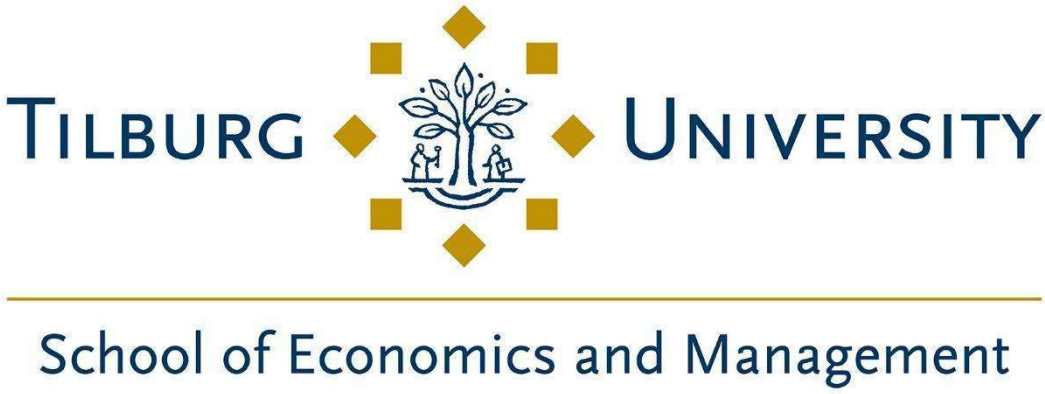

Stock Selection with Random Forest in the Chinese Stock Market: A Momentum-Based Approach

Master Thesis Finance

Author Yuhui Yang
SNR 2131398
ANR 840964

Supervisor: Denis Kojevnikov
Second reader: Gianpaolo Parise
Date: 23-06-2025

Abstract

This thesis explores the effectiveness of a pure momentum-based stock selection strategy in the Chinese A-share market using Random Forest classification models. Given the market's relatively weak-form inefficiency and dominance by individual investors, this thesis focuses exclusively on technical momentum indicators to predict relative stock performance. Stocks are classified based on forward excess returns, and portfolios are constructed by selecting the top 20 stocks with the highest predicted probability of outperformance.

Risk control is integrated into portfolio construction through the application of Minimum Variance and Minimum Maximum Drawdown weighting schemes. Empirical results show that both risk-adjusted methods significantly outperform the equally weighted benchmark portfolio, with notably improved Sharpe and Sortino ratios.

The model is further tested through a comprehensive parameter sensitivity analysis, evaluating the impact of tree number, group number, training period, trading period, and holding period on performance. The analysis highlights that intermediate values for most parameters strike the best balance between adaptability and stability.

The robustness of the model is examined under different market regimes, classified using a rolling-window approach based on relative index movements. By identifying bull, bear, and sideways periods, the strategy's adaptability to varying market conditions is tested. This regime-based evaluation further reveals that the strategy is particularly well-suited to sideways markets.

Overall, the thesis demonstrates that a classification-based momentum strategy, when combined with risk-aware portfolio construction, can consistently generate excess returns in a volatile emerging market environment. The findings offer practical implications for individual investors seeking interpretable and implementable quantitative trading frameworks.

Contents

1. Introduction

How to identify high-quality stocks in the market has long been a famous topic in financial research. Among various approaches, factor models are the most widely used frameworks to explain stock returns and have been extensively applied in both asset pricing and stock selection. Traditionally, researchers have employed linear regression models to capture the relationship between forward returns and firm-specific characteristics. However, the assumption of linearity in such models has been increasingly questioned. A growing body of literature suggests that incorporating nonlinear methods can enhance predictive accuracy and improve the effectiveness of stock selection strategies in identifying equities with superior performance. The recent proliferation of machine learning techniques provides an alternative modeling paradigm which is capable of capturing complex and nonlinear interactions between predictors and returns. Empirical evidence shows that these models can significantly improve forecasting performance. Accordingly, numerous studies in developed markets such as the U.S. have demonstrated that machine learning-based strategies can yield economically significant excess returns. As a result, many financial institutions in advanced markets have integrated machine learning into their trading operations, thereby contributing to increased market efficiency.

While the majority of research has focused on developed markets, emerging markets such as China are attracting growing attention. These markets are generally considered less efficient, and therefore potentially more exploitable using advanced predictive methods. A recent study by Yu (2024) finds that the Chinese stock market does not satisfy the weak-form Efficient Market Hypothesis, suggesting that historical price information may still contain valuable predictive signals. This finding provides the motivation for this thesis to evaluate whether known strategies, such as momentum-based models, can generate excess returns in the Chinese context.

A variety of machine learning models have been applied to the Chinese stock market, but the majority of existing studies adopt a broad feature engineering approach that integrates a wide range of fundamental and technical factors. Such approaches, although comprehensive, often suffer from high search costs and limited interpretability, particularly from the perspective of individual investors. In contrast, this thesis contributes to the real market by focusing on a narrower yet more targeted set of price-based momentum features, which remain potentially predictive given the documented inefficiencies in the Chinese market. Furthermore, by employing a transparent and relatively simple classification model, this thesis offers a

practical and accessible framework that can be readily implemented by retail investors. This not only enhances the real-world applicability of the proposed strategy, but may also promote more informed trading behaviours, ultimately contributing to the gradual improvement of market efficiency in China.

Therefore, the primary research question of this thesis is: Can a random forest model trained solely on momentum-based features produce statistically significant excess returns in the Chinese A-share market? To answer this question, several sub-questions can be explored.

The first is the application of the momentum strategy, which includes the design of the trading scheme as well as the creation of momentum factors. The literature has suggested different kinds of techniques for factor construction, including multi-window rolling returns, volatility-adjusted momentum, cumulative returns over fixed lookback periods, and skip-month adjustments to reduce short-term reversal effects. As recommended by Krauss et al. (2017), this thesis uses the multi-window rolling return approach, which increases predictive flexibility by capturing momentum signals over a range of time horizons. In order to approximate a realistic investment environment, this thesis uses a rolling training and prediction framework for the trading scheme. The strategy is intended to stay flexible in response to shifting market conditions by periodically retraining the random forest model and updating the stock pool. This dynamic structure improves the practical relevance of backtesting results and enables the model to reflect changing patterns in momentum behaviour.

The construction of the portfolio, which involves selecting the Random Forest model's hyperparameters and figuring out the weights of each chosen stock, is another crucial aspect of this study. This thesis employs a thorough parameter sensitivity analysis because ex-ante knowledge about the ideal configuration is lacking. The evaluation of each hyperparameter's marginal contribution to the overall strategy performance is made possible by the independent variation of each one while maintaining the others constant. This methodology facilitates the identification of the most influential hyperparameter settings under the rolling training and trading framework. The parameters subjected to sensitivity testing include the number of trees in the Random Forest, the number of stock groups employed for classification, the length of the training window, the duration of the trading period, and the holding period of the constructed portfolio. The distribution of portfolio weights is just as important to the ultimate success of an investment as the choice of stocks. The top 20 stocks with the highest predicted probability of falling into the highest return class are chosen by the model after it has been trained. The weights of the chosen stocks are then determined by this thesis using a risk-aware optimization process. In order to create the final portfolio, weights are specifically assigned by

minimizing predetermined risk metrics, which are history variance or maximum drawdown over the training period of the portfolio. The efficacy of these risk-controlled portfolios in producing excess risk-adjusted returns is then evaluated by benchmarking their performance against market Index and alternative weighting schemes.

To carry out the empirical analysis, this thesis employs daily trading data for all A-share stocks listed in the Chinese market, sourced from Wind, a widely recognized and authoritative financial data provider. To ensure data quality and consistency, stocks with a trading history of less than one year, as well as those labeled as "ST" (Special Treatment) due to financial distress or regulatory concerns, are excluded from the sample. No additional data preprocessing procedures are applied, given that the Random Forest model is inherently robust to large datasets and capable of handling noisy or high-dimensional input without introducing estimation bias. All stages of data handling, model implementation, and backtesting are conducted using the Python programming language, and a transaction cost of 0.16% per trade is preset.

This thesis adopts the random forest model as the core predictive algorithm. As established in the literature, while a variety of powerful machine learning methods, such as artificial neural networks and deep learning architectures, have been applied in financial forecasting tasks, a study by Ballings et al. (2015) suggests that random forest remains a widely recognized benchmark model, particularly in cross-sectional return prediction. Its ensemble structure, built upon multiple de-correlated decision trees, allows it to capture complex nonlinear relationships while maintaining robustness to noise and overfitting. Beyond its predictive capability for stock returns, random forest has consistently demonstrated superior performance compared to other classification algorithms, such as support vector machines and k-nearest neighbours, especially in high-dimensional and noisy financial datasets. Moreover, empirical evidence from Yuan et al. (2020) suggests that in the context of emerging markets, notably the Chinese stock market, random forest models exhibit enhanced robustness and generalization performance. These advantages make random forest particularly well-suited for applications involving large-scale stock selection based on momentum-related features.

In portfolio construction, it is insufficient to solely determine which stocks to include; the allocation of portfolio weights is equally crucial in shaping the return and risk characteristics of the strategy. While equal weighting serves as a natural baseline, preliminary empirical results suggest that this approach fails to consistently generate excess returns. To address this limitation, the thesis adopts a two-step approach: first, the top-ranked stocks are selected based on predicted return probabilities; second, portfolio weights are optimized by

minimizing historical risk measures. Specifically, two risk-based optimization techniques are employed: one that minimizes portfolio variance, and another that minimizes the maximum drawdown (MDD) over the training period. Comparative analysis reveals that integrating risk control into the weighting scheme substantially improves both return performance and risk-adjusted outcomes. Among the methods tested, the Minimum MDD approach outperforms others by delivering the highest excess returns while also maintaining relatively lower downside risk. However, this improvement comes at the cost of greater portfolio turnover, as evidenced by the higher magnitude of weight changes between rebalancing periods. Despite this instability, the superior performance of the MDD-based method justifies its application. Accordingly, this thesis adopts the minimum maximum drawdown approach as the default weighting strategy in subsequent portfolio construction.

To test the impact of hyperparameters to the proposed model, this thesis conducts a comprehensive hyperparameter sensitivity analysis. By systematically varying key hyperparameters, which include the number of trees in the random forest, the number of classification groups, the lengths of training and trading periods, and the portfolio holding period, the thesis evaluates how model configurations affect predictive accuracy and portfolio performance. The results reveal that strategy outcomes are sensitive to these parameter choices. Specifically, increasing the number of trees enhances model stability and performance up to a saturation point, while too many classification groups lead to overfitting and reduced robustness. Longer training period gives the model more information to make stock selection and yields higher return, while trading period should be maintained at a moderate level to ensure the model is updated. Besides, the holding period should also be at a suitable range, for shorter holding period would increase the turnover of the portfolio and introduce large transaction costs to the portfolio which can eliminate the profitability, while longer holding period would dilute the predictive power. These findings provide important guidance for hyperparameter tuning and demonstrate that careful calibration of model design is essential for achieving strong and stable performance in real-world market conditions. Finally, this thesis adopts the hyperparameters as follows: tree number of 100, group number of 5, training period of 252 days, trading period of 60 days, and holding period of 15 days.

