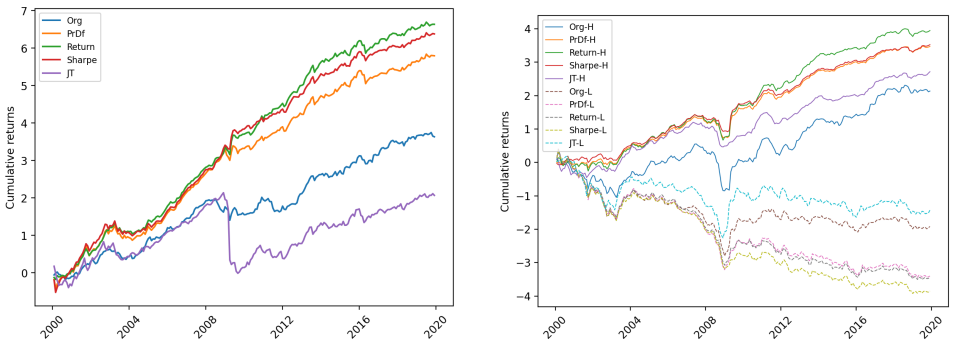
The literature suggests that the momentum strategy shows weak performance in East Asian markets. Our results are consistent with the previous findings. The Sharpe ratios for the momentum strategy in Japan and South Korea are close to zero. Although one-year momentum appears to have little predictive power, momentum features of different periods can have significant predictive power when combined together via a NN. In Japan, DM (Org) is able to generate an annualised mean return of 9% and a Sharpe ratio of 0.92, while DM (Return) yields a mean return of 17% and a Sharpe ratio of 1.11. These values are significantly higher than those for JT, which are 0% and 0.00, respectively. In South Korea, JT yields a mean return of 5% and a Sharpe ratio of 0.15. In contrast, DM (Org) yields a mean return of 33% and a Sharpe ratio of 0.98, while DM (Return) yields a mean return of 49% and a Sharpe ratio of 1.64.

Unlike the US and the UK, JT performs better in Asian countries when stocks are value-weighted. The mean returns and the Sharpe ratios for JT are 2% and 0.09 in Japan, and 3% and 6% and 0.19 in South Korea, respectively. DM performs worse when stocks are value-weighted, although it still outperforms JT. The mean returns and the Sharpe ratios of DM (Return) are 13% and 0.72 in Japan, and 41% and 1.20 in South Korea, respectively. Of the two countries, DM performs best in South Korea in terms of mean return.

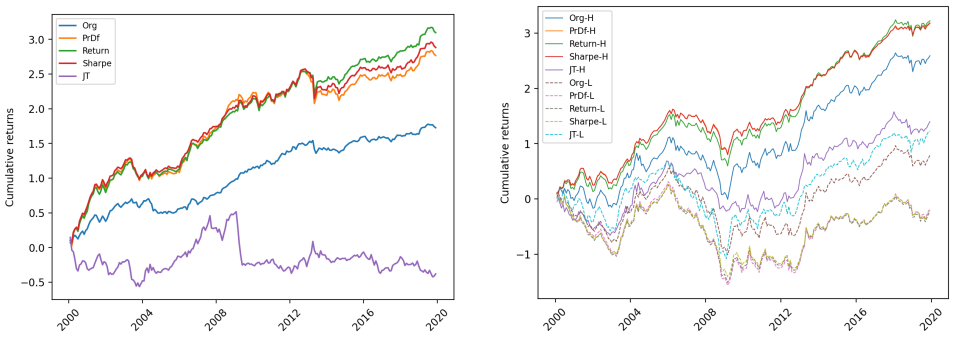
Figure 2.2 presents the cumulative returns of the long-short portfolios and their legs. The figure highlights the effectiveness of reclassification. While DM outperforms JT before reclassification, it still suffers from significant drawdowns at times. In contrast, the long-short portfolio does not suffer any significant loss when stocks are reclassified. It is also notable that the return of the long-short portfolio grows steadily throughout the test period for all countries when stocks are reclassified.

2.4.4 Factor Regression

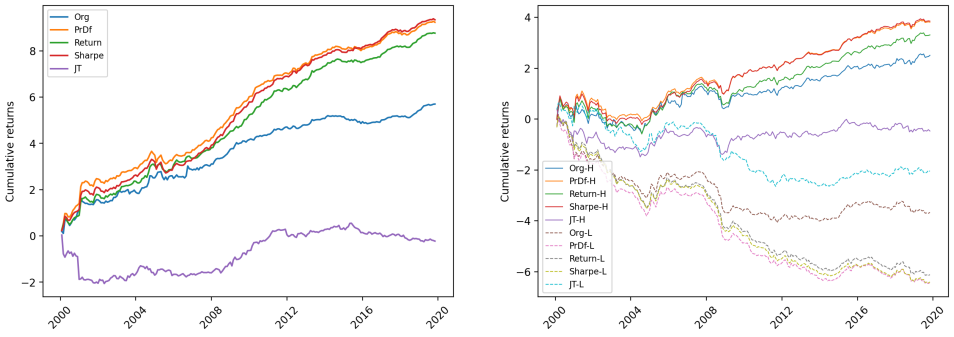
To examine whether the return of the DM long-short portfolio can be explained in terms of factors, we run a factor regression using the Fama-French three factors plus the momentum factor. We use the Fama-French three factors rather than the more recent five factors as the other two factors are difficult to obtain for other countries. The UK factors are derived from a data source at Exeter University, whereas the Japanese factors were obtained from the K. French website, and the Korean factors are courtesy of Handa partners. The sample period covered runs from 2000 to 2019, except for the UK, where the sample period ends in 2017 due to limited data availability. The portfolios constructed by DM (Return) are used for the regression analysis.



(a) UK



(b) Japan



(c) South Korea

Figure 2.2: Cumulative Returns (equal-weighted)

This figure presents the cumulative returns (logarithmic scale) of the equal-weighted long-short portfolios constructed by DM and the traditional momentum strategy (JT). The left panels present the long-short portfolios and the right panels present long (solid line) and short (dashed line) portfolios separately.

Table 2.3: Financial performance
This table reports the annualised mean returns and Sharpe ratios of the long (H), short (L), and long-short (H-L) portfolios as constructed using the DM model and the traditional momentum strategy (JT). The returns are calculated over the test period from 2000.01 to 2019.12.

	Mean return					Sharpe ratio				
	Org	PrDf	Return	Sharpe	JT	Org	PrDf	Return	Sharpe	JT
					UK					
H	0.16	0.20	0.23	0.20	0.16	0.48	0.98	0.92	1.07	0.84
L	-0.05	-0.12	-0.13	-0.15	-0.01	-0.15	-0.42	-0.48	-0.58	-0.04
H-L	0.20	0.32	0.36	0.35	0.18	0.94	1.42	1.66	1.56	0.57
					Japan					
H	0.17	0.18	0.19	0.18	0.10	0.63	0.94	0.84	0.96	0.41
L	0.08	0.03	0.02	0.02	0.10	0.29	0.11	0.09	0.09	0.37
H-L	0.09	0.15	0.17	0.16	0.00	0.92	0.93	1.11	1.03	0.00
					South Korea					
H	0.22	0.27	0.25	0.27	0.03	0.47	0.60	0.57	0.63	0.09
L	-0.12	-0.25	-0.24	-0.25	-0.02	-0.32	-0.69	-0.67	-0.70	-0.04
H-L	0.33	0.52	0.49	0.52	0.05	0.98	1.46	1.64	1.64	0.15

(a) Equal-weighted portfolios

	Mean return					Sharpe ratio				
	Org	PrDf	Return	Sharpe	JT	Org	PrDf	Return	Sharpe	JT
					UK					
H	0.12	0.09	0.14	0.09	0.10	0.32	0.42	0.57	0.49	0.47
L	-0.06	-0.09	-0.04	-0.07	-0.08	-0.15	-0.24	-0.12	-0.21	-0.19
H-L	0.18	0.18	0.18	0.16	0.19	0.53	0.54	0.53	0.49	0.43
					Japan					
H	0.09	0.12	0.11	0.12	0.05	0.32	0.55	0.49	0.55	0.22
L	0.03	-0.02	-0.02	-0.02	0.03	0.12	-0.07	-0.07	-0.08	0.10
H-L	0.06	0.14	0.13	0.14	0.02	0.36	0.70	0.72	0.73	0.09
					South Korea					
H	0.06	0.11	0.11	0.08	-0.07	0.13	0.24	0.23	0.17	-0.17
L	-0.17	-0.28	-0.31	-0.30	-0.13	-0.46	-0.70	-0.79	-0.79	-0.27
H-L	0.23	0.39	0.41	0.39	0.06	0.64	0.93	1.20	1.02	0.19

(b) Value-weighted portfolios

Table 2.4 reports the regression results. When stocks are equally weighted, the long-short portfolio alpha is statistically and economically significant, with the t-statistic greater than 6.0 for all countries, except for Japan, where the t-statistic is 4.38. When stocks are value-weighted, the t-statistic of alpha is greater than 4.0 in South Korea, 2.70 in Japan, and 1.95 in the UK. As discussed later, when size dummies are included in the input feature set, the alpha of the value-weighted long-short portfolio becomes statistically and economically significant with the t-statistic greater than 3.0 in all countries. This result suggests that the return of the DM long-short portfolio cannot be explained by the Fama-French three factors or the momentum factor.

Table 2.4: Factor regression

This table reports the factor regression results of the long (H), short (L), and long-short (H-L) portfolios constructed by DM (Return). The portfolio returns are regressed via the Fama-French three factors (Mkt-Rf, SMB, HML) plus the momentum factor (MOM). For each portfolio, the first row reports the coefficients, while the second row reports the Newey-West adjusted t-statistics. The sample period runs from 2000 to 2019, except for the UK, for which the sample period ends in 2017 due to limited data availability.

		α	Mkt-Rf	SMB	HML	MOM	α	Mkt-Rf	SMB	HML	MOM
		Equal-weighted					Value-weighted				
UK	H	0.019	0.074	0.330	-0.181	-0.051	0.010	0.146	0.341	-0.040	-0.036
		3.54	0.54	2.16	-1.03	-0.45	1.87	1.09	2.27	-0.23	-0.33
	L	-0.012	0.109	0.590	-0.268	-0.206	-0.007	0.361	0.655	-0.053	-0.178
		-2.14	0.75	3.63	-1.43	-1.70	-0.92	2.01	3.24	-0.23	-1.18
	H-L	0.029	-0.027	-0.258	0.077	0.153	0.014	-0.208	-0.311	0.003	0.139
		6.36	-0.23	-1.97	0.51	1.58	1.95	-1.11	-1.48	0.01	0.89
Japan	H	0.014	0.136	0.330	-0.098	-0.075	0.008	0.195	0.257	-0.069	-0.120
		3.17	1.43	1.94	-0.64	-1.01	1.69	2.00	1.47	-0.44	-1.57
	L	0.000	0.099	0.469	-0.175	-0.163	-0.003	0.181	0.444	-0.223	-0.133
		-0.06	0.92	2.45	-1.01	-1.94	-0.73	1.78	2.45	-1.36	-1.68
	H-L	0.013	0.042	-0.132	0.067	0.085	0.010	0.019	-0.180	0.144	0.011
		4.38	0.65	-1.16	0.65	1.71	2.70	0.25	-1.29	1.15	0.18
S. Korea	H	0.015	1.566	1.648	-0.576	-0.160	0.003	1.673	1.337	-0.532	-0.136
		2.78	14.82	11.52	-3.65	-0.82	0.48	13.54	7.99	-2.88	-0.60
	L	-0.026	1.312	1.384	-0.573	-0.169	-0.031	1.344	1.309	-0.678	-0.206
		-5.63	15.02	11.69	-4.39	-1.05	-5.71	13.05	9.39	-4.41	-1.09
	H-L	0.038	0.256	0.263	-0.004	0.007	0.031	0.332	0.026	0.146	0.067
		6.67	2.34	1.77	-0.02	0.04	4.76	2.68	0.15	0.79	0.30

2.4.5 Inclusion of size dummies

When size dummies are included in the input feature set, all evaluation metrics in the test set improve in most countries as shown in Table 2.5. The effects of the size dummies become more evident when we compare financial performance. Table 2.6 reports the annualised mean returns and the Sharpe ratios of the portfolios constructed by DM, including size dummies. In the UK, the mean return of the long-short portfolio increases from 20% to 24% and the Sharpe ratio increases from 0.94 to 1.01 before reclassification. When stocks are reclassified on Return, these values increase from 36% to 38% and from 1.66 to 1.97, respectively. In Japan, the mean return of the long-short portfolio increases from 9% to 15% and the Sharpe ratio increases from 0.92 to 1.56 before reclassification. When stocks are reclassified on Return, these values increase from 17% to 24% and from 1.11 to 1.75, respectively.