The final empirical results confirm the effectiveness of the proposed momentum-based strategy. Over the full evaluation period, the strategy consistently outperforms the market benchmark, delivering sustained excess returns with lower drawdowns and strong overall stability. Across different market regimes, the strategy performs best in sideways markets, where relative price signals are most exploitable. It remains stable in bull markets but tends to

lag behind the index, while performance in bear markets is weaker due to higher volatility and lack of short-side exposure. These findings highlight the strategy's practical value in capturing inefficiencies within the Chinese equity market.

The empirical results of this thesis align with and extend the findings of several key studies in the field. Krauss et al. (2017) applied a range of machine learning models including deep neural networks, gradient-boosted trees, and random forests to U.S. equity markets, and they found that random forest models delivered superior predictive performance. Consistent with their results, this thesis also confirms the effectiveness of random forests in generating profitable trading signals, particularly in the context of a momentum-based stock selection framework. However, while Krauss et al. (2017) focused on the U.S. market and employed a relatively short holding horizon for statistical arbitrage, this thesis extends the analysis to the Chinese stock market, incorporating market-specific constraints of the absence of short-selling and adapting the strategy to a long-only investment framework.

The results also fill some gaps in the context of the Chinese stock market. Yuan et al. (2020) applied machine learning models to factor selection and price forecasting, concluding that random forests outperform other classifiers in terms of accuracy. While their work focused on feature importance across both fundamental and technical domains, this thesis differentiates itself by relying exclusively on technical momentum signals, and by implementing a backtesting framework with dynamic weighting. More importantly, this thesis not only demonstrates the predictive strength of random forests, but also highlights the contribution in incorporating risk-controlled portfolio construction through techniques, and this application significantly enhances the strategy's return-generating capacity.

The empirical results of this thesis demonstrate that a stock selection strategy based on random forest models and price momentum signals can effectively generate excess returns in the Chinese stock market. The incorporation of risk-based portfolio optimization further enhances the strategy's performance and robustness. Importantly, the success of this purely technical approach provides new evidence against the weak-form efficiency of the Chinese market, suggesting that historical price information remains a valuable and exploitable signal.

Importantly, the results challenge the traditional view that momentum strategies perform best in trending markets. While classic studies emphasize the role of strong directional trends in driving momentum profits, the empirical findings here reveal that the proposed strategy achieves its strongest performance during sideways market conditions in Chinese market. This divergence stems from the use of cross-sectional momentum signals and an

overall long-term downward trend, which allow the model to exploit relative performance differences even in the absence of clear market direction in relatively short-term period.

Overall, this thesis proposes a momentum-based stock selection strategy that is practically implementable. The strategy demonstrates strong empirical performance, particularly within the context of an emerging and inefficient market. Its design proves effective not only in generating excess returns but also in managing downside risk. These results provide actionable insights for investors and contribute to the broader literature on the application of data-driven methods in financial decision-making within emerging markets.

The rest of this thesis proceeds as follows. Section 2 describes the literature review. Section 3 overviews the data and software. Section 4 shows the methodology. Section 5 presents the empirical findings and, finally, Section 6 concludes.

2. Literature Review

2.1 Advancements in Modeling Techniques for Stock Selection

In the area of stock selection, first there has been a surge of interest in linear factor models, driven by the extensive development of factors capturing a wide range of technical and fundamental characteristics. One of the most famous asset pricing model was introduced by Fama and French (2015), they proposed and further developed factor asset pricing models that extend the Capital Asset Pricing Model (CAPM) by various factors, in addition to the market factor in a linear way.

However, it remains unclear whether the market behaves in a linear way and whether returns can truly be explained by linear regression or are simply driven by market anomalies. Zhu et al. (2011) questioned the linearity assumption in traditional models, and argue that while linear models are effective in capturing simple relationships, they are not good at modeling the intricate, high-order interactions among financial variables that often drive the stock performance. For stock ranking, they suggested a hybrid strategy that blends logistic regression and decision trees. Since then, this approach has gained popularity as a tool for both industry practice and scholarly research.

Later, more flexibility and variety than traditional models are offered by the subsequent development of machine learning techniques, which offer a fresh approach to investigating the connection between stock prices and company characteristics. It has been shown that powerful machine learning model classes can improve the classification and prediction efficiency of stocks. For example, Zhu et al. (2012) used decision trees (DT), Belciug and Sandita, (2017) applied artificial neural networks (ANN), Chong et al., (2017) used deep neural networks (DNN), Krauss et al. (2017) tested random forests (RF), gradient-boosted trees (GBDT), etc.

These developments inspire this thesis to investigate non-linear and data-driven methods, particularly random forest models, to better capture complex relationships in stock selection tasks.

2.2 Random Forest in Stock Selection

It has already been established that random forests demonstrate strong predictive performance and offer robustness to noise and minimal parameter tuning requirements. Krauss et al., (2017) applied deep neural networks, gradient-boosted trees, and random forests to construct a short-term statistical arbitrage strategy for U.S. equities. They came to the conclusion that random forests perform better than gradient-boosted trees and deep neural networks in their application. Similarly, Gu et al. (2020) discovered that “shallow” learning

(including boosted trees and random forest) outperforms "deep" learning (including neural networks), which deviates from the typical conclusions in other fields. This indicates that the random forest model is regarded as a standard technique in cross-sectional stock selection.

Random forest model also outperforms various classification algorithms. Ballings et al., (2015) evaluated the performance of random forests (RF), Support Vector Machine (SVM), K-nearest Neighbors (KNN), and neural networks, for predicting the direction of stock price movements, and concluded that random forests consistently outperform other classifiers in terms of both accuracy and robustness. Kavin et al. (2018) conducted a comparative study evaluating the performance of random forests (RF), Support Vector Machine (SVM), and Extreme Learning Machine (ELM) for intrusion detection. Their results show that Random Forest consistently performed better than the other two models in terms of classification accuracy and robustness.

With respect to the application of the random forest model, broad literature suggests that classification-based framework may yield more superior prediction over regression-based approaches in financial forecasting when applying machine learning models. The study Enke and Thawornwong (2005) demonstrate that predicting the directional movement of stock returns often yields more superior risk-adjusted returns compared to forecasting exact return magnitudes. This benefit is especially noticeable in markets like China that are noisy and have a lot of variation.

These findings reinforce the rationale for employing Random Forest as a suitable classification model in financial prediction tasks, particularly when dealing with complex and high-variance datasets such as those observed in emerging markets. Prior studies suggest that a classification-based framework, which focuses on predicting the direction or relative rank of excess returns, can outperform traditional regression approaches, which is designed to estimate exact return magnitudes. This is especially relevant in less developed and noisier markets, such as Chinese market, where overfitting and instability are common challenges. These findings also inspire this thesis to adopt classification models to reach greater robustness and more favorable risk-adjusted performance.

2.3 Price Momentum Strategy in Stock Selection

The most basic and essential kind of momentum strategy is price momentum, which is frequently applied in stock selection and asset pricing. Although it is not a new idea, it is still a straightforward and straightforward investment strategy. The most well-known research is carried out by Jegadeesh and Titman (1993), who established the existence of a momentum effect in U.S. stock returns. Their work is among the most significant empirical studies in asset

pricing. The findings indicated that while previous losers underperformed, stocks that had performed well over the previous three to twelve months (referred to as "winners") tended to continue outperforming in the ensuing three to twelve months. The Capital Asset Pricing Model and other conventional risk-based models were unable to account for the notable positive abnormal returns that this momentum strategy produced. A large amount of research on return anomalies and behavioral finance was spurred by their findings, which also called into question the Efficient Market Hypothesis.

Behavioral finance offers a crucial theoretical framework for comprehending momentum strategies that goes beyond risk-based explanations. A prominent view is that investor underreaction to new information contributes to the persistence of stock price trends. As suggested by Barberis et al. (1998), investors may initially underweight unexpected news due to cognitive biases such as conservatism, leading to a delayed price adjustment. Daniel et al. (1998) argued that overconfidence and biased self-attribution can cause investors to overreact to private signals while underreacting to public information, reinforcing momentum effects. Hong and Stein (1999) suggested that positive feedback trading, a form of trend-chasing behavior, can exacerbate price continuation patterns. When investors extrapolate past price movements into the future, they may collectively push prices further away from fundamental values, thereby contributing to short-term and medium-term momentum.

There was a growing body of research which investigated how historical stock returns can be systematically transformed into momentum factors. Jegadeesh and Titman (1993) implemented the momentum strategy in a direct way by sorting stocks based on their past J-month returns and forming long-short portfolios that buy past winners and sell past losers, held for K months. Carhart (1997) introduced a refined version of the momentum factor in his well-known four-factor model by excluding the most recent month's return from the cumulative return calculation. This version defines the momentum signal as the return from month $t-12$ to $t-2$, effectively skipping one month. Krauss et al. (2017) constructed momentum features by calculating the past stock returns over multiple rolling time windows including short-term windows and monthly-scale windows, allowing the model to capture return dynamics at different frequencies.

These theoretical perspectives offer further justification for the use of momentum signals in data-driven stock selection frameworks adopted in this thesis. And this thesis follows the momentum design of Krauss et al. (2017), using multiple rolling return windows to construct price-based momentum features for use in the random forest model.

2.4 Stock Selection and Model Application in the Chinese Market

A growing amount of research indicates that the weak-form efficiency assumed in many developed markets may not apply to the Chinese stock market. Notably, Yu (2024) examined the predictability of stock returns through a combined liquidity-based trading strategy and provided strong evidence against the weak-form efficiency assumption of the Chinese stock market. Yu's findings support the broader view that the Chinese market is still in a developmental stage, which creates persistent mispricing that can be exploited by data-driven strategies. These findings justify the use of momentum-based strategy to generate alpha in the Chinese stock market.

The application of machine learning models in the Chinese stock market has also gained significant attention in recent years. Yuan et al. (2020) employed a range of machine learning techniques, including Support Vector Machines (SVM), Random Forests (RF), and Artificial Neural Networks (ANN), to perform factor selection and stock price trend prediction, which utilized both fundamental and technical indicators. Their results indicated that the random forest model consistently outperforms other methods in terms of improving both feature selection effectiveness and prediction accuracy.

Existing studies in Chinese market tend to emphasize broad-based factor selection across a wide array of inputs, often combining both fundamental and technical indicators, in which some factors are not easy for individual investors to get. Relatively few studies have focused specifically on technical momentum signals as the sole basis for machine learning-based stock selection in the Chinese market. Thus, this thesis adopts a pure momentum strategy which only uses public prices of the stocks.

2.5 Risk Control in Portfolio Construction

The finance literature has long stressed the importance of incorporating risk control into portfolio construction, and researchers have looked into a variety of risk-aware constraints and strong optimization frameworks.

Jagannathan and Ma (2003) demonstrated that the imposition of seemingly restrictive constraints, such as prohibiting short-selling or bounding portfolio weights, can substantially reduce portfolio risk and improve performance. They argue that such constraints act as a form of implicit regularization, preventing the optimizer from overreacting to estimation noise and reducing the sensitivity of portfolio weights to extreme parameter values. These constraints, while theoretically "wrong," are practically beneficial by stabilizing the solution and reducing outlier exposure.

Building on these insights, recent literature has explored more tailored risk measures. Chekhlov et al. (2005) proposed a formal framework for incorporating maximum drawdown as a constraint or objective in portfolio optimization. Their study demonstrated that optimizing portfolios with respect to drawdown can yield allocations that are more consistent with the preferences of risk-averse investors, particularly over multi-period horizons. Compared to volatility-based methods, the drawdown-based approach provides better control over cumulative loss paths, reflecting a more intuitive and behaviorally relevant dimension of risk.

DeMiguel et al. (2009) critically compared the performance of optimized portfolios against simple diversification strategies, where the former refers to the Mean-variance optimization proposed by Markowitz, and the latter is the 1/N equally weighted approach. Their findings reveal that, in many realistic settings, simple diversification outperforms optimized portfolios unless robust estimation techniques or risk controls are applied. This observation highlights the role of assigning the weights of different stocks in a portfolio in improving robustness.

Based on the regulation of Chinese stock market and findings above, this thesis excludes short selling and testes different weight optimization methods while constructing the portfolio.

3. Data and Software

3.1 Data Source

The entire range of stock price data from every listed company on the Chinese stock market is used in this thesis. Because it is the most popular reference index in China and its index futures are among the most actively traded, the CSI 300 Index is selected as the benchmark. All data are based on daily frequency, including stock prices, market capitalization, and turnover rates.

The dataset is sourced from the Wind Financial Terminal, which is one of the most authoritative databases in China, offering comprehensive coverage of Chinese equity market data.

3.2 Data Pre-processing

To enhance the robustness and reliability of the sample, this thesis excludes stocks that have been listed for less than one year, thereby avoiding distortions associated with the IPO-induced listing premium. In addition, stocks marked as "ST" (Special Treatment) are also removed due to their heightened financial risk and abnormal trading behavior, which could introduce excessive noise into the model training process.

Regarding data preprocessing, no normalization, winsorization, or feature transformation is applied prior to model training. This decision is based on the well-documented capacity of the Random Forest algorithm to handle high-dimensional datasets and maintain reliable classification performance without requiring strict distributional assumptions. However, to address missing values in the input features, this thesis adopts a simple forward-filling approach, whereby each missing value is imputed using the immediately preceding value in the same time series. This approach preserves the temporal structure of the data while minimizing information loss.

After applying these exclusion criteria and preprocessing steps, the resulting dataset consists of 4,683 unique stocks with daily closing prices spanning from October 8, 2019, to December 31, 2024. The dataset is free of missing values in the core variable of closing price after forward-filling was applied.

3.3 Stock Pool Construction

According to Section 3.2, after excluding certain stocks, over 4,000 stocks still remain in the training set. Using all of them as model inputs would lead to prohibitively long training times. Therefore, prior to each model training cycle, a subset of stocks is selected to form a stock pool, which is then used for both model training and stock selection.

Mimicking the methodology used in index construction, stocks are ranked based on their market capitalization and turnover rate, and the top 300 stocks are selected to constitute the stock pool. This stock pool serves as the basis for both model training and subsequent portfolio construction, ensuring a balance between data richness, liquidity, and computational tractability.

3.4 Software

All data handling will be conducted using Python.

4. Methodology

4.1 Design of Rolled Training and Trading Scheme

In this thesis, a rolled training and trading framework is adopted to evaluate the performance of the trading strategy over time, and all back-tests performed and discussed bear a transaction cost preset of 0.16% for each trade. A training period refers to a fixed number of business days used to fit the model. Following each training period, a trading period is defined, during which the trained model is applied to generate output and construct portfolios. The precise lengths of both the training and trading periods are treated as hyperparameters, meaning that their values are not predetermined but will be optimized based on performance through the later parameter sensitivity analysis. The overall backtesting window is a 600-day length window which extends to December 31, 2024. The total trading window is divided into 10 non-overlapping sub-periods, and each sub-period has 60 trading days and serves as an independent trading evaluation set.

For each selected stock, a holding period of n business days is applied, during which the stock remains in the portfolio before being replaced. The holding period is another hyperparameter that would be assessed, as different durations may affect the trade-off between responsiveness and transaction cost.

To prevent the use of future information when constructing the model, a gap of n business days is introduced between the end of each training period and the beginning of its corresponding trading period. This temporal separation ensures that the model does not inadvertently use information from the future, thereby preserving the integrity of the out-of-sample evaluation.

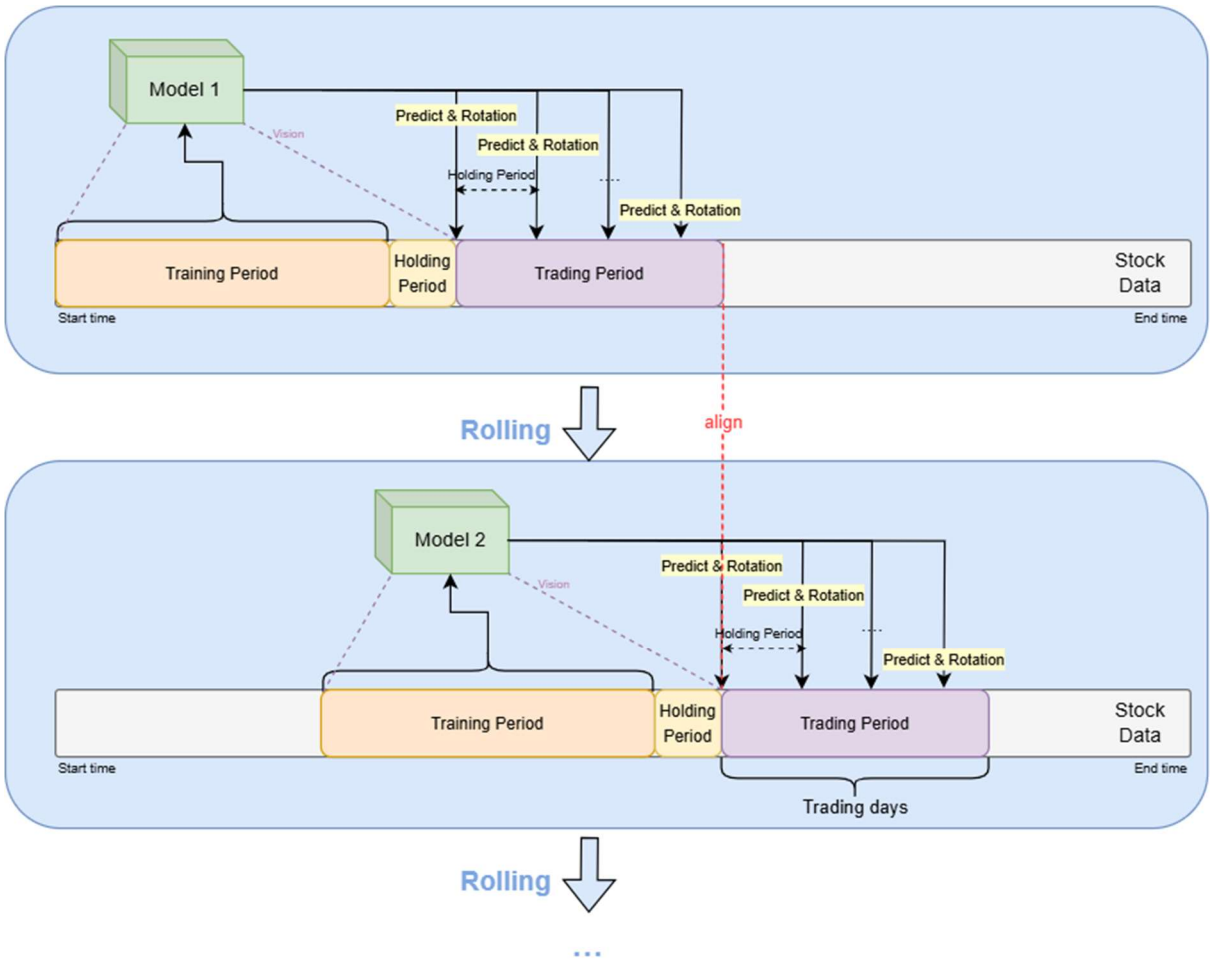

Figure 1: Rolled training and trading scheme

4.2 Momentum Strategy

The momentum strategy can be dated back to the research of Jegadeesh and Titman (1993), based on a well-documented phenomenon in equity markets. They found that stocks that have performed well in the recent past (referred to as "winners") tend to continue to outperform those that have performed poorly (referred to as "losers") over a subsequent period. Momentum strategy has since become a distinctly and widely recognized investment style across the U.S. and other well developed stock markets. The momentum effect is most pronounced over intermediate horizons, ranging from three to twelve months.

Among various types of momentum, such as price and earnings momentum, this thesis focuses exclusively on price momentum. As suggested by empirical findings by Chan et al. (1996), the price momentum effect is generally stronger and more persistent than that of earnings momentum. Unlike classical momentum strategies that construct long-short portfolios by taking long positions in past winners and short positions in past losers, this thesis adopts a novel approach by embedding the momentum signal into a machine learning classification

framework. Specifically, the historical return features serve as inputs to a random forest model, which is trained to predict whether a stock will fall into the top return group during the upcoming holding period. Based on the model's predicted probabilities, a subset of the most promising stocks is selected for portfolio inclusion.

4.3 Random Forest Model

Random Forest is an emerging machine learning technique that belongs to the powerful class of models exemplified by tree-based algorithms and neural networks.

According to Kavin et al. (2018), the original Random Forest model is well-suited for handling a large number of variables in the dataset and has been shown to produce unbiased estimates for classification problems in real-world applications, including those in the financial domain. Therefore, this thesis adopts the standard Random Forest construction method proposed by Breiman (2001) without modifications.

In principle, a random forest consists of an ensemble of deep but uncorrelated decision trees, each trained on a different random sample of the original dataset. The construction process is as follows: for each individual tree, a random subset of the training data is first drawn (with replacement) to form the bootstrap sample. A decision tree is then grown on this sample up to a maximum depth J_{RF}. At each split in the tree, a random subset of m_{RF} features is selected from the total p available features, and the best split is chosen based on a criterion like Gini impurity or information gain. This procedure is repeated n_{RF} times to generate a forest of n_{RF} trees. The final output of the model is obtained by aggregating the predictions of all trees using majority voting in classification tasks. The computational complexity of training the random forest can be approximately estimated as $O(nRF \cdot p \cdot n_{ins} \cdot logn ins)$, where n_{ins} represents the number of instances (samples) in the training dataset.

Three key hyperparameters are typically tuned to optimize model performance, they are: n_{RF} (the number of trees in the forest), J_{RF} (the maximum depth of each tree) and m_{RF} (the number of features randomly selected at each split).