In South Korea, the mean return of the long-short portfolio increases from 33% to 42% and the Sharpe ratio increases from 0.98 to 1.70 before reclassification. When stocks are reclassified on Return, these values increase from 49% to 78% and from 1.64 to 2.76, respectively. The performance of DM is remarkable, especially when we consider that the momentum strategy performs poorly in Asian countries.

Size dummies improve the performance of the value-weighted portfolios as well, especially in the UK and South Korea. When stocks are reclassified on Return, the mean return and the Sharpe ratio increase from 18% to 32% and from 0.53 to 1.09 in the UK; from 13% to 15% and from 0.72 to 0.78 in Japan; and from 41% to 55% and from 1.20 to

1.88 in South Korea, respectively. It is notable that size dummies do not always add value in DM when stocks are not reclassified. In the UK, the Sharpe ratio increases only slightly from 0.53 to 0.66. This result highlights the importance of reclassifying stocks based on their predicted financial performance.

Table 2.5: Classification performance: including size
This table reports the classification performance of DM including size dummies. The evaluation metrics are defined in Section 2.3.5. The training set (Train) and validation set (Valid) metrics are the averages of the metrics obtained from yearly retraining, while the test set (Test) metrics are obtained from the test period, 2000.01 to 2019.12.

	Train				Valid				Test			
	Loss	Acc	MSE	MAE	Loss	Acc	MSE	MAE	Loss	Acc	MSE	MAE
UK	2.23	15.41	14.83	3.07	2.24	14.41	14.99	3.09	2.22	15.02	14.59	3.04
Japan	2.24	15.26	15.72	3.17	2.26	14.08	16.01	3.21	2.26	14.22	15.96	3.20
S. Korea	2.23	16.30	15.70	3.17	2.26	14.05	15.84	3.19	2.26	14.16	15.99	3.20

The factor regression results presented in Table 2.7 reveal that the returns of the long-short portfolios cannot be explained solely by these factors and confirm the important role of size dummies. The t-statistic of the long-short portfolio alpha increases in all countries. When stocks are equally weighted, the t-statistic is greater than 6 and as high as 11.71 (South Korea). In contrast, when stocks are value-weighted, it is greater than 3 and as high as 7.81 (South Korea).

The significant difference in performance arising between equal-weighted portfolios and value-weighted portfolios, and the effect of size dummies suggests that the predictive power of the momentum features varies across the size of the firm. It is also well known that a large part of the performance of the momentum strategy is derived from small stocks.

Banz (1981) first document the size effects in the US market. The smaller firms tend to have higher risk-adjusted returns than larger firms. Fama and French (1992) suggest that the size factor and book-to-market ratio provide a powerful characterisation of the cross-section of average stock returns. In their future work (Fama and French, 2012), they also find supportive evidence for the size effect in the international equity market.

2.4.6 Performance during financial crisis

The momentum strategy is known to crash in the US market when the market recovers from the financial crisis in 2008. To examine the performance of DM and JT more closely around the financial crisis, we report in Table 2.8 their performance before, during, and after the financial crisis. In the table, DM refers to the DM model and includes size dummies and reclassified on Return. As a benchmark, we also include the following market indices: FTSE100 for the UK, Nikkei225 for Japan, and KOSPI for South Korea. Figure 2.3 presents cumulative returns around the period of the financial crisis.

During the financial crisis, the momentum strategy performs poorly in the UK and Japan. It renders a negative return in these countries and underperforms the market index in the UK. The only exception is South Korea, where the momentum strategy performs well with a Sharpe ratio of 1.031. Figure 2.3 reveals that the momentum strategy behaves

Table 2.6: Financial performance: including size

This table reports the annualised mean returns and Sharpe ratios of the long (H), short (L), and long-short (H-L) portfolios constructed by the DM model including size dummies and the traditional momentum strategy (JT). The returns are calculated over the test period from 2000.01 to 2019.12.

	Mean return					Sharpe ratio				
	Org	PrDf	Return	Sharpe	JT	Org	PrDf	Return	Sharpe	JT
					UK					
H	0.16	0.19	0.22	0.19	0.16	0.47	1.01	0.91	1.23	0.84
L	-0.08	-0.15	-0.16	-0.15	-0.01	-0.26	-0.53	-0.60	-0.62	-0.04
H-L	0.24	0.34	0.38	0.34	0.18	1.01	1.66	1.97	1.79	0.57
					Japan					
H	0.23	0.23	0.25	0.23	0.10	0.93	1.14	1.11	1.16	0.41
L	0.09	0.01	0.00	-0.00	0.10	0.32	0.05	0.02	-0.01	0.37
H-L	0.15	0.21	0.24	0.23	0.00	1.56	1.43	1.75	1.63	0.00
					South Korea					
H	0.29	0.52	0.50	0.54	0.03	0.73	1.15	1.24	1.10	0.09
L	-0.13	-0.26	-0.28	-0.26	-0.02	-0.36	-0.73	-0.83	-0.73	-0.04
H-L	0.42	0.78	0.78	0.79	0.05	1.70	2.07	2.76	2.01	0.15

(a) Equal-weighted portfolios

	Mean return					Sharpe ratio				
	Org	PrDf	Return	Sharpe	JT	Org	PrDf	Return	Sharpe	JT
					UK					
H	0.10	0.14	0.20	0.15	0.10	0.24	0.80	0.84	1.02	0.47
L	-0.13	-0.15	-0.12	-0.12	-0.08	-0.35	-0.45	-0.39	-0.43	-0.19
H-L	0.23	0.30	0.32	0.27	0.19	0.66	0.91	1.09	1.11	0.43
					Japan					
H	0.13	0.11	0.12	0.13	0.05	0.51	0.50	0.49	0.58	0.22
L	0.03	-0.03	-0.03	-0.04	0.03	0.10	-0.13	-0.12	-0.16	0.10
H-L	0.10	0.14	0.15	0.16	0.02	0.62	0.71	0.78	0.85	0.09
					South Korea					
H	0.18	0.31	0.30	0.32	-0.07	0.44	0.79	0.73	0.82	-0.17
L	-0.14	-0.26	-0.25	-0.22	-0.13	-0.35	-0.73	-0.72	-0.64	-0.27
H-L	0.32	0.57	0.55	0.54	0.06	1.06	2.00	1.88	2.09	0.19

(b) Value-weighted portfolios

very differently in the three countries during the financial crisis. It performs well in the UK and Japan during the crisis but crashes sharply when the market starts to rebound. However, in South Korea, it performs steadily without being affected by the financial crisis.

In contrast, the performance of DM during the financial crisis is remarkable. It performs consistently throughout all three periods and performs even better during the crisis in Japan and South Korea. In South Korea, DM achieves a Sharpe ratio of 4.53 for the period of the financial crisis. DM is able to identify losers very well during the crisis and the following recovery period. Figure 2.3 confirms that DM does not suffer from any significant drawdowns during the financial crisis. Han (2022) also finds that DM performs better during the crisis in the US.

Table 2.7: Factor regression: including size

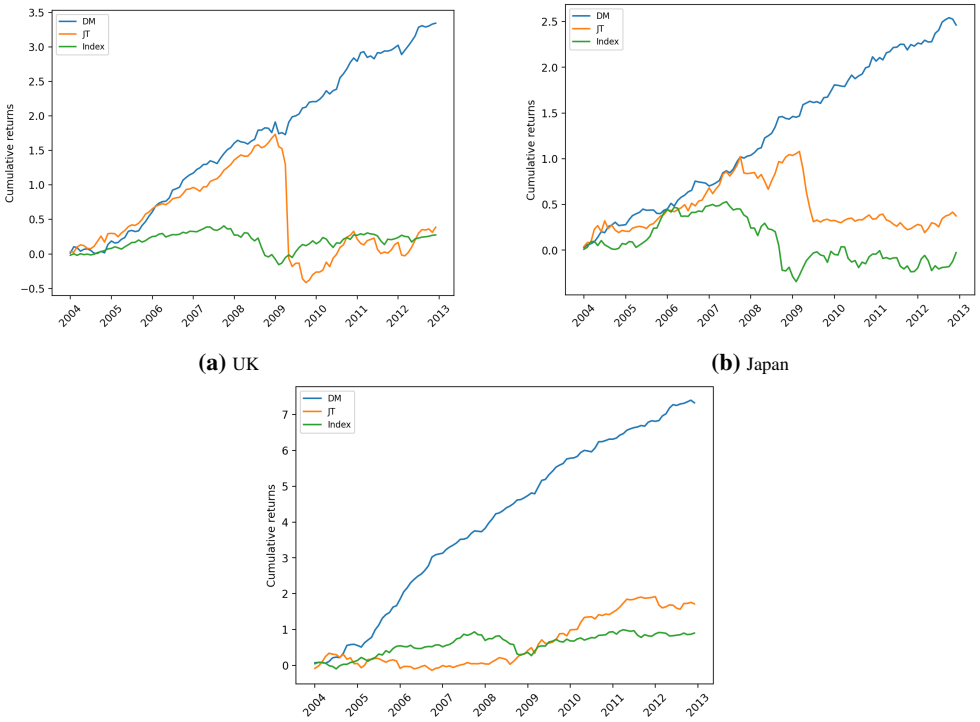
This table reports the factor regression results for the long (H), short (L), and long-short (H-L) portfolios constructed by DM (Return) including size dummies. The portfolio returns are regressed via the Fama-French three factors (Mkt-Rf, SMB, HML) plus the momentum factor (MOM). For each portfolio, the first row reports the coefficients and the second row reports the Newey-West adjusted t-statistics. The sample period is from 2000 to 2019, except for the UK, for which the sample period ends in 2017 due to limited data availability.

		α	Mkt-Rf	SMB	HML	MOM	α	Mkt-Rf	SMB	HML	MOM
		Equal-weighted					Value-weighted				
UK	H	0.018	0.064	0.431	-0.170	-0.030	0.014	0.207	0.591	-0.212	-0.042
		3.42	0.48	2.88	-0.99	-0.27	3.00	1.70	4.32	-1.35	-0.41
	L	-0.014	0.108	0.499	-0.203	-0.226	-0.009	0.226	0.647	-0.276	-0.366
		-2.60	0.79	3.23	-1.15	-1.97	-1.44	1.45	3.68	-1.37	-2.80
	H-L	0.030	-0.037	-0.066	0.023	0.194	0.021	-0.012	-0.054	0.053	0.322
		7.35	-0.35	-0.56	0.17	2.24	3.47	-0.07	-0.31	0.27	2.47
Japan	H	0.019	0.081	0.319	-0.190	-0.083	0.008	0.144	0.567	-0.193	-0.136
		4.45	0.87	1.90	-1.26	-1.13	1.65	1.42	3.12	-1.18	-1.71
	L	-0.002	0.125	0.427	-0.118	-0.165	-0.005	0.247	0.326	0.053	-0.076
		-0.40	1.22	2.33	-0.72	-2.06	-1.15	2.36	1.73	0.31	-0.92
	H-L	0.019	-0.039	-0.101	-0.081	0.080	0.012	-0.099	0.248	-0.255	-0.063
		7.35	-0.67	-0.96	-0.86	1.75	3.16	-1.20	1.69	-1.93	-0.98
S. Korea	H	0.037	1.434	1.602	-0.657	-0.337	0.020	1.587	1.421	-0.880	-0.236
		7.36	14.97	12.34	-4.59	-1.91	3.68	15.10	9.98	-5.60	-1.22
	L	-0.029	1.341	1.304	-0.579	-0.133	-0.026	1.382	1.067	-0.534	-0.151
		-7.34	17.54	12.59	-5.06	-0.95	-6.28	17.34	9.88	-4.48	-1.03
	H-L	0.063	0.096	0.297	-0.079	-0.206	0.043	0.208	0.353	-0.347	-0.087
		11.71	0.93	2.12	-0.51	-1.08	7.81	1.95	2.45	-2.18	-0.45

Table 2.8: Performance around the global financial crisis

This table reports the annualised mean returns (Mean), standard deviations (Std), and Sharpe ratios (Sharpe) of the long-short portfolios constructed by DM (Return) including size dummies and the traditional momentum strategy (JT) around the Global Financial Crisis. The benchmark indexes are FTSE100 for the UK, Nikkei225 for Japan, and KOSPI for South Korea.

	2004.01 - 2006.12			2007.01 - 2009.12			2010.01 - 2012.12		
	Mean	Std	Sharpe	Mean	Std	Sharpe	Mean	Std	Sharpe
					UK				
DM	0.400	0.545	2.261	0.381	0.752	1.600	0.405	0.720	1.931
JT	0.328	0.491	1.996	-0.174	1.746	-0.413	0.262	0.865	1.033
Index	0.113	0.254	0.930	-0.030	0.627	-0.347	0.038	0.474	0.244
					Japan				
DM	0.253	0.377	2.041	0.349	0.480	2.364	0.255	0.527	1.670
JT	0.221	0.584	1.134	-0.052	1.035	-0.244	0.025	0.429	0.197
Index	0.170	0.491	0.996	-0.131	0.853	-0.616	0.014	0.674	0.067
					South Korea				
DM	1.119	0.996	3.757	0.934	0.677	4.531	0.551	0.680	2.651
JT	0.017	0.976	-0.078	0.338	0.986	1.031	0.340	0.988	1.083
Index	0.208	0.631	0.925	0.091	0.946	0.173	0.071	0.573	0.240



(a) UK

(b) Japan

(c) South Korea

Figure 2.3: Cumulative Returns around the global financial crisis

This figure presents the cumulative returns (logarithmic scale) of the equal-weighted long-short portfolios constructed by DM (Return) including size dummies (DM) and the traditional momentum strategy (JT), and a market index (Index) around the time of the Global Financial Crisis. The benchmark indexes are FTSE100 for the UK, Nikkei225 for Japan, and KOSPI for South Korea.

2.5 Conclusion

Han (2022) documents that momentum stocks in the US market exhibit a bimodal distribution of relative returns. This suggests that winners are likely to become losers in the future, and vice versa. To address this bimodality, Han developed a novel machine learning return prediction model known as deep momentum, which was found to significantly outperform the traditional momentum strategy.

In order to investigate the presence of bimodality and the profitability of deep momentum in a wider context, we examined stocks from the UK, Japan, and South Korea. Our findings reveal that bimodality is also observed in the UK, Japan, and South Korea. Moreover, deep momentum proves to be an effective means of addressing the bimodality of momentum. The portfolios based on the deep momentum model demonstrate significant outperformance as compared to the traditional momentum strategy, with Sharpe ratios of 1.97, 1.75, and 2.76 for the UK, Japan, and South Korea, respectively.

One avenue for future research would involve extending this study to other countries and asset classes. Additionally, it is important to note that our study did not consider transaction costs. Future researchers should take this factor into account when training machine learning models and conducting further evaluations.

Chapter 3

Exploring the Non-linearity of Momentum in Chinese Stocks with Machine Learning

ABSTRACT

This chapter examines the effectiveness of the momentum in Chinese stock markets from both linear and non-linear perspectives. Specifically, we compare the performance of the traditional linear JT model, XG-Boost model, neural network models and neural network reclassification models as developed by Han (2022). The article focuses on comparing the performance of long-only and long-short portfolios due to the strict short-selling restrictions of the Chinese market. Our findings demonstrate that machine learning models based on momentum features outperform the traditional JT linear regression model. This indicates that a non-linear relationship exists between momentum features and stock returns in China, and thus the momentum features may prove to be a valuable source of information for stock price forecasting. Notably, the Han reclassification models exhibit the most favourable financial performance. This could be due to the reduction of bimodality in the true target distribution within high-return deciles. Additionally, we observe a significant positive correlation between the return of the long-only portfolio, as developed using the momentum feature in the machine learning framework, and both the size and degree of change in the sentiment index. Despite being regressed with the Fama-French three-factor and change in sentiment index, our machine-learning-based portfolios still achieve a significant positive alpha.

3.1 Introduction

Empirical studies on return predictability have attracted considerable interest from researchers for some time. Since the publication in 1993 of Jegadeesh and Titman's seminal study, the momentum effect has been one of the most persistent and popular anomalies in both academia and industry. The momentum effect suggests that those assets which have generated positive returns in the past tend to have positive returns in the future, while those that have had negative returns in the past tend to have negative returns in the future. The predictive ability of the momentum effect has challenged the classical efficient markets hypothesis (EMH).

Such momentum is one of the most puzzling anomalies in the financial literature, a conundrum over which researchers have not yet been able to reach consensus. Momentum has been investigated via various studies across a wide range of countries (Fama and French, 2012; Rouwenhorst, 1998). It also has been explored across a diverse array of financial assets (Moskowitz et al., 2012; Gorton et al., 2013). In fact, the momentum phenomenon is not entirely stable and tends to occur during times of market stress. During bear market phases, momentum can have a significant negative beta Grundy and Martin (2001). In the past, the momentum effect has tended to suffer from momentum crash, which can lead to significant losses. In 2016, Daniel and Moskowitz document two major momentum crash periods in the US market – June 1932 to December 1939 and March 2009 to March 2013. During such a crash period, the past–loser decile portfolio gains more than the past-winner decile portfolio – indeed the gap can be as large as 200%. Within recent machine learning-based asset pricing models (Madge and Bhatt, 2015; Peachavanish, 2016), momentum has emerged as a significant factor. Specifically, Han (2022) has identified the risks of a momentum strategy from a cross-sectional perspective and has developed a machine-learning-based model to address such momentum crashes. In his research, he has found that the U-shaped bimodal cross-sectional relative return distribution of momentum stocks in the US market implies that, while past winners may belong to the highest-return group, they also have a higher probability of becoming part of the lowest-return group in the future. This makes the momentum strategy fundamentally risky.

Our current study differs from previous research conducted in developed countries, as it focuses on the Chinese stock market – currently ranked as the second-largest stock market globally after the United States (as of 2017). Given China's prior status as a developing country, its stock market exhibits unique features. Previous research has indicated that momentum profits are frequently attributed to irrational investor behaviour. In contrast to the stock markets of developed countries, in which institutional investors largely dominate, the stock market in China is characterised by a significant proportion of retail investors. As of the end of 2015, individual investors held 88% of all free-floating shares (Shenwan Hongyuan Research Co.). In comparison to institutional investors, individual investors lack investment experience and knowledge, making them more susceptible to emotions. Unlike institutions, individual investors in developing countries are inclined to sell winning stocks and hold losing stocks Brzeszczyński et al. (2015). Given the

inconsistent evidence for momentum in China within the existing literature, this study seeks to provide further proof of the effectiveness of the momentum features using a machine learning approach.

The current academic discourse on the implementation of machine learning for momentum in developed countries predominantly utilises NN algorithms. However, the stock market in China has a shorter history, owing to its delayed entry into the global economy, thus creating a data deficiency that may adversely affect the model's performance. Therefore, this study employs XG-Boost algorithms to evaluate the predictability of momentum features. If both NN and XG-Boost models exhibit favourable predictability, this could imply that momentum features carry useful information for non-linear models. Considering the fundamental differences apparent in constructing these two models, it is certainly worth exploring further.

One purpose of this study is to examine the predictive power of the momentum feature in forecasting the direction of stock performance. Our methodology involves constructing both long and short portfolios through the conventional JT approach, as well as utilising machine learning algorithms to build portfolios. We approach the problem as a supervised learning classification problem and employ two machine learning frameworks, NNs and XG-Boost. Additionally, we employ three models based on NN reclassification Han (2022), including the PrDf, Return, and Sharpe reclassification methods.

Yang et al. (2019) explore the time-variant nature of momentum feature in China's stock market. Throughout the testing period from February 2000 to November 2019, we observe no explicit profitability of the traditional JT strategy in China. Nonetheless, the momentum-based machine learning models demonstrated considerable predictive capabilities in the areas of both model and financial performance. The return of long-only and long-short portfolios, established using the classification labels predicted by the machine learning model, cannot be fully explained by the Fama-French three-factor. Our machine learning portfolio exhibits a statistically significant positive alpha — a desirable outcome for investors. Our research provides an alternative perspective on the time-variant nature of momentum by demonstrating that the correlation between momentum and stock return is non-linear in China.

Simultaneously, we document that the return of our momentum-based portfolios exhibits a statistically significant positive correlation with the SMB factor. This implies that our portfolio is inclined towards yielding positive gains while small-cap stocks outperform large-cap stocks. However, this relationship necessitates cautious assessment when investing in the Chinese market. China has a history of employing the smallest listed firms as shell targets for reverse mergers to bypass restrictive IPO regulations. As a result, these small firms often possess a market value that mirrors their potential use as a shell rather than their actual business worth.

Additionally, our research reveals a statistically significant positive correlation between the returns of our momentum-based long-only portfolios and the changes in the Investor Sentiment Index (ISI). It is worth noting that our long-only

portfolios, which are constructed utilising machine learning techniques, exhibit superior performance as compared to the long-only portfolio constructed using the conventional momentum linear regression model. Specifically, our portfolio has the ability to capture a positive alpha that is statistically significant, even after taking into account changes in the ISI.

The present study also highlights the superior performance of reclassification models over original NNs and XG-Boost models. The Sharpe ratio of long-short portfolios constructed using reclassification models can attain values of up to 1.87. This is attributed to the ability of reclassification to mitigate the bimodal distribution phenomenon in classification. We uncover evidence that such a bimodal cross-sectional relative return distribution arises not only in developed markets, but also in China. However, we only observe the bimodal distribution in the winner decile. This indicates that investing based on the traditional JT strategy could result in losses, given that the stocks purchased by investors are highly likely to become low-return stocks. Under reclassification models, the true return class distribution shifts from bimodal to right-skewed. Notably, less than 6% of extreme losers are included in the winner group, on average.

This chapter is organised as follows: In section 2, we review the previous literature about momentum and machine learning in asset pricing. In section 3, the methodology of Han (2022) is displayed. Section 4 describes the data. In this section, we also present and analyse the performance of the DM model from both machine learning modelling and financial profitability perspectives. Finally, we conclude in Section 5.

3.2 Literature review

Jegadeesh and Titman (1993) document the persistence and pervasiveness of the momentum effect in the US market. They rank stocks based on their previous returns, group them into deciles, and construct a long-short portfolio by buying the last decile (winners) and selling the first decile (losers). They find that the portfolio can generate significant abnormal returns over the next few months. The momentum has been tested extensively in the US market. Green et al. (2017) examine 94 firm characteristics in the US market and find the price momentum is still significant.

However, the performance of the momentum strategy is not stable. The study by Chordia and Shivakumar (2002) indicates that the profit gained from momentum investment during economic expansions exhibits statistically significant positivity, yet insignificant negativity during economic recessions. The authors suggest that the momentum premium in the US stock market is expected to be pro-cyclical. However, it is important to note that their research is based on data from July 1987 to December 1994, which may no longer be up-to-date. Blitz et al. (2011) also finds similar evidence, which documents that the return of the momentum strategy exhibits a substantial average performance of 14.7% per annum during expansionary periods. Conversely, the performance drops to -8.73% annually during recessionary periods. The authors suggest that the poor performance of total return momentum during economic contractions can be attributed

to the stylised fact that the largest market reversals tend to take place during recessionary periods. They document that the losses of total return momentum during recessions are indeed concentrated in the second half of recessions when the market tends to revert.

The study conducted by Daniel and Moskowitz (2016) reveals that a momentum portfolio exhibits a significant difference in its up-market and down-market betas during a bear market. Specifically, the up-market beta is more than double the down-market beta, with values of -1.51 and -0.70, respectively (t-statistic of the difference = 4.5). The majority of this asymmetry in beta values is attributed to stocks with past losers. The findings suggest that the behaviour of the momentum strategy in bear markets is consistent with the written call option. This similarity means that these strategies yield slight gains when the market falls and significant losses when it rises. The momentum strategy is tested across various asset classes, such as equity, index futures, commodity, fixed income, and currency, and also in different countries, including the US, Europe, Japan, and the UK. The research findings indicated that the momentum strategy in all markets and all asset classes suffer from crashes due to the conditional beta and option-like feature of losers. To analyse this, the researchers utilise variance swap returns imputed with the volatility index (VIX) and discover that the momentum strategy payoff exhibits a robust negative correlation with market variance innovations in bear markets. Therefore, the researchers recommend the design of an optimal dynamic momentum strategy in which the winner-minus-loser (WML) portfolio is levered up or down over time to maximise the unconditional Sharpe ratio of the portfolio. This is achieved by Daniel and Moskowitz (2016) by dynamically adjusting the weight on the WML momentum strategy using the forecasted return and variance of the strategy. They use a simple linear regression model to forecast the expected return and variance of the WML momentum strategy based on lagged returns and variances. They then use these forecasts to adjust the weight of the WML momentum strategy. When expected returns are high and volatility is low, they increase their exposure to the momentum portfolio. When expected returns are low and volatility is high, they decrease their exposure. By dynamically adjusting their weight on the momentum portfolio in this way, they are able to generate a Sharpe ratio which is more than double that of a baseline $1long/1$ short WML strategy.