In this thesis, J_{RF} is set to be unlimited, allowing each tree to grow until either all nodes are pure or each terminal node contains fewer than two samples. For feature subsampling, this thesis follows James et al. (2023) and sets the standard practice of $m_{RF} = \sqrt{p}$. The influence of the number of trees n_{RF} on both classification accuracy and out-of-sample performance is then systematically explored through sensitivity analysis.

4.4 Input and Output of the model

To implement the momentum-based strategy, this thesis generates the input features for each training period based on historical price movements. Following the methodology proposed by Krauss et al. (2017), a series of momentum indicators are constructed to capture the recent performance of individual stocks.

Specifically, for each stock $s \in \{1,2,\ldots n\}$, let P_t^s denote the closing price of stock s at time t, and $m \in \{\{1,\ldots 20\} \cup \{40,60,\ldots,240\}\}$, the momentum factor $Mom_{t,m}^s$ at time t over a look-back window of m days can be defined by

$$Mom_{t,m}^s = \frac{P_t^s}{P_{t-m}^s}.$$

Here the thesis focuses on the stock return of the last 20 days and then switch to the preceding 11 months, which in total are 31 factors. These features are designed to capture different return patterns over various time horizons.

Unlike many traditional approaches that normalize or winsorize input variables, this thesis applies the raw momentum values directly. This is because the random forest algorithm is known for its robustness to outliers and its ability to handle unscaled input features, while still providing reliable and unbiased predictions in classification settings.

Consistent with findings of Enke and Thawornwong (2005), classification models have demonstrated superior performance compared to regression models when applied to financial data. Therefore, this thesis adopts a classification framework where the prediction target is not the exact return, but the class to which the stock belongs, based on its forward excess return.

Hence, the input and output of the model can be summarized as follows.

Input (factors) is a $u \times v$ matrix, where u is the sample number and is calculated by the number of stocks multiply the number of days, v is the number of factors.

Output (probability of being classified as the best class) is a $u \times 1$ matrix. First, let $P^S = (P_t^s)_{t \in T}$ denote the close price series of stock $s \in \{1,2,\ldots,n\}$ over the time period T, at each time point t, the forward return $R_{t,m}^s$ of stock s over the subsequent m trading days is calculated as

$$R_{t,m}^s = \frac{P_{t+m}^s}{P_t^s} - 1.$$

Second, let P^I denotes the CSI index price series, the benchmark return $R^I_{t,m}$ over the same period is calculated by

$$R^I_{t,m} = \frac{P^I_{t+m}}{P^I_t} - 1.$$

Third, the excess return for stock s over the period $[t, t+m]$ is defined as

$$ER^s_{t,m} = R^s_{t,m} - R^I_{t,m}.$$

Last, after computing the excess returns for all stocks at time t, the stocks are ranked in descending order based on their $ER^s_{t,m}$. The ranked stocks are then evenly divided into N groups, where N is also a hyperparameter that will be tested in the sensitivity analysis.

4.5 Risk Control and Portfolio Construction

There are some assumptions while assigning the weights of the stocks. The thesis incorporates a realistic market constraint by excluding short selling from the trading strategy. This decision is informed by both practical and theoretical considerations. On the one hand, short selling is heavily restricted or unavailable to retail investors in the Chinese stock market, making long-only strategies more implementable in practice. On the other hand, several empirical studies have shown that the exclusion of short selling can enhance strategy robustness and reduce sensitivity to misclassification errors, particularly in noisy or inefficient markets.

To enhance the robustness and profitability of the momentum-based stock selection framework, this thesis incorporates risk-aware optimization methods into the portfolio construction process. Following the prediction step, where the trained random forest model ranks stocks based on the likelihood of belonging to the optimal return class, the top 20 stocks are selected to form the candidate portfolio. Instead of assigning equal weights, the thesis adopts an optimization procedure to assign stock-level weights that minimize specific risk objectives. Specifically, two commonly used risk metrics, variance and maximum drawdown, are employed to inform the allocation of weights among the selected stocks.

On one hand, the thesis adopts a traditional approach for each rebalancing period, which applies a minimum variance optimization approach. This method constructs the portfolio by minimizing the overall variance of portfolio returns over the training window while imposing constraints to ensure full investment and prohibit short selling. Specifically, let $\mathrm{w} \in \mathbb{R}^{\mathrm{n}}$ denote the portfolio weight vector and $\Sigma \in \mathbb{R}^{\mathrm{n}\times\mathrm{n}}$ the covariance matrix of historical daily returns of the selected stocks, the optimization problem can be formulated as

$$\min_{w} w^{\top}\Sigma w,$$

where $1^{\top}w = 1$, and $w \geq 0$.

The covariance matrix Σ is estimated using the return matrix $R \in \mathbb{R}^{t \times n}$, where each row represents a trading day and each column corresponds to a selected stock. Let $\bar{R} \in \mathbb{R}^{1 \times n}$ denote the row vector of mean returns over the training window of T trading days, the Σ is defined by

$$\Sigma = \frac{1}{T-1}(R - \bar{R})^{\top}(R - \bar{R}).$$

On the other hand, the thesis also adopts the minimum maximum drawdown strategy. Recognizing that investors often care more about potential losses than volatility alone, this method aims to minimize the historical maximum drawdown of the portfolio. The optimization follows the framework of Chekhlov et al. (2005), where a linear programming approach is used to minimize drawdown risk, under the same constraints of full investment and no short selling. Specifically, for a given weight vector w and historical return matrix $R \in \mathbb{R}^{t \times n}$, the cumulative return time series C_t is calculated as

$$C_t = \prod_{\tau}^{t}(1 + R_{\tau} \times w).$$

The maximum drawdown is then calculated as

$$MDD(w) = \max_{1 \le t \le T} \left(\frac{\max_{1 \le s \le t} C_s - C_t}{\max_{1 \le t \le T} C_t} \right).$$

Then, the optimization problem is therefore non-convex and involves minimizing the MDD with respect to w, subject to the standard portfolio constraints. It can be defined as

$$\min_{w} MDD(w),$$

where $1^{\top}w = 1$ and $w \geq 0$.

Both optimization methods are applied out-of-sample at each rebalancing point using a rolling window approach, ensuring that only past information is used in weight determination. The resulting weights are applied to the top-20 stocks to construct the portfolio for the upcoming holding period.

To evaluate the effectiveness of different risk control approaches, this thesis compares the performance of portfolios constructed under three weighting schemes: Equal Weight, Minimum Variance, and Minimum Maximum Drawdown. The evaluation is based on cumulative net asset value (NAV), daily return distributions, and a set of standard risk-adjusted performance metrics, including the Sharpe Ratio, Sortino Ratio, and Maximum Drawdown. Besides, the turnover of the portfolio using different risk control approaches would also be examined.

The Sharpe Ratio measures how much excess return can be earned per unit of total risk. Let $\bar{R}_p$ denote the mean daily return of the portfolio, R_f denote the risk-free rate, and σ_p is the standard deviation of daily returns. The Sharpe Ratio can be calculated as

$$Sharpe\ Ratio = \frac{\bar{R}_p - R_f}{\sigma_p}.$$

The Sortino Ratio refines the Sharpe Ratio by considering only downside risk instead of total risk. Let σ_d denote the downside deviation calculated from negative return observations, it can be defined as

$$Sortino\ Ratio = \frac{\bar{R}_p - R_f}{\sigma_d},$$

where $\sigma_d = \sqrt{\frac{1}{T}\sum_{t=1}^{T} min\left(R_{p,t} - R_f, 0\right)^2}$.

The Calmar Ratio evaluates performance relative to drawdown risk and is calculated as

$$Calmar\ Ratio = \frac{R_{annual}}{MDD}.$$

The portfolio turnover is analyzed by calculating the average L1 norm of weight changes between consecutive periods and is calculated as

$$Turnover_t = \sum_{i=1}^{n} \left| w_{i,t} - w_{i,t-1} \right|.$$

4.6 Hyperparameters Sensitivity Test

To evaluate the robustness of the proposed momentum-based classification strategy, this thesis undertakes a comprehensive hyperparameter sensitivity analysis situated within a rolling backtesting framework. This analysis serves two primary purposes: first, to examine how the variation in model parameters influences both predictive accuracy and investment performance; and second, to identify parameter settings that yield stable and superior results, particularly under the conditions of the Chinese stock market, which is characterized by high volatility and structural inefficiencies.

The sensitivity analysis adopts a ceteris paribus approach, where each hyperparameter is independently varied while holding all other parameters constant. This design isolates the individual impact of each parameter, thereby facilitating a more precise understanding of its marginal effect on model behavior and trading outcomes. The parameters under investigation include the number of trees used in the random forest model, the number of stock groups used for classification and ranking, the length of the training window, the trading window duration, and the portfolio holding period. Each parameter is tested across a range of values that are theoretically motivated and practically relevant for financial modeling.

For every parameter configuration, the trading strategy is executed over the entire test period. During each iteration, two return series are recorded: the net asset value (NAV) of the portfolio and the hedged NAV, which accumulates the portfolio's daily excess return over the selected market benchmark. These series are plotted to enable a visual comparison of performance across alternative settings. To complement the visual evidence, a series of commonly used performance indicators are calculated, including annualized return, maximum drawdown, Sharpe ratio, Sortino ratio, and Calmar ratio. These metrics provide a well-rounded perspective on how each configuration influences not only the level of return but also the magnitude and frequency of downside risk.

In addition to return-based metrics, model accuracy is evaluated through the out-of-bag (OOB) score, an internal validation method built into the random forest algorithm (Breiman, 2001). For each rolling training window, the OOB score reflects the model's in-sample classification accuracy without requiring a dedicated holdout dataset. Specifically, during training, each decision tree in the forest is fitted on a bootstrap sample of the original data, leaving approximately one-third of the observations "out-of-bag." These OOB samples are then passed through their respective trees to generate predictions, and the aggregate prediction across all trees where a sample was OOB is compared to the true class label. Let N denote the total number of training samples, y_i denote the true label, $\hat{y}_i^{oob}$ denote the majority vote prediction from trees where sample i was not included during training, and $\mathbb{I}(\cdot)$ denote the indicator function, the OOB score is computed as

$$OOB\ Score = 1 - \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(\hat{y}_i^{oob} \neq y_i\right).$$

The average OOB score across all sub-periods is used to gauge the model's predictive performance over time, offering a robust proxy for generalization ability under changing market conditions.

By systematically analyzing the influence of each hyperparameter on both predictive accuracy and portfolio performance, this methodology provides a rigorous foundation for optimizing the strategy. It ensures that the final parameter configuration is not only theoretically sound but also practically effective in enhancing both the stability and profitability of momentum-based strategies in the Chinese stock market context.

4.7 Market Regime Classification

To better understand the behavior of the momentum strategy under different market conditions, this thesis classifies market states into bull, bear, and sideways regimes. The classification is based on price dynamics of the CSI 300 Index. While the academic literature

offers various methods for regime detection, this thesis follows the peak-trough framework proposed by Pagan and Sossounov (2003), which is widely adopted in empirical studies.

Specifically, a rolling window approach is used: for each trading day, the past 60 trading days are examined to detect relative local extrema. A bull market is defined if the current index price is more than 10% above the lowest price in the past 60 trading days, q bear market is defined if the current index price is more than 10% below the highest price in the past 60 trading days, otherwise, the regime is classified as sideways. The definition are

$$Bull\ if\ \frac{P_t - m\quad(P_{t-60:t})}{min(P_{t-60:t})} \geq 10\%, Bear\ if\ \frac{P_t - max(P_{t-60:t})}{max(P_{t-60:t})} \leq -10\%.$$

While Pagan and Sossounov (2003) used a ±20% threshold to define bull and bear markets, this setting is found to be overly conservative in the context of the Chinese equity market. As illustrated in **Table 1**, applying a ±20% threshold leads to a near-complete dominance of sideways classifications, with only 3 days labeled as bear markets and less than 100 days as bull markets. In contrast, the ±10% threshold provides a more balanced segmentation: 346 bull days, 221 bear days, and 646 sideways days. This distribution ensures sufficient observations within each regime to allow meaningful statistical comparisons of strategy performance. Then the analysis focuses on the longest observed bull, bear, and sideways periods within the dataset.

Table 1: Number of days when applying different thresholds

	±20% threshold	±10% threshold
Sideways	1,117	646
Bull	93	346
Bear	3	221

5. Main Results

5.1 Analysis of Risk Control Methods

The initial design of this thesis focused on constructing an equally weighted portfolio composed of the 20 stocks with the highest predicted probability of belonging to the top-performing class. This simple approach aimed to test whether machine learning-based stock selection alone could consistently generate excess returns. However, the preliminary empirical results revealed that the equal-weighted strategy delivered unsatisfactory performance, suggesting that naïvely allocating capital without accounting for risk may limit the effectiveness of the predictive model.

In response to these findings, the thesis introduces an additional layer of portfolio optimization by incorporating risk control into the weighting scheme. Although this adjustment was not part of the original research design, subsequent analysis showed that managing risk can significantly improve both the return and stability of the portfolio. This motivates the use of a risk-aware weighting approach, as formally presented in the methodology section.

As shown in **Figure 2**, both risk-based portfolios significantly outperform the benchmark and the equal-weighted portfolio in terms of cumulative Net Asset Value. The portfolio constructed using the Minimum Maximum Drawdown approach exhibits the strongest performance, peaking at over 1.4× initial NAV and maintaining a higher level throughout most of the investment horizon. The Minimum Variance strategy also delivers notable outperformance relative to both the benchmark and the equal-weight scheme, with a smoother upward trajectory and visibly lower drawdown periods.

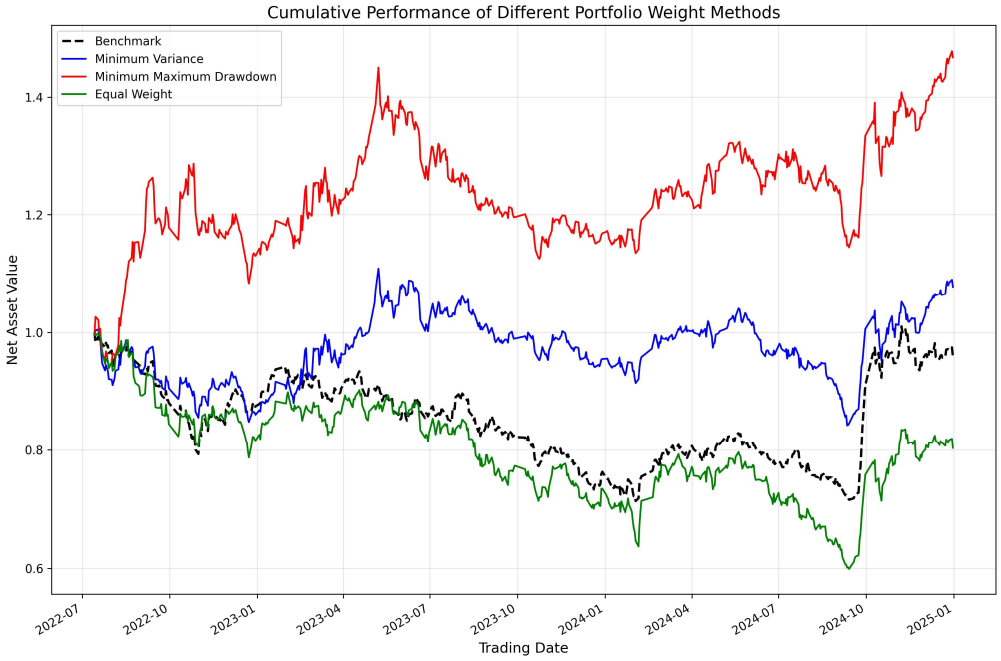

Figure 2: Cumulative Performance of Different Portfolio Weight Methods

Figure 3 presents the boxplots of daily returns across the three methods. Both the Minimum MDD strategy and Minimum Variance achieve higher median and upper quartile daily returns, while Minimum MDD strategy exhibits higher volatility. In contrast, the equal-weighted portfolio not only has a lower median return but also shows a wider interquartile range and more pronounced negative outliers, indicating higher tail risk.

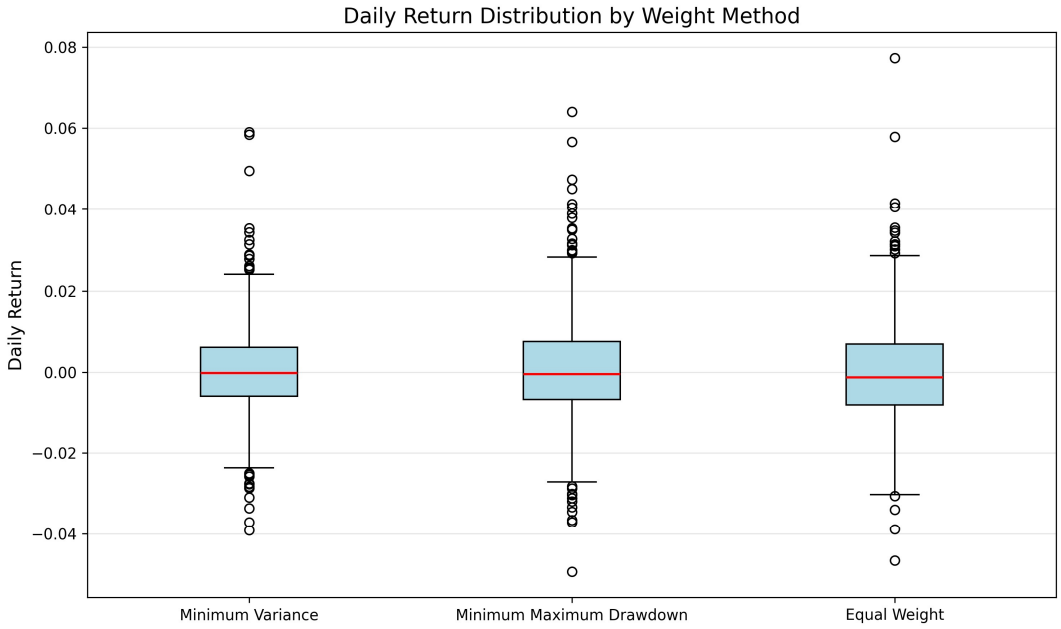

Figure 3: Daily Return Distribution by Different Weight Methods

While risk-based portfolios improve return metrics, they may introduce instability in portfolio composition. **Figure 4** plots the average magnitude of portfolio weight changes across rebalancing periods. The Minimum MDD strategy, although most profitable, exhibits the largest fluctuations in weight allocation, with a weight change of more than 0.9 all the time. This suggests that the strategy frequently reallocates capital in response to market conditions, which could incur higher transaction costs in practice. In comparison, the Minimum Variance strategy strikes a balance between performance and stability, with a moderate average weight change of 0.6-0.7. The Equal Weight portfolio, while underperforming, remains the most stable with the lowest weight adjustment magnitude.

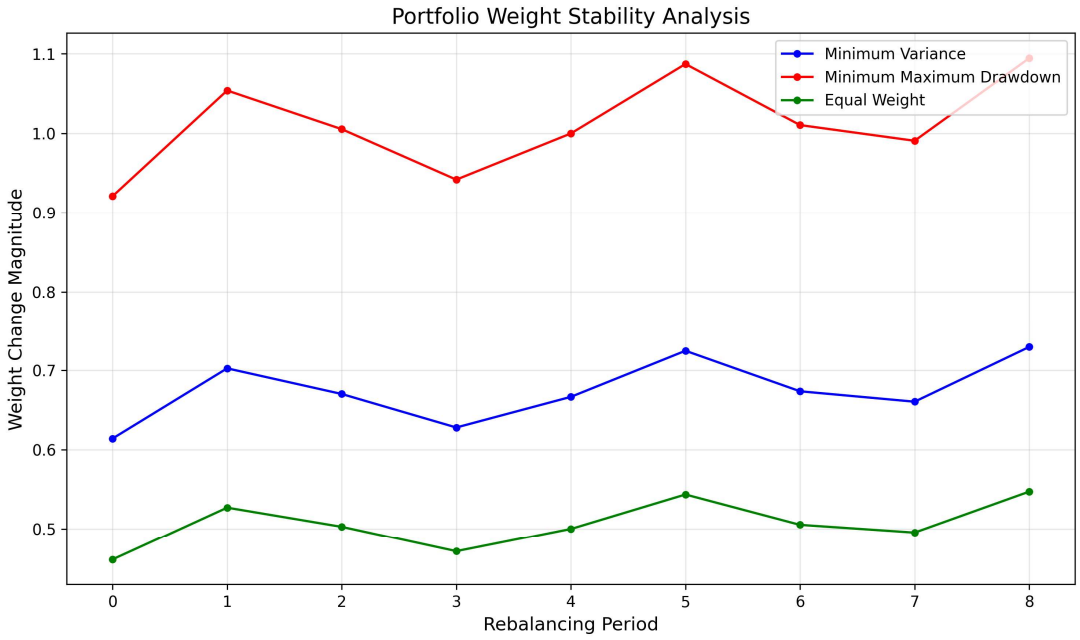

Figure 4: Portfolio Weight Stability Analysis

Figure 5 presents a comparative visualization of key risk adjustment metrics across the three portfolio weighting methods. The results clearly indicate that both Minimum Variance and Minimum Maximum Drawdown strategies outperform the Equal Weight method across all metrics. This provides strong evidence that incorporating risk-based optimization not only mitigates downside risk but also improves the efficiency of return generation.

Specifically, the Minimum Maximum Drawdown strategy achieves the highest profitability with good risk adjustment metrics and lowest maximum drawdown. These findings suggest that the following sections would adopt the method of minimizing MDD in the portfolio construction.