On the other hand, Barroso and Santa-Clara (2015) predict the risk of momentum by using its own lagged risk as a predictive variable. It is worth noting that the main difference arising between Daniel and Moskowitz (2016) and Barroso and Santa-Clara (2015) is that the dynamic strategy of Daniel and Moskowitz (2016) adjusts the weight on the momentum portfolio based on expected returns and volatility, while the constant volatility strategy of Barroso and Santa-Clara (2015) maintains a fixed level of risk exposure. In more recent research, Han (2022) documents the U-shaped bimodal cross-sectional distribution. He double-sorted US stocks on the on-year price momentum and the one-month ahead returns over the period from 1955 to 2016 and acquired the time-series averages of the resulting cross-sectional distributions. He suggests that the momentum strategy is fundamentally risky because, despite the high probability that the stocks of the winner group will remain winners, they also have a high likelihood of becoming losers.

He presents a DNN model using momentum-related features as inputs, and his momentum strategy did not suffer the momentum crash.

Beyond the US market, many studies on momentum have been conducted in different regions. Earlier momentum in different countries' tests dates back to 1998. In a different approach from testing country indices, Rouwenhorst (1998) finds evidence of momentum within markets and cross markets at the individual stock level of 12 European countries from 1978 to 1995. After that, Rouwenhorst (1999) examines 1,705 firms from 20 emerging markets and suggests that the stocks exhibit momentum.

Chui et al. (2000) document the profitability of momentum in seven Asian markets, including Hong Kong, Indonesia, Korea, Malaysia, Singapore, Taiwan and Thailand. Although there is some overlap between their study and that of Rouwenhorst (1999), they expand the data size by including all the listed companies and further periods. After expanding the data, their sample includes those periods occurring both before and after many of these markets first opened to foreign investors. This is important since previous research suggests that foreign investors tend to be momentum investors (Grinblatt and Keloharju, 2000).

In the subsequent research, more scholars went on to test the momentum profit in different countries. Griffin et al. (2003) calculated the average winner minus loser profits for each country tested in local currency and found that the Asian countries display the weakest momentum profits. Compared with Asia, 14 of the 17 European countries displayed positive mean momentum profits over the test period. Meanwhile, the profitability of momentum is not limited to the stock market, but also exists in relation to other financial assets. Asness et al. (2013) studied momentum with value features together across eight markets. They test the individual stocks in the US, UK, Japan and continental EU, government bonds, currencies, equity index futures, and commodity futures. They claim to have found ubiquitous evidence of value and momentum return premia across all eight tested markets and asset classes. In the previous chapter, we also test the profitability of momentum in the UK, Japan and South Korea. Unfortunately, during our testing period, the traditional JT portfolio does not display good profitability in Asian countries.

China has a relatively short history of involvement in the stock market, with its inception occurring only in December 1990. Nevertheless, China has captured significant attention from international investors, particularly after A-shares were added to the MSCI emerging market index in June 2017. The Chinese stock market also has unique characteristics. In comparison to developed economies, Chinese stocks exhibit higher volatility. Furthermore, investors find it challenging to hedge their position in the Chinese stock market. Until 2018, listed companies were not allowed to delist and short selling was not entirely open. Additionally, the Chinese stock market is primarily dominated by individual investors, who tend to be more emotional than institutional investors Kaniel et al. (2008). Individual investors exhibit disposition behaviour, which entails buying stocks following declines in the previous month and selling following price increases.

Besides, during the market's downturn, investors tend to exhibit a significantly stronger herding tendency (Fu and Lin, 2010). Therefore, it is worthwhile investigating the performance of the momentum in this unique market.

The evidence of the momentum effect in the Chinese stock market is controversial. By testing the A-shares from 1993 to 2000, Kang et al. (2002) find statistically significant intermediate-horizon momentum profits in the China stock market. Wang and Chin (2004) suggest that momentum is exhibited in low-volume stocks while high-volume stocks exhibit return reversals. Wang and Chin (2004) posit that the observed differences in outcomes may be attributed to the unique characteristics of the Chinese stock market, such as the prohibition of short selling and the dominance of inexperienced individual investors. This implies that the Chinese stock market is more susceptible to behavioural issues.

According to Cheema and Nartea (2014), momentum profits in China are anticipated to be lower as compared to the United States. The authors have expanded the sample period up to June 2013 and obtained relevant supporting evidence, stating that momentum returns in China are not necessarily higher for those companies with greater information uncertainty. This phenomenon is attributed in part to restrictions on short selling. Further, the authors' findings indicate that the Fama-French factor is inadequate in explaining the momentum of profits in China. They note a significant loss of momentum strategy in the Chinese market during the GFC subperiod test (2007-2013). However, recent research by Chen et al. (2018) reveals that the low-frequency momentum strategy remains profitable in China, even when controlling for the effects of size and bid-ask bounce.

Yang et al. (2019) reviewed the majority of momentum research in China and claim that the previous literature employs different filters, which may lead to a "designed" result. The authors have provided confirmatory evidence for the time-variant nature of momentum returns, which they explore further by splitting the sample period into five sub-periods: 1993–1996, 1997–2000, 2001–2004, 2005–2008, and 2009–2012. In each sub-period, momentum strategies, including purchasing past winners and selling past losers, are implemented for a maximum of 12 months. No significant findings are identified for the initial stages of Chinese market development (1993–2000). On the other hand, during the sub-sample periods of 2005–2008 and 2009–2012, momentum strategies result in negative returns, signifying the non-existence of momentum effects during these two sub-periods. Conversely, contrarian trading has the potential to generate substantial profits. In contrast, for the sub-sample period of 2001–2004, significant momentum is observed when formation periods exceed three months and for holding periods near 12 months (6–12 months for most strategies). Therefore, long-term momentum is evident during the sub-sample period of 2001–2004.

Han and Shi (2022) examine 62 cross-sectional anomalies in the Chinese market and test nine different momentum-related factors. Their findings suggest that the overall performance of the momentum category in the Chinese market is poor. During their total sample period from 1995 to 2017, there is no statistical evidence for traditional JT in equal-weighted and value-weighted portfolios. It is noteworthy that, in their sub-sample test, the equal-weighted

individual return of momentum is positively significant from 1995 to 2006. However, this relationship reversed in the sub-period from 2007 to 2017, with momentum exhibiting a significant negative correlation with returns. This finding provides further evidence that momentum in China behaves differently in terms of its linearity over time.

The inconsistency of the linear relationship across different periods is why we are more interested in studying the momentum effect in the non-linear relationship. Due to the absence of any unanimous agreement among Chinese scholars regarding the identification of momentum, our paper aims to facilitate the establishment of robust findings concerning the existence of momentum effects in China.

In recent years, machine learning has gained more attention from researchers in the asset pricing arena. The earlier wave comes from the application from a computer science aspect. Compared with other data sources, the information from the financial market has a noisy nature, which means the learning process needs to be guided (Abu-Mostafa and Atiya, 1996; Deboeck, 1994). Therefore, a previous paper about machine learning in asset pricing is more focused on the analysis of the accuracy of model designing.

Like other financial researchers, we want to focus on financial analysis by introducing machine learning as a useful tool only. Gu et al. (2020) define machine learning from an empirical asset pricing angle. They claim that machine learning is a diverse collection of high-dimensional models for statistical prediction. This method combines the "regularisation" technique to mitigate the effect of overfitting and efficient algorithms which can acquire information from potential model specifications. In the paper, they compare different machine learning methods for predicting stock returns in the US market. The models that they tested include linear regression, generalised linear models with penalisation, principal components regression (PCR), partial least squares (PLS), regression trees, and NNs. Gu et al. (2020) suggest that DNNs have shown superior performance in the asset pricing arena. Messmer (2017) propose that the performance of their long-short portfolio, based on deep feed-forward NNs, can produce considerably appealing risk-adjusted returns as compared to a linear benchmark. Further, they assert that their findings are robust to size, weighting schemes and portfolio cut-off points. Feng et al. (2018) construct a deep learning framework and build in a new factor. They try to capture the non-linearity and interactions arising among financial predictors. They also suggest that their deep factors can significantly improve the predictive ability of stock returns. Han (2022) also document favourable results based on the deep-learning structure. That research provides supportive evidence for applying DNNs in the asset pricing arena.

Based on equilibrium theory, the difference in expected returns reflects compensation for different degrees of risk. Previous specific linear risk factors can work on certain single-sort portfolios.Chen et al. (2023) utilises a DNN to estimate a general non-linear asset pricing model for all U.S. equity data. The model incorporates a considerable proportion of macroeconomic and firm-specific characteristics. One innovative approach that this study employs

is the inclusion of the no-arbitrage condition within the NN algorithm. Over recent years, there have been various contributions to the literature which document a significantly large number of different predictors of future returns.

Green et al. (2013) analyse the population of more than 330 public return predictive signals over the 40-year period from 1970 to 2010. They find that the correlations arising between those features are high. The high-dimensional nature of the machine learning method can improve the ability to draw useful information from past returns. Therefore, there is an abundance of existing research about the application of machine learning in asset pricing which tends to use extensive features as inputs based on 60 financial and economic features over a 10-year period, Zhong and Enke (2017) try to predict the daily direction of S&P 500 Index ETF (SPY) returns. They applied the fuzzy c-means method (FCM) and principal component analysis (PCA) to reduce the dimensionality. They test artificial neural networks (ANNs) and logistic regression models as the training model. They find that combining ANNs with PCA provides significantly higher risk-adjusted profits.Gu et al. (2020) conduct extensive scale input features in their research which include 94 characteristics for each stock. Feng et al. (2018) use 56 published firm characteristics as predictors in their deep learning factor model.

Previous research on machine learning in asset pricing has tended to rely on branch predictors and use machine learning as a dimension-reducing tool. However, Han (2022) has focused more on financial phenomena, using fewer factors. His research documents the existence of bimodality in the cross-sectional distributions of both high- and low-momentum stocks. This finding has sparked our interest in investigating whether the same phenomenon exists in China.

3.3 Methodology

3.3.1 Data

The data groups utilised in this study are acquired from the Datastream database and comprise monthly returns, prices, volumes, and market capitalisation. Notably, the available data range for China is comparatively shorter, beginning in December 1990. To facilitate comparison with a local benchmark, we collect returns of the primary index, Shanghai Securities Composite Index, from the China Stock Market and Accounting Research (CSMAR) database. To be consistent with other related research, we only study stocks in one stock exchange per country. In this study, we have opted to analyse stocks listed on the Shanghai Stock Exchange instead of the Shenzhen Stock Exchange. The main reason for this decision is that the Shanghai Stock Exchange holds the distinction of being the largest and most established stock exchange in China. It is important to note that the Shanghai Stock Exchange also is the fourth largest stock exchange globally. We utilise the stock data from A-shares listed on the Shanghai Stock Exchange (SSE), quoted only in the Chinese Renminbi (RMB) and historically closed to foreign investors. This constraint serves to narrow the effect of investor type further. To avoid survivorship bias, the samples comprise both listings and delisted stocks.

The provided samples have been categorised into three distinct groups, specifically the training set, the validation set, and the test set. The training set is used to train the model parameters. The training set is essential for the model to learn the patterns or relationships arising, which will ultimately enable it to make accurate predictions in the future. It serves as the foundation for the model's parameter learning process. On the other hand, the validation set, constituting 30% of the total training data, plays a crucial role in providing an impartial evaluation of the model's performance and refining its parameters. Additionally, in the realm of supervised learning, validation sets are instrumental in identifying instances of overfitting. Meanwhile, the test set measures the actual performance of the final model, spanning from February 2000 to November 2019. It is noteworthy that the model undergoes retraining yearly during this period, with all historical data available before February 2000 utilised for training, while only the past 80 months of data are considered for training after that time frame.

3.3.2 Input features and test models

Table 3.1(a) provides a summary of the input features used, while Table 3.1(b) presents the combinations of input features and models that are tested within the empirical analysis. In this study, momentum features are constructed using standardised momentum features, $nMOM_m$, and also the cross-sectional means, M_{MOM_m}. The concept of the m-month price momentum, denoted as $MOM_{m,i}$, refers to the accumulation of returns over a period of m months for each individual stock i in the given dataset. However, when $m = 1$, it represents the return from the previous month. Five momentum characteristics are computed using the equation proposed by Jegadeesh and Titman (1993):

$$MOM_{m,i} = \prod_{j=t-m}^{t-2} (r_{j,i} + 1) - 1, \quad m = 3, 6, 9, 12, \tag{3.1}$$

$$MOM_{1,i} = r_{t-1,i}, \quad m = 1, \tag{3.2}$$

where $r_{j,i}$ denotes the return of stock i in month j. Then we standardised the price momentum characteristics for each month:

$$nMOM_{m,i} = \frac{MOM_{m,i} - M_{MOM_m}}{S_{MOM_m}}, \tag{3.3}$$

where M_{MOM_m} and S_{MOM_m} respectively denote the cross-sectional mean and standard deviation of MOM_m. Our momentum includes one feature per stock, as they reflect the past performance of each individual stock relative to its own history. However, they are also influenced by the market, as they are standardised by the cross-sectional mean and standard deviation of the momentum features for each month.