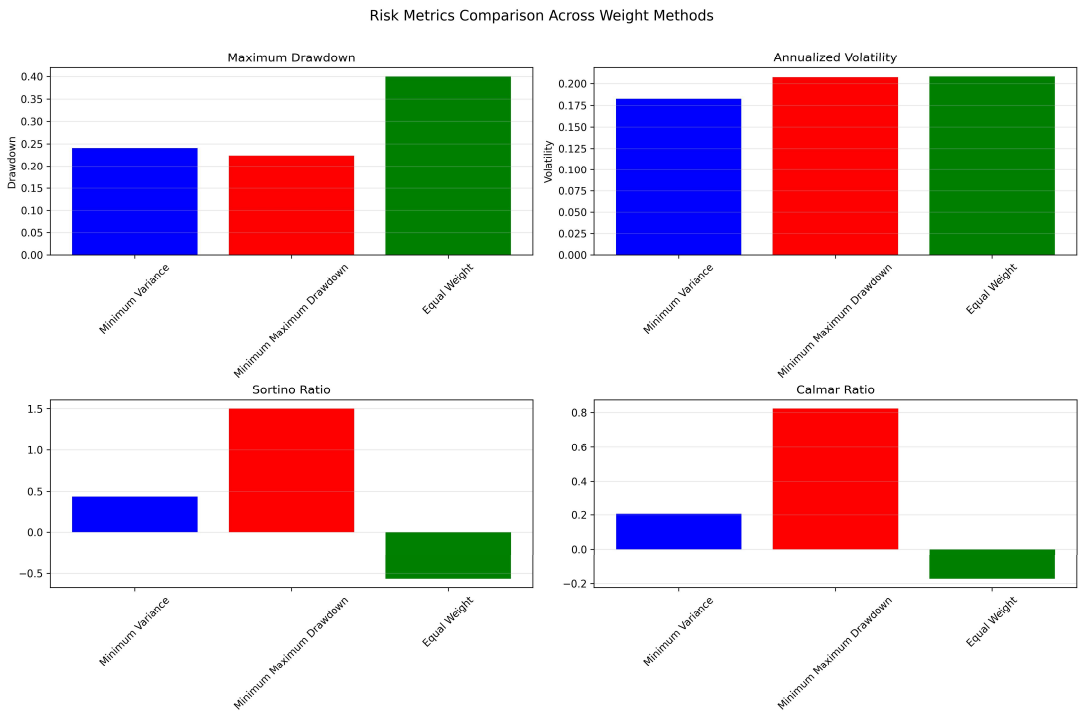

Figure 5: Risk Metrics Comparison Across Different Weight Methods

5.2 Dependence of Strategy Performance on Model Parameters

The thesis thoroughly examines how much the strategy's performance depends on modal parameters in order to examine the Random Forest model's classification efficacy on the Chinese stock market and the strategy's viability in taking advantage of excess return.

5.1.1 Tree number

The tree number controls how many decision trees are included in the random forest ensemble. This parameter plays a crucial role in determining the model's stability and predictive accuracy. A larger number of trees typically leads to more stable and accurate predictions, as the model can better average out the variance from individual trees, thereby reducing overfitting. However, increasing the number of trees also raises computational cost and training time, which becomes particularly relevant in rolling schemes with frequent model retraining.

The results presented in **Figure 6** demonstrate that increasing the number of trees enhances both the predictive accuracy and portfolio performance of the Random Forest model. As the number of estimators increases, the strategy's cumulative Net Asset Value and excess returns improve notably, with diminishing marginal gains observed beyond 100 trees. In particular, the configuration using 100 trees achieves the best overall trade-off between return

and risk-adjusted metrics, including the highest Sharpe and Sortino ratios, while maintaining competitive performance in terms of maximum drawdown.

Although the out-of-bag score continues to improve slightly when increasing the number of trees to 120, the incremental improvement is marginal relative to the additional computational cost incurred, especially under a rolling retraining framework. By contrast, using 40 trees or 60 trees results in significantly lower classification accuracy, weaker excess returns, and less favorable risk profiles.

Therefore, the configuration of 100 decision trees is identified as the optimal setting for this strategy.

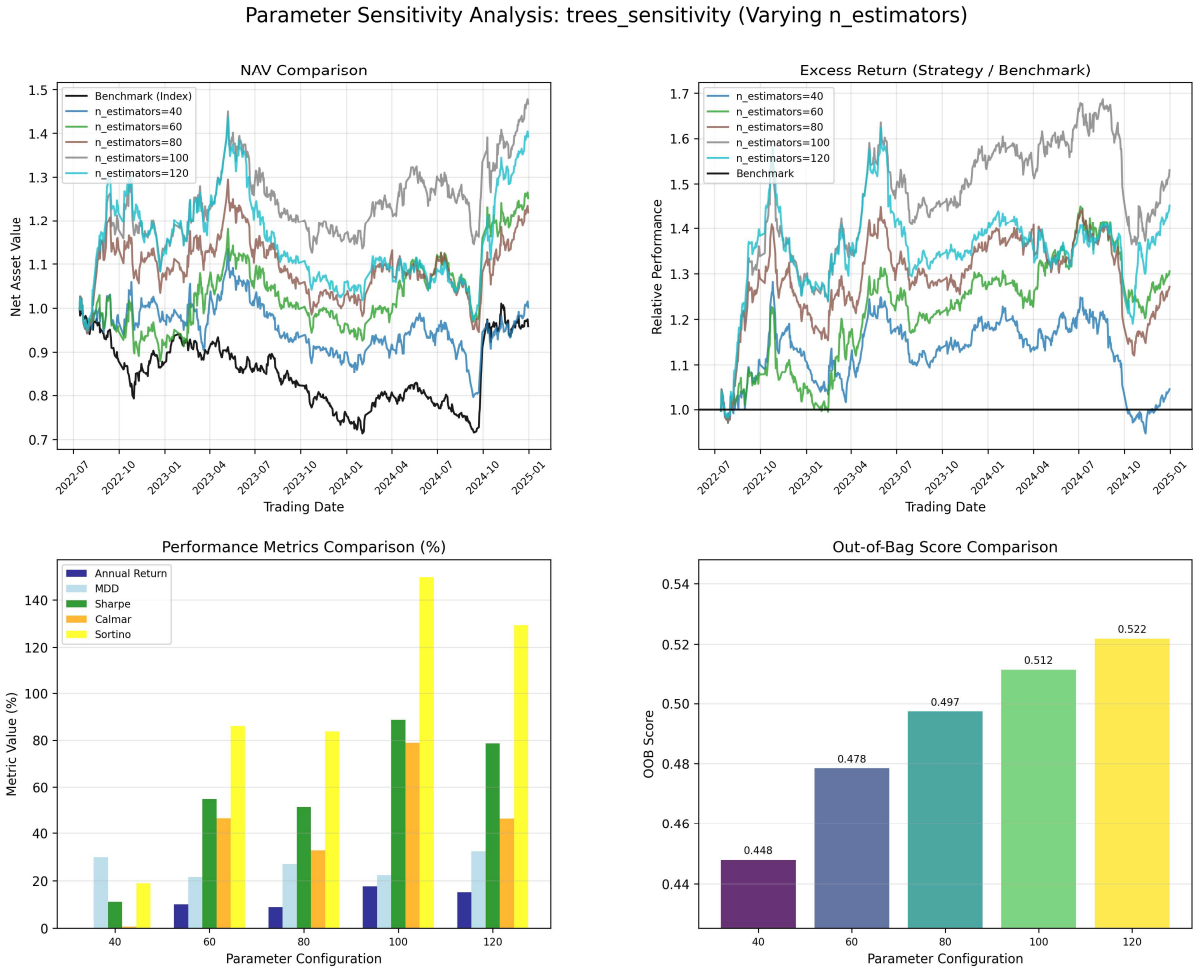

Figure 6: The strategy's performance is dependent on the number of trees, which is set at {40, 60, 80, 100, 120}. (a) The relationship between strategy performance and the number of trees as indicated by the net asset value portfolio; (b) The relationship between hedged net asset value and the number of trees; (c) The relationship between the net asset value portfolio estimators and the number of trees; and (d) The relationship between the Oob score and the number of trees.

5.1.2 Group number

The number of groups refers to how the stocks are categorized based on their forward excess returns during model training. This grouping determines the classification labels for the random forest model. A higher number of groups allows for finer granularity in distinguishing stock performance levels, which may help the model learn more nuanced patterns. However, too many groups can also lead to imbalanced classes and make the classification task more difficult, possibly reducing model accuracy and increasing noise.

The empirical results in **Figure 7** reveal that the number of groups used to define classification labels has a substantial impact on model performance and risk-adjusted returns. Among all configurations, dividing the stocks into 5 groups consistently yields the highest cumulative NAV and excess returns, along with the best performance across key metrics such as Sharpe, Sortino, and Calmar ratios. This setting also achieves the highest out-of-bag score, indicating superior in-sample classification accuracy.

While increasing the number of groups theoretically allows for finer resolution in performance ranking, the results suggest that configurations with 10 or more groups suffer from a significant decline in both predictive accuracy and return performance. This may due to the increased class imbalance and noise introduced by overly granular groupings, which makes it harder for the model to distinguish meaningful differences across classes. In particular, the poorest outcomes are observed 20, where the strategy exhibits severe drawdowns and negative annual returns.

In summary, the five-group classification strikes the optimal balance between granularity and class stability.