Size Dummy:

The size dummies improve the model performance significantly in the US market, especially for the value-weighted group. In each month, stocks would be separated into ten size groups based on their market capitalisation. The top 10% would be marked in group 0, and the remainder are classified into groups from 1 to 9 in descending order. After the group classification, the one-hot encoding process transfers the class group to its categorical features in binary.

Table 3.1: Model specifications
This table lists the specifications of the test models. Panel (a) lists the input features of the neural network models, while panel (b) lists the models tested in the empirical study.

(a) Input Features

Price momentum features	
$nMOM_m$	Standardised m-month price momentum.
M_{MOM_m}	Mean of MOM_m.
Size Dummies	
D_s	Size dummy.

(b) Test Models

Model	Input features
Neural networks-NOM	$nMOM_m$ + M_{MOM_m}
Neural networks-SZ-NOM	$nMOM_m + M_{MOM_m} + D_s$
XG-Boost-SZ-NOM	$nMOM_m + M_{MOM_m} + D_s$

3.3.3 Methodology of neural networks and the DM model

This research follows the methodology from Han (2022) referred to as the DM model. This methodology is separated into two parts. In the first part, the model estimates the cross-sectional distribution arising through the NN. After that, in the second part, the results from the first part are reclassified by three different reclassification criteria. In this research, the DNN is employed for the purposes of stock classification. A NN includes an input layer, a hidden layer(s), and the output layer. The DNN model requires more than one hidden layer. We apply the supervised learning methodology, which means the input data is labelled, and the target task is to make the prediction as close to the actual labels as possible. The output is labelled into the K class group (in this paper, we have K=10 class groups). The DNN will estimate the probability for each class and select the class with the highest probability of each class as the predicted class.

The target of training the DNN model is to minimise the cost function. The cost function tries to measure the difference arising between the prediction value and the true value. Rumelhart et al. (1986) develops the back-propagation process which can exploit the chain rule. During the training process, the model tries to find a function which can minimise

the cost function and best maps the inputs to the outputs. During the back-propagation process, the model can keep updating the parameter set to decrease the loss. This optimisation process is also known as the training process. Given a sample of N observations, $\{x^i, y^i\}_{i=1}^N$, the cost functions are represented as follows:

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} L\left(y^i, \hat{y}^i\left(x^i; \theta\right)\right), \tag{3.4}$$

in the cost function, θ is a set of model parameters which the machine learning model learns from. $\hat{y}(x;\theta)$ is the output of the DNN, whose k-th elements is given by the softmax function. The softmax function, also called the normalised exponential function, converts the vector components to values between 0 and 1.

$$\hat{y}_k(x;\theta) = \frac{e^{z_k(x;\theta)}}{\sum_{k=1}^{K} e^{z_k(x;\theta)}}, \quad k = 1, \ldots, K, \tag{3.5}$$

where K represents the number of classes in the classification problem. The output vector and the softmax function are both K-dimensional, as they correspond to the K possible classes in the classification problem.

The loss function, $L\left(y, \hat{y}(x;\theta)\right)$, is the cross-entropy loss. $L(\theta)$ represents the discrete probability distribution of the target classes.

$$L\left(y, \hat{y}(x;\theta)\right) = -\sum_{k=1}^{K} y_k \log \hat{y}_k(x;\theta). \tag{3.6}$$

We apply the reclassification methods on the NNs and construct portfolio results based on the reclassification results. There are three reclassification methods employed in this chapter, specifically $Prdf$, Ret and $Sharpe$. The method $PrDf$ uses probability difference as a proxy for the expected return which tries to capture the relative return. It tries to capture the information from the law of total expectation and then estimates the expected stock returns(μ_k) in each decile portfolio. It can be shown that the mean return of a stock is proportional to:

$$PrDf = \sum_{k=1}^{K/2} \left(\hat{y}_k - \hat{y}_{K+1-k}\right) \left(\frac{K}{2} + 1 - k\right). \tag{3.7}$$

Thus PrDf reclassifies stocks using this value.

The second method is more intuitive and is known as the Ret reclassification. This reclassification is based on the mean return. The mean return of each class is estimated by the sample analogue. The mean stock return is estimated from the

following equation. $\hat{\mu}_k$ donates the sample mean of class k. The estimate of stock i's mean return is derived as follows:

$$\hat{\mu}^i = \sum_{k=1}^{K} \hat{y}_k^i \hat{\mu}_k \tag{3.8}$$

The third method is based on the prediction of the Sharpe ratio. Using the law of total variance, the variance of stock return is given by:

$$\sigma^2 = V[r] = E\left[V[r|c]\right] + V\left[E[r|c]\right] = \sum_{k=1}^{K} P(k)\left(\sigma_k^2 + \mu_k^2\right) - \mu^2, \tag{3.9}$$

where σ_k^2 is the variance of class k's return. Substituting μ_k and σ_k^2 with their sample analogues $\hat{\mu}_k$ and $\hat{\sigma}_k^2$, the variance of a stock's return can be estimated using the equation:

$$\hat{\sigma}^2 = \sum_{k=1}^{K} \hat{y}_k\left(\hat{\sigma}_k^2 + \hat{\mu}_k^2\right) - \hat{\mu}^2. \tag{3.10}$$

The Sharpe ratio of the stock is then estimated by $\hat{SR} = \hat{\mu}/\hat{\sigma}$.

3.3.4 XG-Boost

We also present a predictive modelling approach, utilising the XG-Boost algorithm, to forecast stock positions in the financial market. During training, our target output corresponds to the NN model. The objective function is specified as "Softmax," while the loss function is specified as "Multiclass logloss." The logloss is defined as the negative log-likelihood of a logistic model that returns y_{pred} probabilities for its training data y_{true}. For example, if sample i has label k taken from a set of K labels. Let P be a matrix of probability estimates, with $p_{i,k} = Pr(y_{i,k} = 1)$. Then the log loss of the whole set is:

$$L_log(Y, P) = -\frac{1}{N}\sum_{i=0}^{N-1}\sum_{k=0}^{K-1} y_{i,k} log p_{i,k} \tag{3.11}$$

In order to attain optimal performance, we have implemented the randomised search cross-validation technique to optimise the model's hyperparameters. Thus, a randomised search cross-validation technique was performed on the training set, whereby the XG-Boost Classifier was utilised with various hyperparameters, including maximum depth and minimum child weight. The XGBoost model is trained on the hyperparameters using the train set and is evaluated on both a validation and test set.

The performance of the model is evaluated using the balanced accuracy scoring method. This metric takes an average of recall scores per class or raw accuracy. Additionally, each sample is assigned a weight based on the inverse prevalence of its true class, which is helpful for imbalanced datasets. The formula for balanced accuracy is calculated as half of the

sum of sensitivity and specificity. Sensitivity represents the correct identification of true positives, while specificity measures the correct identification of true negatives.

Based on the chosen evaluation metric, we select the optimised hyperparameters for the model. Subsequently, the model is employed to predict the stock's class in the test set.

3.3.5 Model Evaluation

In the beginning, we apply a conventional classification performance metric to evaluate the performance of machine learning models. The metric includes accuracy, precision, recall, and F1 score. The accuracy can evaluate how accurately the model predicts future class.

$$\text{Average Loss} = \frac{1}{N}\sum_{i=1}^{N} L\left(y^i, \hat{y}^i\right), \tag{3.12}$$

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{N}. \tag{3.13}$$

The precision, recall, and F1 score can share more details of how the DNN model classifies samples. Each evaluation indicator is calculated in each decile from 0 to 9. The performance of 0 and 9 was particularly important as it can affect the financial performance of the long-short portfolio.

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{3.14}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{3.15}$$

$$\text{F1-score} = \frac{2}{1/\text{Precision} + 1/\text{Recall}}, \tag{3.16}$$

where TP, FP, and FN, respectively, denote true positive, false positive, and false negative.

3.4 Empirical Analysis

3.4.1 Classification Performance

Table 3.2 displays the accuracy scores of each machine learning model. The equivalence of the test set accuracy and validation set accuracy to the training set accuracy indicates that our model does not suffer from overfitting. Prior research on DM model classification in developed countries such as the United States, United Kingdom, Japan, and South Korea, has demonstrated a test accuracy exceeding 14%. However, our test accuracy in China is marginally

lower, at 13.58%. Nevertheless, this figure is significantly superior to the accuracy of a random guess, at 10%. The inclusion of size dummies in the input feature set results in a modest improvement in the test set's accuracy score, further validating the effectiveness of the size factor. It is worth noting that both the original NN and reclassification models, which are used for the subsequent analysis in this article, incorporate the size dummy. This chapter does not encompass any additional verification procedures concerning the effectiveness of the size dummies.

Chapter One showcases the strong performance of the XG-Boost model in classification problems. Nevertheless, the accuracy score of the XG-Boost model falls short in comparison to that of NNs. This disparity may arise due to the intricacy of the stock data in contrast to the fund data. In certain circumstances, NNs may require more data to achieve optimal performance, whereas XG Boost may be better suited to smaller datasets.

Table 3.2: Classification performance
This table reports the classification performance of three machine learning models, specifically neural networks trained with momentum, neural networks trained with momentum and size, and XG-Boost trained with momentum and size.

Model	Train set	Validation set	Test Set
Neural networks train with momentum (NW-MOM)	14.64	13.54	13.58
Neural networks train with momentum and Size (NW-MOM-SZ)	14.23	13.41	13.65
XG-Boost train with momentum and Size (XGB-MOM-SZ)	13.12	12.21	11.55

Table 3.3 displays the performance of the XB-Boost model (XGB-MOM-SZ), the NN model (NW-MOM-SZ), and the reclassified NN models in predicting the H class over the test period from February 2000 to November 2019. The model evaluation matrix presents various metrics, namely Accuracy, Precision, Recall, and F1 score, to evaluate the respective models' performance.

Upon examining the table, it is evident that the original NNs model has the highest accuracy of 12.93%, followed by XG-Boost at 11.31%. The three models that resulted from different reclassifications, PrDf, Return, and Sharpe, have accuracy scores lower than 11%. It is noteworthy that the accuracy score of the model in the H class is lower than the mean accuracy score of the entire test set. Therefore, the model's capacity to differentiate and screen high-return stocks in the top class is inferior as compared to other classes. In terms of precision, the original NN model again stands out, with the highest score at 1.67%. The other models have precision scores lower than 1.28%.

In conclusion, based on the information provided in the table, the original NN model appears to be the best model in predicting the H class among the five models tested. This model has the highest scores in terms of Accuracy, Precision, Recall, and F1 score. However, it is necessary to conduct further analysis and testing before drawing any definitive conclusions.

Table 3.3: H decile performance
This table reports the H class test set classification performance of neural networks (NW), XG-Boost (XGB), neural networks following reclassifications (PrDfm, Return, Sharpe). The testing phase covers the period from 2000.02 to 2019.11. The 'PrDf', 'Return', and 'Sharpe' denote reclassification on probability difference, reclassification on mean return, and reclassification on Sharpe ratio, respectively.