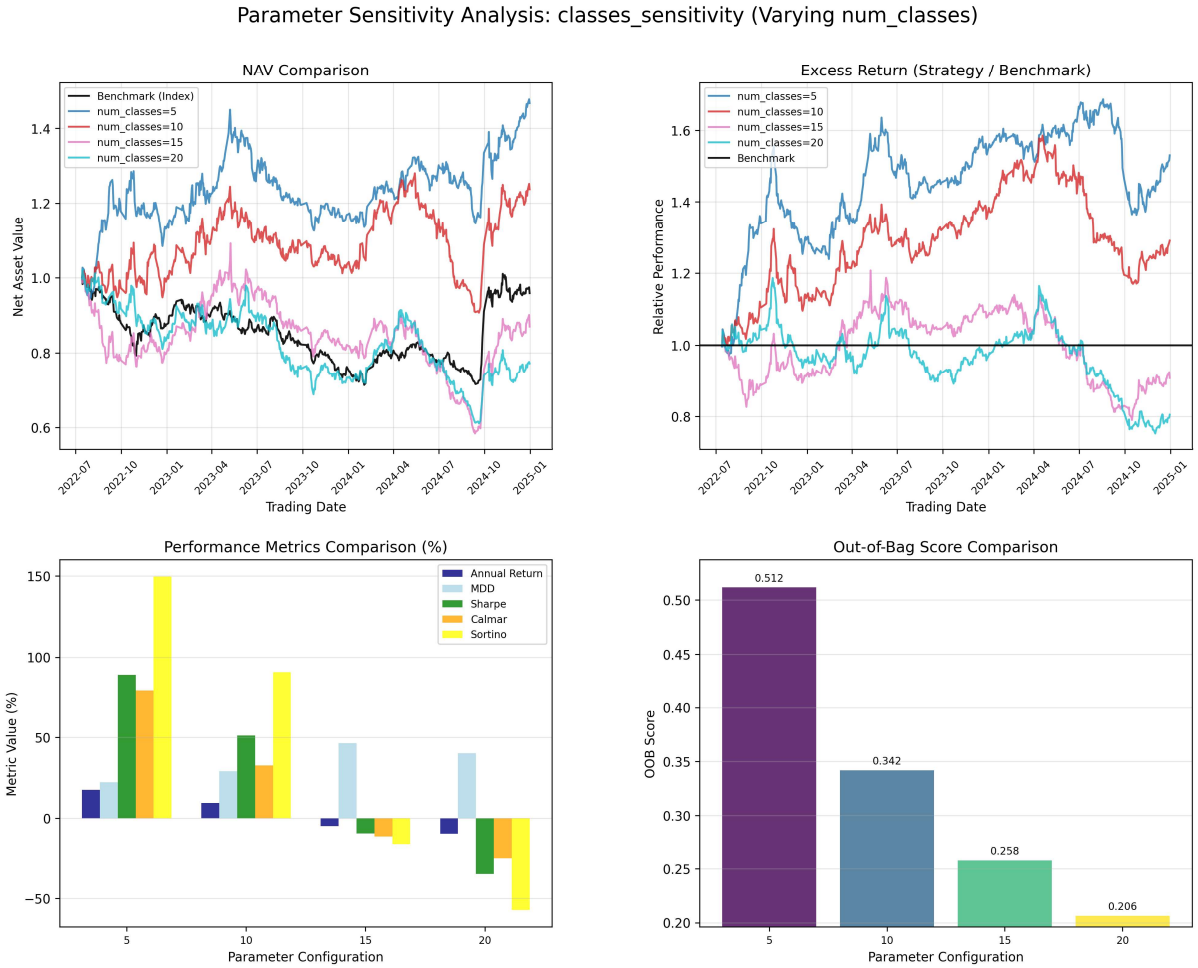

Figure 7: The strategy's performance is dependent on the number of groups, which is set at {5, 10, 15, 20}. (a) The relationship between strategy performance and the number of groups as indicated by the net asset value portfolio; (b) The relationship between hedged net asset value and the number of groups; (c) The relationship between the net asset value portfolio estimators and number of groups; and (d) The relationship between the Oob score and the number of groups.

5.1.3 Training Period

The training period refers to the length of historical data (measured in trading days) used to train the Random Forest model before each trading round. A longer training period allows the model to learn from a larger dataset, potentially capturing more stable and long-term patterns. However, it may also include outdated information that no longer reflects current market dynamics. On the other hand, a shorter training period focuses more on recent data, which may be more relevant to current market conditions but also more volatile and noisier.

The results indicate that extending the training period significantly enhances overall portfolio performance. Among all tested configurations, the model trained with 252 days of historical data achieves the highest cumulative NAV and excess return, along with superior

risk-adjusted metrics such as the Sharpe, Sortino, and Calmar ratios. These outcomes suggest that a full-year training window enables the Random Forest model to capture more stable and representative patterns in stock behaviour, which ultimately translates into more robust predictions and trading performance.

Although shorter training periods result in slightly higher out-of-bag classification scores, their trading performance is notably inferior. This reflects a trade-off between in-sample accuracy and real-world effectiveness: while shorter periods may better fit recent data, they are more susceptible to overfitting market noise and fail to generalize across different market conditions. Conversely, the 252-day configuration exhibits a slight reduction in OOB score but delivers the most consistent and profitable strategy over time.

Therefore, this thesis adopts 252 trading days as the optimal training period.

Figure 8: The strategy's performance is dependent on the training period, which is set at {75, 125, 175, 252}. (a) The relationship between strategy performance and the training period as indicated by the net asset value portfolio; (b) The relationship between hedged net asset value and training period; (c) The relationship

between the net asset value portfolio estimators and training period; and (d) The relationship between the Oob score and the training period.

5.1.4 Trading Period

The trading period refers to the length of time (in trading days) during which the trained Random Forest model is used to make predictions and select stocks before being retrained with updated data. A longer trading period reduces the frequency of model retraining, which helps to limit transaction costs and computational overhead. However, it may also lead to performance degradation if the model fails to capture rapidly changing market dynamics. Conversely, a shorter trading period increases the frequency of model updates, potentially improving adaptability to evolving market conditions and enhancing prediction accuracy.

In this thesis, the total number of trading days is fixed at 600 to ensure comparability across different configurations. As such, a shorter trading period implies that the entire 600-day horizon is divided into more frequent retraining intervals. This design allows for evaluating how frequently the model needs to be updated to maintain optimal performance under realistic market volatility, while holding the total evaluation period constant.

Based on the results shown in **Figure 9,** the length of the trading period has a clear impact on the strategy's performance. Among all configurations, a trading period of 60 days achieves the most favorable outcomes in both absolute and risk-adjusted terms. It delivers the highest annual return, as well as the best Sharpe, Sortino, and Calmar ratios, while maintaining a moderate level of maximum drawdown. This suggests that updating the model approximately every three months strikes a good balance between adaptability to changing market conditions and stability in model predictions.

Shorter trading periods of 20 and 40 days allow for more frequent model retraining, which may help the model stay responsive to short-term market shifts. However, these configurations also introduce higher model turnover and potentially more noise, which weakens the overall performance. Meanwhile, a longer trading period of 80 days results in lower returns and less favorable risk-adjusted metrics, which may be due to the model becoming stale over time and failing to incorporate timely market updates.

Importantly, the out-of-bag scores remain relatively stable across all settings, indicating that in-sample classification accuracy is not significantly affected by how frequently the model is retrained. This reinforces the idea that the performance variation is mainly driven by the

model's responsiveness to changing market dynamics during the out-of-sample trading phase. Thus, a 60-day trading period offers the best trade-off between predictive freshness and transaction efficiency in this strategy.

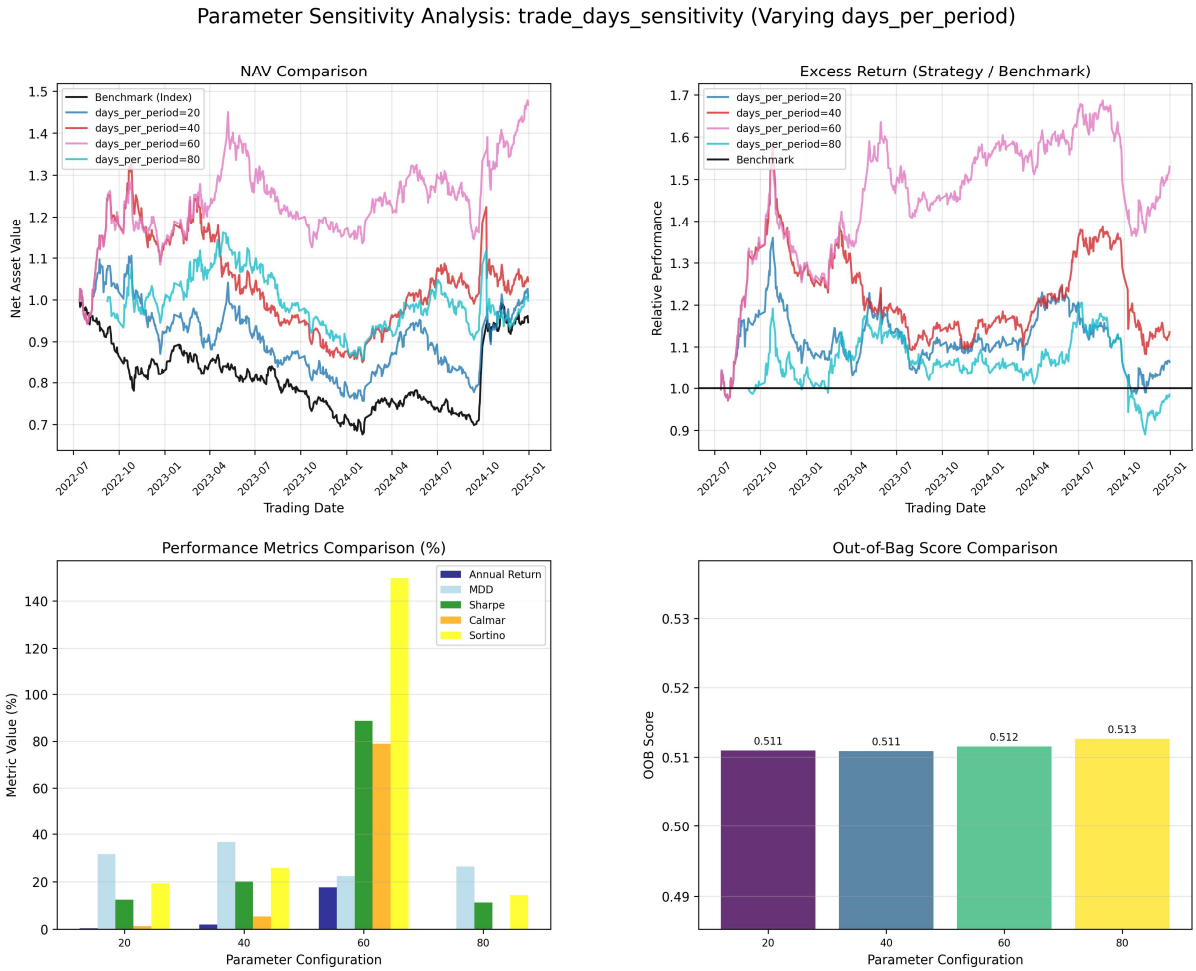

Figure 9: The strategy's performance is dependent on the trading period, which is set at {20, 40, 60, 80}. (a) The relationship between strategy performance and the trading period as indicated by the net asset value portfolio; (b) The relationship between hedged net asset value and trading period; (c) The relationship between the net asset value portfolio estimators and trading period; and (d) The relationship between the Oob score and the trading period.

5.1.5 Holding Period

The holding period refers to the number of consecutive trading days for which a selected stock remains in the portfolio before it is sold or replaced in the next rebalancing cycle. The holding period directly influences the turnover rate and trading frequency of the portfolio. A shorter holding period allows the strategy to quickly respond to changes in stock predictions, which can enhance responsiveness to short-term market movements. However, it may also lead to higher transaction costs and portfolio instability. On the other hand, a longer holding period

reduces trading frequency and costs, but may dilute the predictive power of the model if market conditions shift rapidly or the momentum signal weakens over time.

As illustrated in **Figure 10**, the holding period has a substantial impact on the performance of the strategy. Among the tested configurations, a holding period of 15 trading days emerges as the optimal setting, achieving the highest annual return, the best Sharpe and Sortino ratios, and the lowest maximum drawdown. This suggests that rebalancing the portfolio every 15 trading days provides an effective balance between capturing short-term momentum signals and maintaining portfolio stability.