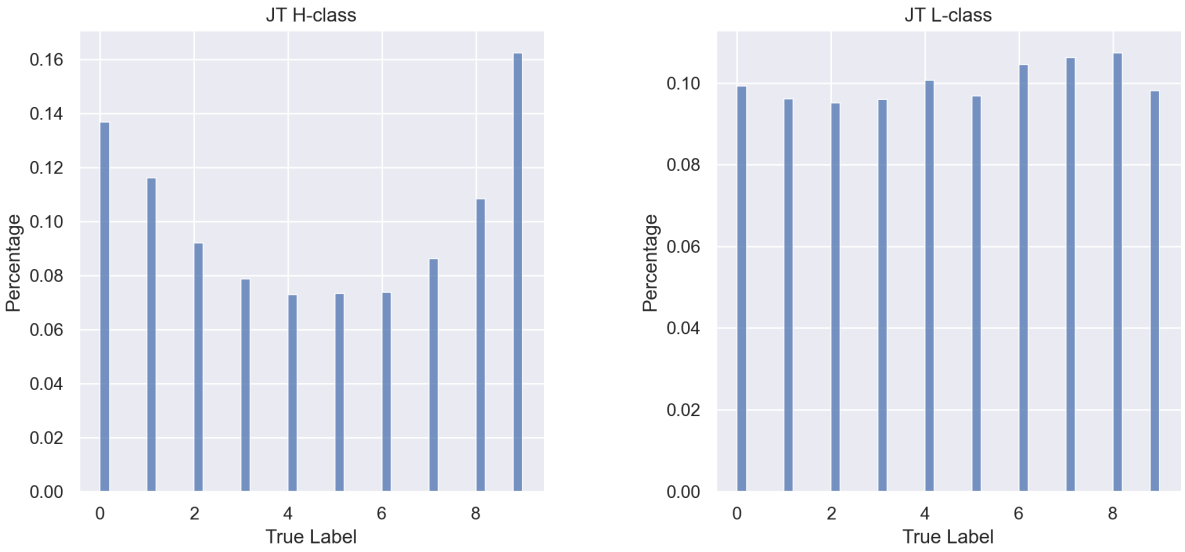
Model	Accuracy	Precision	Recall	F1 Score
Neural Networks	12.93%	1.67%	12.93%	2.96%
XG-Boost	11.31%	1.28%	11.31%	2.30%
Reclassfication				
PrDf	10.16%	1.03%	10.16%	1.87%
Return	10.64%	1.13%	10.64%	2.05%
Sharpe	10.01%	1.00%	10.01%	1.82%

3.4.2 Evidence of bimodality in the H decile

Figure 3.1 depicts the distributions of the true labels in the high-return (H) and low-return (L) deciles of the traditional JT model. Notably, the figure of JT reveals the bimodal distribution inherent within the H decile, indicating that many stocks predicted to fall into H may actually end up in L. It is worth mentioning that, unlike other developed countries, the existence of bimodality in the L in China is not conspicuous. Additionally, the actual frequency of each group is roughly equally proportioned. The observed variance in China's stock market may be attributed to its unique mechanism. Firstly, the Shanghai and Shenzhen Stock Exchanges have implemented a daily price limit rule since December 1996. Regular stocks are only allowed to increase or decrease by 10 percent relative to the previous day's closing price, while special treatment (ST) stocks can only fluctuate by up to 5 percent. Secondly, short sales have been relatively uncommon in the Chinese stock market, with limited permissions granted by the China Securities Regulatory Commission (CSRC) in March 2010. During the Chinese stock market crash of 2015, many trading firms voluntarily stopped short-selling activities due to government pressure. The CSRC also began to act against "malicious" high-frequency traders suspected of engaging in market manipulation. Additionally, in August of the same year, the practice of same-day transaction settlements for short-sellers was discontinued (Morah). The unique features of China's stock market may contribute to the lack of any evident bimodal distribution in the L. In order to enhance the practicality of the study, our focus has thus shifted towards evaluating the performance within the H decile.

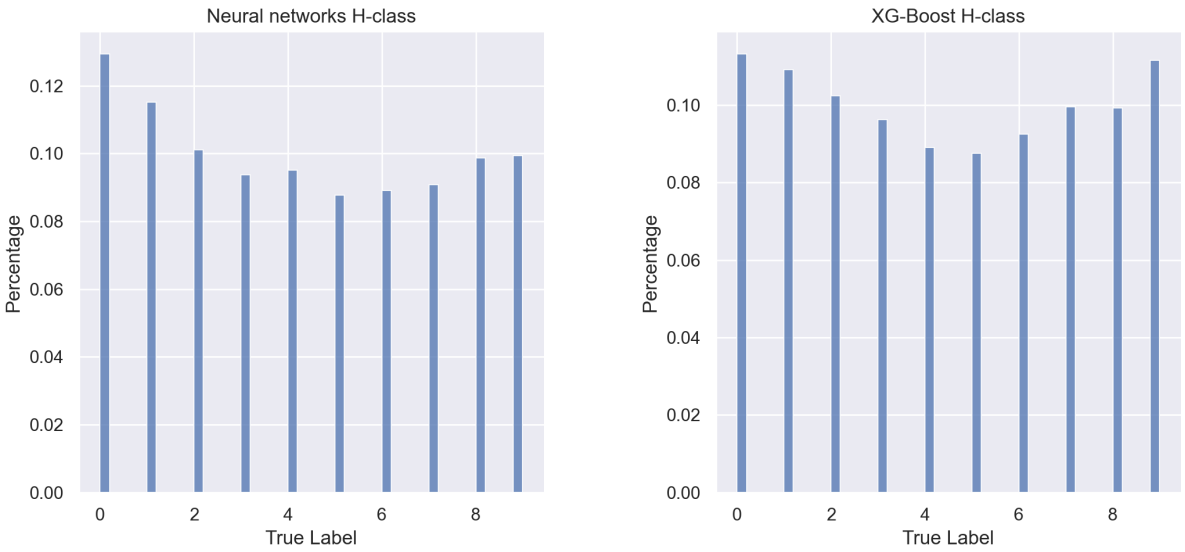
The distribution of true labels for the H decile as predicted by the NN and XG-Boost models is presented in Figure 3.2. While bimodality is not fully eradicated, the machine learning model exhibited less bimodality than the linear model. Preliminary observations from the figure suggest that the NN model outperforms the XGB model by assigning more stocks with good performance to Group H and fewer stocks with poor performance to the same group.

Based on the results presented in Figure 3.3, it is evident that the reclassification process has led to a notable enhancement in the predictive capability of the NN model. Specifically, the distribution of truth labels within the H decile exhibits a right skew. Despite a decrease in the number of top winners classified into the H decile, an even greater decrease in



(a) H decile **(b)** L decile

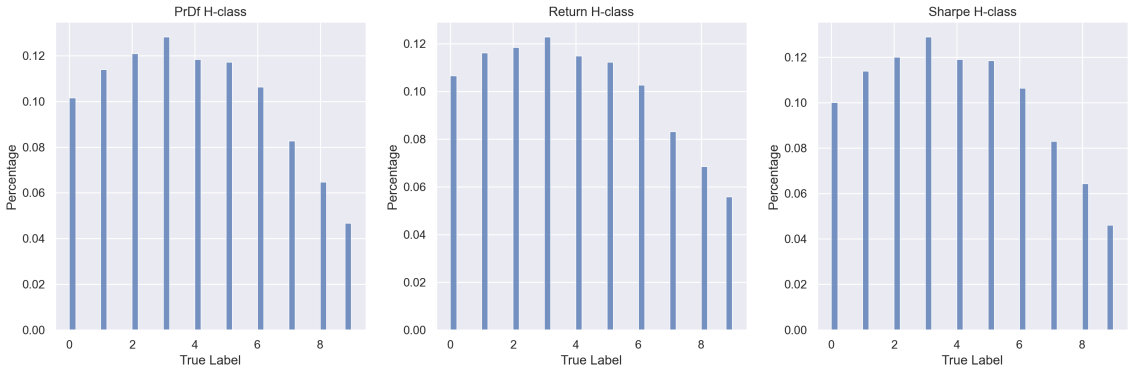
Figure 3.1: JT model
The diagram presented depicts the true label distribution within the JT model prediction group. It is inferred that stocks belonging to the H decile are anticipated to be winners, while those belonging to the L decile are predicted to be losers.



(a) Neural networks model **(b)** XG-Boost model

Figure 3.2: True target distribution in H decile (Original)
The diagram presented depicts the true label distribution within the H decile group of the neural networks and XG-Boost models.

the number of losers classified into the H decile is indicative of a relatively desirable outcome. It has been observed that the probability of the H decile including extreme loser stocks in all reclassification models is less than 6%. Given that extreme losers have a more adverse impact on returns for long-only portfolios as compared to stocks under other labels, a reduction in the number of extreme loser stocks within the H decile could significantly improve the overall profitability of the long-only portfolio.



(a) PrDf Reclassification **(b)** Return Reclassification **(c)** Sharpe Reclassification

Figure 3.3: True target distribution in H decile (Reclassification)
The diagram presented depicts the true label distribution within the H decile group as predicted by the reclassification models.

3.4.3 Financial Performance

Table 3.4 reports the financial performance of the long-only portfolios and the long-short portfolios. According to the preceding literature, the JT strategy lacks strong supportive evidence in the Asian stock market. Our study's findings corroborate this view throughout the testing period. Consistent with Japan and South Korea, the Sharpe ratio of the JT model-based long-short portfolio in China is near zero. Combined with the performance in the H decile, we find that the cause of this outcome is similar to Japan, where the gap between the high-return (H) and low-return (L) deciles is minute. Specifically, the return of L is excessively high. In fact, the selling of the high-yield L can lead to a negative return on the long-short portfolio.

The annualised returns of both the original NN and XG-Boost models are higher than those of the traditional JT model and market index for both long-only and long-short portfolios. However, it has been noted that the annualised returns of long-short portfolios are lower than those of long-only portfolios due to the negative impact of short-side stocks on the overall portfolio's annual returns, particularly in the NN model. In long-short portfolios, the XG-Boost model's annualised return of 0.122 is slightly higher than the original NN model's returns of 0.113. This relationship changes in long-only portfolios, where the NN model's annualised return is as high as 0.173.

Long-short portfolios exhibit significantly less volatility than long-only portfolios, resulting in higher risk-adjusted performance. The Sharpe ratio of the long-short portfolio based on the XG-Boost model is nearly twice that of the long-only portfolio. Specifically, the Sharpe ratio for the former is 0.82, as compared to 0.42 for the latter.

After the reclassification of the stocks, the return on the long side increased further. Specifically, the application of the PrDf reclassification method resulted in an increase in return from 0.173 to 0.317, which is particularly attractive to investors in light of the difficulty in shorting stocks in China. Meanwhile, the annual return of the long-short portfolio doubled from 0.113 to 0.330, while the Sharpe ratio of the long-short portfolio increased significantly to 1.87. The reclassification of DM had a dual effect in augmenting returns on the long side and improving the performance on the short side, thereby enhancing the overall financial performance of the long-short portfolio.

Moreover, it is noteworthy that the PrDf reclassification portfolio exhibited marginally superior performance in China as compared to the return reclassification portfolio, which contrasts with the findings from developed nations. Nevertheless, during the time of investigation, the degree of discrepancy in financial performance among the reclassification models is modest.

Table 3.4: Financial performance

This table reports the financial performance of the equal-weighted long-only portfolio and long-short portfolios constructed by neural networks (NW), XG-Boost (XGB), neural networks following reclassifications (PrDfm, Return, Sharpe), and the traditional momentum strategy (JT). The testing phase covers from 2000.02 to 2019.11. In the matrix, we report the annualised return (AR), standard deviation (std), Sharpe ratio, Sortino ratio, cumulative return (cumulative) and maximum drawdown (MaxDD).

Benchmarks	**AR**	**Std**	**Sharpe ratio**	**Sortino ratio**	**Cumulative return**	**MaxDD**
Index	0.066	0.261	0.20	0.29	0.63	-2.00

(a) Market index portfolios

Model	**AR**	**Std**	**Sharpe ratio**	**Sortino ratio**	**Cumulative**	**MaxDD**
Neural Networks	0.113	0.188	0.52	0.76	1.87	-0.60
XG-Boost	0.122	0.130	0.82	1.46	2.24	-0.18
JT	-0.012	0.211	-0.13	-0.18	-0.68	-2.70
Reclassfication						
PrDf	0.330	0.169	1.87	3.88	6.18	-0.38
Return	0.315	0.168	1.79	3.55	5.90	-0.25
Sharpe	0.326	0.167	1.86	3.89	6.12	-0.34

(b) Long short portfolios

Model	**AR**	**Std**	**Sharpe ratio**	**Sortino ratio**	**Cumulative**	**MaxDD**
Neural Networks	0.173	0.392	0.40	0.62	1.88	-1.43
XG-Boost	0.159	0.345	0.42	0.66	1.97	-1.23
JT	0.104	0.320	0.28	0.41	1.02	-2.59
Reclassfication						
PrDf	0.317	0.381	0.79	1.41	4.83	-0.86
Return	0.311	0.385	0.77	1.36	4.69	-0.87
Sharpe	0.314	0.380	0.79	1.40	4.78	-0.88

(c) long-side portfolios

We present the cumulative returns of the long-only and long-short portfolios in Figure 3.4. In Figure 3.4a, it is observed that the cumulative returns for all long-only portfolios of the models experienced a significant decline during the late stages of the financial crisis, which is in alignment with previous research (Blitz et al., 2011; Chordia and Shivakumar, 2002). Despite the frequent fluctuations in the cumulative returns of our long-only portfolio, our portfolio performs better than both the traditional JT model and the market index for a significant part of the time. Notably, the cumulative return trend of XG-boost is more akin to the traditional JT model and the market index as compared to the original NN. This suggests that the stocks in the high-return group, as classified by the XG-Boost model, share a higher correlation with the market index.

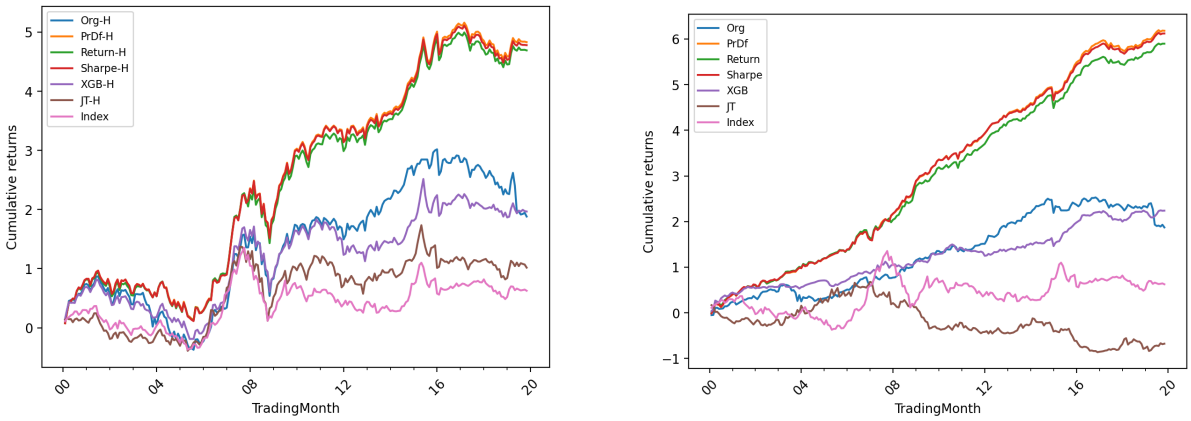
In Figure 3.4b, the cumulative return performance of the long-short portfolios is illustrated. Upon combining the short decile, the performance of the machine learning model improves significantly, resulting in a substantial outperformance of the traditional JT portfolio. The traditional JT model suffers significant losses during the financial crisis, consistent with previous results documented by Cheema and Nartea (2014). Furthermore, when comparing the portfolio of the original NN model with that of XG-Boost, the latter exhibits greater stability in terms of cumulative return. The maximum drawdown experienced by the original NN model is more than three-fold that of the XG-Boost model, a situation which is typically considered unfavourable by investors.

Moreover, the study findings suggest that the application of reclassification methods has proven to be effective in both long-only and long-short portfolios in China. Long-only portfolios based on reclassification have exhibited a significantly higher cumulative return than other portfolios. Moreover, reclassification portfolios have demonstrated the ability to drop less than other portfolios during the financial crisis period. On the other hand, the long-short portfolio based on original NNs has, at times, endured considerable losses, with the cumulative return of the original NN portfolio falling even lower than that of the JT portfolio during certain time periods. However, the long-short portfolios constructed after reclassification have demonstrated no significant losses and have exhibited steady growth throughout the test period.

3.4.4 Fama-French Regression

To determine whether the return of the long-only and long-short portfolio could be explained by the Fama-French three factors, a factor regression analysis is conducted, following previous research in developed countries. The relative factors in China are obtained from the CSMAR database, and the test period covered the years from 2000 to 2019. The results from our regression analysis are presented in Table 3.5.

The portfolios constructed by traditional JT, including the long-only and long-short portfolios, do not exhibit a statistically significant positive alpha after regression with the Fama French three factors. Conversely, the models



(a) Long-only portfolio

(b) Long-short portfolio

Figure 3.4: Cumulative Returns (Equal-weighted)
The given figure demonstrates the cumulative returns (logarithmic scale) of the equal-weighted long-only portfolio and long-short portfolios as constructed by neural networks (NW), XG-Boost (XGB), neural networks following reclassifications (PrDfm, Return, Sharpe), and the traditional momentum strategy (JT). The testing phase covers the period from 2000.02 to 2019.11.

constructed through machine learning algorithms demonstrate statistically significant positive alpha, indicating that these portfolios generate excess returns independent of market, size, or value factors.

The long-only portfolio of XG-Boost displays a positive and statistically significant Rm-Rf coefficient, providing further support for the finding of cumulative returns. Moreover, all of our long-only portfolios display a statistically significant positive coefficient with the SMB factor, suggestive of positive exposure to small-cap stock premiums. Notably, the relationship arising between the SMB factor and the long decile is consistent with prior regressions conducted in other developed nations such as the US, UK, Japan, and South Korea. This relationship also exists in our long-short portfolio as constructed through machine learning algorithms, potentially due to the addition of size dummies to the training process. It is advisable to monitor the portfolio's exposure to small stocks carefully. In China, the smallest listed firms are frequently utilised as shell targets for reverse mergers due to the stringent regulations on IPOs. Consequently, these small firms often possess a market value that reflects their prospective use as a shell rather than their actual business value. This cross-contamination of small-firm stock prices with shell value renders the size factor an inadequate gauge of stock risk and return in China. As per Liu et al. (2019), 83% of the reverse mergers in China involve shells sourced from the smallest 30% of stocks. In the XG-Boost long-short portfolio, the HML factor shows a positive coefficient of 0.20 with a t-value of 2.18, indicating positive exposure to a high book-to-market ratio stock premium. However, the NN-based portfolios did not demonstrate statistically significant exposures to these HML factors.

The effectiveness of reclassification is apparent, as the alpha doubled after reclassification. With the exception of the SMB factor, reclassification portfolios do not exhibit a statistically significant relationship with other Fama-French factors, indicating that portfolio returns for these portfolios could not be explained by market or book-to-market ratio factors.

Table 3.5: Fama-French factor regression

This table reports the factor regression results of the equal-weighted long-only portfolio and long-short portfolios constructed by neural networks (NW), XG-Boost (XGB), neural networks following reclassifications (PrDfm, Return, Sharpe), and the traditional momentum strategy (JT). The portfolio returns are regressed on the Fama-French three factors (Rm-Rf, SMB, HML). For each portfolio, the first row reports the coefficients and the second row reports the Newey-West adjusted t-statistics. The testing phase covers the period from 2000.02 to 2019.11.

Model		Constant	Rm-Rf	SMB	HML
JT	coef	0.00	0.99	0.41	-0.24
	t-values	-0.10	27.98	6.66	-2.39
Neural networks	coef	0.01	0.06	0.37	0.00
	t-values	1.36	0.59	2.18	0.01
XG-Boost	coef	0.01	0.11	0.33	0.01
	t-values	1.37	1.25	2.27	0.02
Reclassfication					
PrDf	coef	0.02	0.09	0.36	-0.01
	t-values	3.06	0.93	2.22	-0.05
Return	coef	0.02	0.08	0.35	-0.01
	t-values	2.97	0.82	2.15	-0.04
Sharpe	coef	0.02	0.09	0.36	-0.01
	t-values	3.03	0.96	2.23	-0.03

(a) Long-only portfolios

Model		Constant	Rm-Rf	SMB	HML
JT	coef	0.00	0.00	-0.02	0.01
	t-values	-0.70	-0.06	-0.21	0.05
Neural networks	coef	0.01	-0.02	0.12	0.12
	t-values	1.85	-0.38	1.49	0.93
XG-Boost	coef	0.01	-0.01	0.13	0.20
	t-values	2.88	-0.33	2.40	2.18
Reclassfication					
PrDf	coef	0.02	-0.04	0.19	0.12
	t-values	7.68	-0.98	2.59	1.04
Return	coef	0.02	-0.05	0.17	0.11
	t-values	7.37	-1.22	2.34	0.90
Sharpe	coef	0.02	-0.04	0.18	0.13
	t-values	7.65	-0.96	2.56	1.09

(b) Long-short portfolios

3.4.5 Sentiment analysis

The regression results of portfolio returns and the change in sentiment factor are presented in Table 3.6. Monthly data from the ISI spanning from January 2003 to November 2019 are collected from CSMAR. The selected ISI version is custom-made for China and created by Wei et al. (2014). This measure encompasses several factors, including closed-end fund discount, average first-day return, number of IPOs, new stock accounts, market turnover rate, and consumer confidence index.

The impact of investor sentiment on momentum-based long-only portfolios is worth noting. These portfolios exhibit statistically significant positive correlations with changes in the sentiment index. Particularly noteworthy is our machine learning-based long-only portfolio, which demonstrates a higher correlation and t-value as compared to the traditional

JT long-only strategy. This indicates that our model outperforms the JT long-only portfolio in capturing the effect of sentiment. Additionally, our model is more sensitive and responsive to fluctuations in sentiment, enabling it to generate higher returns during periods of increased sentiment and lower returns during periods of decreased sentiment. Moreover, the constant of the JT long-only portfolio is not significant. However, with the exception of the original NN model, all of our machine learning-based long-only portfolios exhibit a positive and significant constant. These results suggest that sentiment fluctuations alone cannot fully explain the returns of our portfolio.

Table 3.6: Sentiment regression

This table reports the sentiment factor regression results of the equal-weighted long-only portfolio and long-short portfolios constructed by neural networks (NW), XG-Boost (XGB), neural networks following reclassifications (PrDfm, Return, Sharpe), and the traditional momentum strategy (JT). The portfolio returns are regressed on the change in sentiment ($ISI_t - ISI_{t-1}$). For each portfolio, the first row reports the coefficients and the second row reports the Newey-West adjusted t-statistics. The testing phase covers the period from 2003.02 to 2019.11.

Model		**Constant**	$ISI_t - ISI_{t-1}$
JT	coef	0.01	0.02
	t-values	1.58	2.02
Neural networks	coef	0.01	0.05
	t-values	1.84	5.96
XG-Boost	coef	0.01	0.05
	t-values	2.01	6.75
Reclassfication			
PrDf	coef	0.03	0.05
	t-values	3.71	6.31
Return	coef	0.03	0.06
	t-values	3.63	6.43
Sharpe	coef	0.03	0.05
	t-values	3.69	6.31

Long-only portfolios

3.5 Conclusion

This study contributes to the existing literature on the momentum effect by examining its predictability in the Chinese stock market using machine learning models. Although we do not find significant evidence for the effectiveness of the traditional JT strategy, our research demonstrates that momentum-based machine learning models have considerable predictive power in forecasting stock performance in China. Our research provides an alternative perspective on the time-variant nature of momentum and suggests that momentum features contain effective information for non-linear models.

Furtherm, we highlight the superior performance of reclassification models over traditional NNs and XG-Boost models. Our findings indicate that the bimodal distribution phenomenon occurs in China's stock market, with the winner decile exhibiting a higher probability of shifting to the loser group. As such, investing based on the traditional JT strategy could lead to significant losses. Reclassification models can mitigate the bimodality in the High return class and improve overall portfolio performance.

Overall, our study contributes to a deeper understanding of the momentum effect in China's stock market and highlights the importance of using machine learning models to mitigate the risks associated with momentum investing. Whilst our long-only portfolio performed well under most market conditions, its performance plummeted during the period of the financial crisis. In future research, scholars may wish to delve into the possibility of controlling portfolio risk through machine learning techniques. This could potentially enhance the portfolio's risk-adjusted return, thus warranting further discussion and analysis.

Bibliography

Abu-Mostafa, Y.S., Atiya, A.F., 1996. Introduction to financial forecasting. Applied Intelligence 6, 205–213.

Agyei-Ampomah, S., 2007. The post-cost profitability of momentum trading strategies: further evidence from the UK. European Financial Management 13, 776–802.

Ahmadpour, K., Frömmel, M., 2022. The role of gender for the risk-shifting behavior of hedge fund and cta managers. Finance Research Letters 47, 102635.

Amstad, M., Sun, G., Xiong, W., 2020. The handbook of China's financial system. Princeton University Press.

Andreu, L., Puetz, A., 2017. Choosing two business degrees versus choosing one: What does it tell about mutual fund managers' investment behavior? Journal of Business Research 75, 138–146.

Andreu, L., Sarto, J.L., Serrano, M., 2019. Risk shifting consequences depending on manager characteristics. International Review of Economics & Finance 62, 131–152.

Asness, C., 2011. Momentum in Japan: The exception that proves the rule. The Journal of Portfolio Management 37, 67–75.

Asness, C.S., Moskowitz, T.J., Pedersen, L.H., 2013. Value and momentum everywhere. The Journal of Finance 68, 929–985.

Atkinson, S.M., Baird, S.B., Frye, M.B., 2003. Do female mutual fund managers manage differently? Journal of Financial Research 26, 1–18.

Babalos, V., Caporale, G.M., Philippas, N., 2015. Gender, style diversity, and their effect on fund performance. Research in International Business and Finance 35, 57–74.

Banz, R.W., 1981. The relationship between return and market value of common stocks. Journal of Financial Economics 9, 3–18.

Barroso, P., Santa-Clara, P., 2015. Momentum has its moments. Journal of Financial Economics 116, 111–120.

Beckmann, D., Menkhoff, L., 2008. Will women be women? analyzing the gender difference among financial experts. Kyklos 61, 364–384.

Berk, J.B., Green, R.C., 2004. Mutual fund flows and performance in rational markets. Journal of political economy 112, 1269–1295.

Berk, J.B., Van Binsbergen, J.H., 2015. Measuring skill in the mutual fund industry. Journal of financial economics 118, 1–20.

Bernasek, A., Shwiff, S., 2001. Gender, risk, and retirement. Journal of economic issues 35, 345–356.

Bliss, R.T., Potter, M.E., 2002. Mutual fund managers: does gender matter? The Journal of Business and Economic Studies 8, 1.

Blitz, D., Huij, J., Martens, M., 2011. Residual momentum. Journal of Empirical Finance 18, 506–521.

Breiman, L., 1996. Bagging predictors. Machine learning 24, 123–140.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Brzeszczyński, J., Gajdka, J., Kutan, A.M., 2015. Investor response to public news, sentiment and institutional trading in emerging markets: A review. International Review of Economics & Finance 40, 338–352.

Carhart, M.M., 1997. On persistence in mutual fund performance. The Journal of Finance 52, 57–82.

Chae, J., Eom, Y., 2009. Negative momentum profit in Korea and its sources. Asia-Pacific Journal of Financial Studies 38, 211–236.

Charness, G., Gneezy, U., 2012. Strong evidence for gender differences in risk taking. Journal of Economic Behavior & Organization 83, 50–58.

Cheema, M.A., Nartea, G.V., 2014. Momentum returns and information uncertainty: Evidence from China. Pacific-Basin Finance Journal 30, 173–188.

Chen, L., Pelger, M., Zhu, J., 2023. Deep learning in asset pricing. Management Science, Ahead of Print .

Chen, Q., Hua, X., Jiang, Y., 2018. Contrarian strategy and herding behaviour in the Chinese stock market. The European Journal of Finance 24, 1552–1568.

Chen, Z., Xiong, P., Huang, Z., 2015. The asset management industry in china: its past performance and future prospects. The Journal of Portfolio Management 41, 9–30.

Chevalier, J., Ellison, G., 1999. Are some mutual fund managers better than others? cross-sectional patterns in behaviour and performance. The journal of finance 54, 875–899.

Chiang, W.C., Urban, T.L., Baldridge, G.W., 1996. A neural network approach to mutual fund net asset value forecasting. Omega 24, 205–215.

Chordia, T., Shivakumar, L., 2002. Momentum, business cycle, and time-varying expected returns. The Journal of Finance 57, 985–1019.

Chou, P.H., Wei, K.J., Chung, H., 2007. Sources of contrarian profits in the Japanese stock market. Journal of Empirical Finance 14, 261–286.

Chui, A.C., Titman, S., Wei, K.J., 2010. Individualism and momentum around the world. The Journal of Finance 65, 361–392.

Chui, A.C., Wei, K.C., Titman, S., 2000. Momentum, legal systems and ownership structure: An analysis of Asian stock markets. Unpublished Working Paper .

Clare, A., 2017. The performance of long-serving fund managers. International Review of Financial Analysis 52, 152–159.

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

Deboeck, G.J., 1994. Trading on the edge: neural, genetic, and fuzzy systems for chaotic financial markets. Wiley, New York.

DeMiguel, V., Gil-Bazo, J., Nogales, F.J., Santos, A.A., et al., 2021. Can machine learning help to select portfolios of mutual funds?

Elton, E.J., Gruber, M.J., Blake, C.R., 1996. The persistence of risk-adjusted mutual fund performance. Journal of Business , 133–157.

Estes, R., Hosseini, J., 1988. The gender gap on wall street: an empirical analysis of confidence in investment decision making. The journal of psychology 122, 577–590.

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

Fama, E.F., French, K.R., 2010. Luck versus skill in the cross-section of mutual fund returns. The journal of finance 65, 1915–1947.

Fama, E.F., French, K.R., 2012. Size, value, and momentum in international stock returns. Journal of Financial Economics 105, 457–472.

Fang, Y., Wang, H., 2015. Fund manager characteristics and performance. Investment Analysts Journal 44, 102–116.

Feng, G., Polson, N.G., Xu, J., 2018. Deep learning factor alpha. Unpublished Working Paper .

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics , 1189–1232.

Fu, T., Lin, M., 2010. Herding in China equity market. International Journal of Economics and Finance 2, 148–156.

Getmansky, M., Lo, A.W., Makarov, I., 2004. An econometric model of serial correlation and illiquidity in hedge fund returns. Journal of Financial Economics 74, 529–609.

Golec, J.H., 1996. The effects of mutual fund managers' characteristics on their portfolio performance, risk and fees. Financial Services Review 5, 133–147.

Gorton, G.B., Hayashi, F., Rouwenhorst, K.G., 2013. The fundamentals of commodity futures returns. Review of Finance 17, 35–105.

Gottesman, A.A., Morey, M.R., 2006. Manager education and mutual fund performance. Journal of empirical finance 13, 145–182.

Green, J., Hand, J.R., Zhang, X.F., 2013. The supraview of return predictive signals. Review of Accounting Studies 18, 692–730.

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

Griffin, J.M., Ji, X., Martin, J.S., 2003. Momentum investing and business cycle risk: Evidence from pole to pole. The Journal of Finance 58, 2515–2547.

Grinblatt, M., Keloharju, M., 2000. The investment behavior and performance of various investor types: a study of Finland's unique data set. Journal of Financial Economics 55, 43–67.

Grundy, B.D., Martin, J.S.M., 2001. Understanding the nature of the risks and the source of the rewards to momentum investing. The Review of Financial Studies 14, 29–78.

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

Han, C., 2022. Bimodal characteristic returns and predictability enhancement via machine learning. Management Science 68, 7701–7741.

Han, C., Shi, Y., 2022. Chinese stock anomalies and investor sentiment. Pacific-Basin Finance Journal 73, 101739.

Hanauer, M., 2014. Is Japan different? evidence on momentum and market dynamics. International Review of Finance 14, 141–160.

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

Hendricks, D., Patel, J., Zeckhauser, R., 1993. Hot hands in mutual funds: Short-run persistence of relative performance, 1974–1988. The Journal of Finance 48, 93–130.

Hon, M.T., Tonks, I., 2003. Momentum in the UK stock market. Journal of Multinational Financial Management 13, 43–70.

Indro, D.C., Jiang, C., Patuwo, B., Zhang, G., 1999. Predicting mutual fund performance using artificial neural networks. Omega 27, 373–380.

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

Jiang, X., Xu, N., Yuan, Q., Chan, K.C., 2018. Mutual-fund-affiliated analysts and stock price synchronicity: Evidence from china. Journal of Accounting, Auditing & Finance 33, 435–460.

Jones, C.S., Mo, H., 2021. Out-of-sample performance of mutual fund predictors. The Review of Financial Studies 34, 149–193.

Kacperczyk, M., Nieuwerburgh, S.V., Veldkamp, L., 2014. Time-varying fund manager skill. The Journal of Finance 69, 1455–1484.

Kang, J., Liu, M.H., Ni, S.X., 2002. Contrarian and momentum strategies in the China stock market: 1993–2000. Pacific-Basin Finance Journal 10, 243–265.

Kaniel, R., Lin, Z., Pelger, M., Van Nieuwerburgh, S., 2022. Machine-learning the skill of mutual fund managers. Technical Report. National Bureau of Economic Research.

Kaniel, R., Saar, G., Titman, S., 2008. Individual investor trading and stock returns. The Journal of finance 63, 273–310.

Lakonishok, J., Shleifer, A., Vishny, R.W., Hart, O., Perry, G.L., 1992. The structure and performance of the money management industry. Brookings Papers on Economic Activity. Microeconomics 1992, 339–391.

Li, B., Rossi, A.G., 2020. Selecting mutual funds from the stocks they hold: A machine learning approach. Available at SSRN 3737667 .

Li, H., Zhang, X., Zhao, R., 2011. Investing in talents: Manager characteristics and hedge fund performances. Journal of Financial and Quantitative Analysis 46, 59–82.

Li, J., Li, S., et al., 2018. An empirical analysis of the impact of fund manager¡¯ s personal characteristics on fund performance in china¡¯ s fund market-based on dea model and threshold panel model. International Journal of Financial Research 9, 216–226.

Liu, J., Stambaugh, R.F., Yuan, Y., 2019. Size and value in china. Journal of Financial Economics 134, 48–69.

Lo, A.W., MacKinlay, A.C., 1990. An econometric analysis of nonsynchronous trading. Journal of Econometrics 45, 181–211.

Ludwig, R.S., Piovoso, M.J., 2005. A comparison of machine-learning classifiers for selecting money managers. Intelligent Systems in Accounting, Finance & Management: International Journal 13, 151–164.

Lui, W., Strong, N., Xu, X., 1999. The profitability of momentum investing. Journal of Business Finance and Accounting 26, 1043–1091.

Madge, S., Bhatt, S., 2015. Predicting stock price direction using support vector machines. Unpublished Working Paper .

Madison Sargis, L.P.L., 2016. Women fund managers are scarce worldwide. URL: https://www.morningstar.com/articles/781996/women-fund-managers-are-scarce-worldwide.

Maxam, C.L., Petrova, M., Nikbakht, E., Spieler, A.C., 2005. Managerial characteristics and hedge fund performance. Journal of Applied Finance, Forthcoming .

Messmer, M., 2017. Deep learning and the cross-section of expected returns. Unpublished Working Paper .

Morah, C., . Is short selling banned in china? URL: https://www.investopedia.com/ask/answers/09/short-selling-china.asp. July, 2022.

Moskowitz, T.J., Ooi, Y.H., Pedersen, L.H., 2012. Time series momentum. Journal of Financial Economics 104, 228–250.

Park, K.I., Kim, D., 2014. Sources of momentum profits in international stock markets. Accounting and Finance 54, 567–589.

Peachavanish, R., 2016. Stock selection and trading based on cluster analysis of trend and momentum indicators. International MultiConference of Engineers and Computer Scientists 1, 317–321.

Powell, M., Ansic, D., 1997. Gender differences in risk behaviour in financial decision-making: An experimental analysis. Journal of economic psychology 18, 605–628.

Prather, L., Bertin, W.J., Henker, T., 2004. Mutual fund characteristics, managerial attributes, and fund performance. Review of financial economics 13, 305–326.

Pyo, U., Yong, J.S., 2013. Momentum profits and idiosyncratic volatility: the Korean evidence. Review of Accounting and Finance 12, 180–200.

Rao, Z.u.R., Ahsan, T., Tauni, M.Z., Umar, M., 2018. Performance and persistence in performance of actively managed chinese equity funds. Journal of Quantitative Economics 16, 727–747.

Ray, P., Vina, V., 2004. Neural network models for forecasting mutual fund net asset value.

Rouwenhorst, K.G., 1998. International momentum strategies. The Journal of Finance 53, 267–284.

Rouwenhorst, K.G., 1999. Local return factors and turnover in emerging stock markets. The Journal of Finance 54, 1439–1464.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature Journal 323, 533–536.

Sharpe, W.F., 1966. Mutual fund performance. The Journal of business 39, 119–138.

Sharpe, W.F., 1998. The sharpe ratio. Streetwise–the Best of the Journal of Portfolio Management 3, 169–185.

Shukla, R., Singh, S., 1994. Are cfa charterholders better equity fund managers? Financial Analysts Journal 50, 68–74.

Siganos, A., 2010. Can small investors exploit the momentum effect? Financial Markets and Portfolio Management 24, 171–192.

Sortino, F.A., Price, L.N., 1994. Performance measurement in a downside risk framework. the Journal of Investing 3, 59–64.

Su, R., Zhao, Y., Yi, R., Dutta, A., 2012. Persistence in mutual fund returns: Evidence from china. International Journal of Business and Social Science 3.

Sun, Z., Wang, A.W., Zheng, L., 2014. Hedge fund performance persistence over different market conditions. Unpublished working paper .

Wang, C., Chin, S., 2004. Profitability of return and volume-based investment strategies in China's stock market. Pacific-Basin Finance Journal 12, 541–564.

Wei, X., Xia, W., Sun, T., 2014. Research on the measurement of investor sentiment in a-share market based on bw model. Management Observer (Mandarin) 33.

Wermers, R., 1997. Momentum investment strategies of mutual funds, performance persistence, and survivorship bias. Technical Report. Unpublished working paper.

Wu, W., Chen, J., Yang, Z., Tindall, M.L., 2021. A cross-sectional machine learning approach for hedge fund return prediction and selection. Management Science 67, 4577–4601.

Yang, Y., Gebka, B., Hudson, R., 2019. Momentum effects in China: A review of the literature and an empirical explanation of prevailing controversies. Research in International Business and Finance 47, 78–101.

Zhao, X.j., Wang, S.y., 2007. Empirical study on chinese mutual funds' performance. Systems Engineering-Theory & Practice 27, 1–11.

Zhong, X., Enke, D., 2017. Forecasting daily stock market return using dimensionality reduction. Expert Systems with Applications 67, 126–139.