Shorter holding period of 10 days tends to produce unstable results. Although it allows for more rapid responsiveness, it also results in higher turnover and potentially excessive sensitivity to daily price fluctuations. This is reflected in its worse performance across almost all metrics. Conversely, longer holding periods of 20 and 25 days lead to diminished return metrics and increased drawdown risk, which means the predictive power of short-term momentum signals deteriorates over time and becomes less aligned with actual price movements.

Notably, the out-of-bag score increases with the holding period, reaching its peak at 25 days. This indicates that longer horizons may help the classifier achieve more accurate in-sample predictions. However, this higher classification accuracy does not translate into better out-of-sample trading performance, reinforcing the idea that predictive sharpness needs to be aligned with the practical timing of signal execution.

Overall, a 15-day holding period best harnesses the predictive signals from the model.

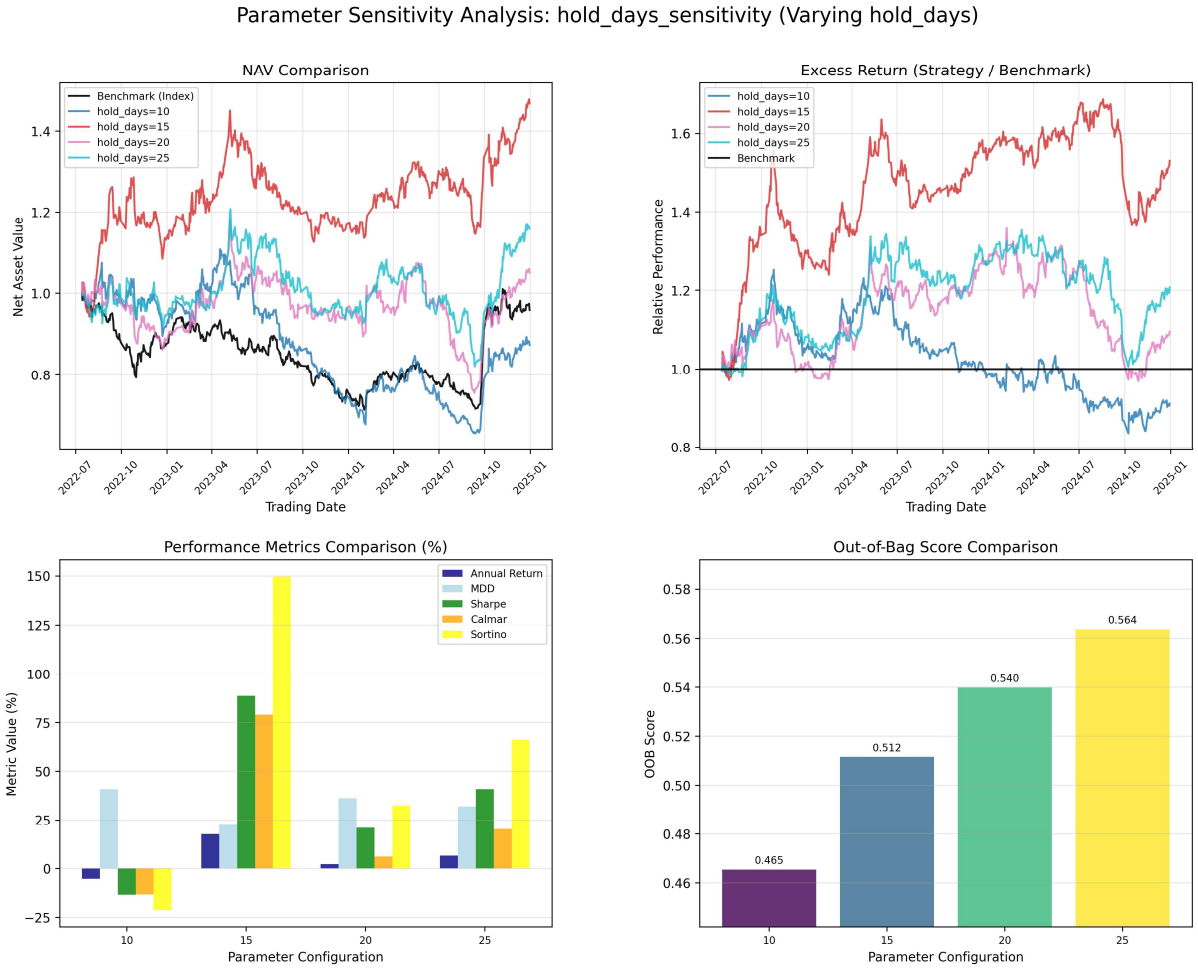

Figure 10: The strategy's performance is dependent on the holding period, which is set at {10, 15, 20, 25}. (a) The relationship between strategy performance and the holding period as indicated by the net asset value portfolio; (b) The relationship between hedged net asset value and holding period; (c) The relationship between the net asset value portfolio estimators and holding period; and (d) The relationship between the Oob score and the holding period.

5.3 Overall Performance Comparison

By employing the parameters provided in the previous sections (**Table 2**), the comparison of the performance is as follows.

Table 2: Modal parameters for the momentum strategy

Parameter	Value
Number of Trees	100
Number of groups	5
Training period	252
Trading period	60
Holding period	15

Figure 11 plots the cumulative net asset value (NAV) of the strategy and the benchmark index over the full backtesting period. The strategy exhibits consistent outperformance, achieving a steadily rising NAV trajectory while the benchmark experiences prolonged decline. **Figure 12** shows the cumulative excess return. The excess return steadily increases throughout most of the sample, peaking around mid-2024 before experiencing a temporary contraction. Even after the contraction, the strategy recovers and preserves a strong relative advantage, ending the period with substantial alpha.

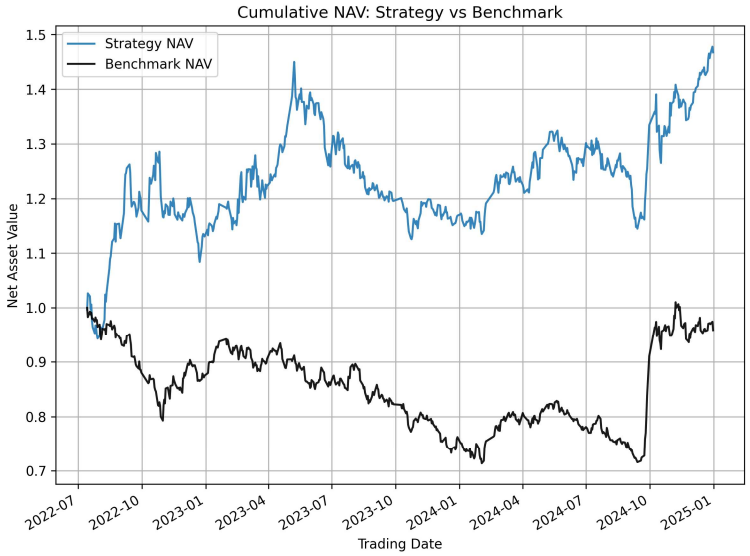

Figure 11: Comparison of the momentum strategy and benchmark performances

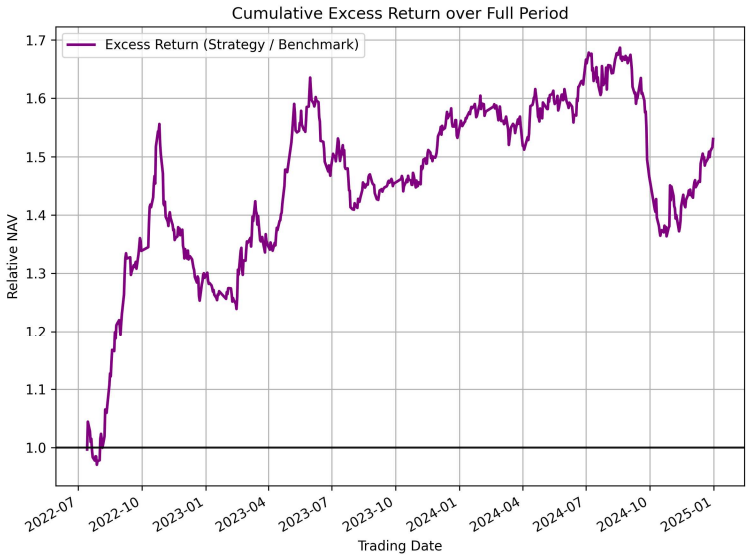

Figure 12: Cumulative excess return of the momentum strategy

5.4 Performance under Different Market Regimes

To investigate the robustness of the proposed momentum strategy, the performance across different market regimes is analysed. As illustrated in **Figure 13**, the different market condition is classified, the average daily returns for bull, bear and sideway market are 0.21%, -0.30% and -0.04%.

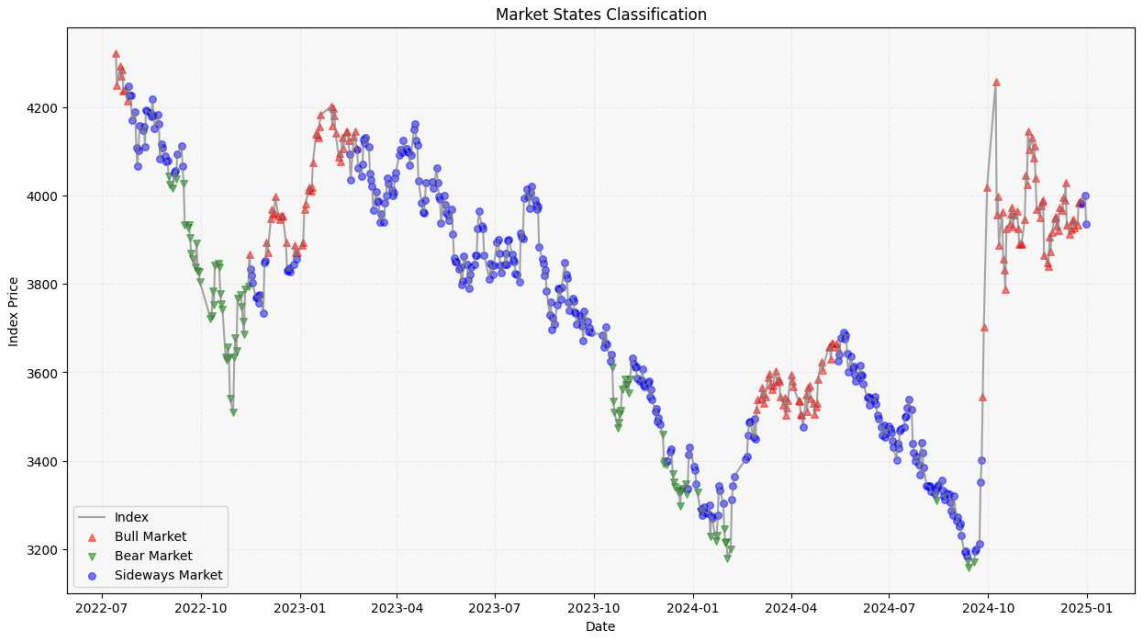

Figure 13: Market states classification of the CSI 300 Index using ±10% threshold over 60-day rolling window

During bull market periods, the momentum strategy demonstrates overall stability but fails to consistently generate significant excess returns over the benchmark. As shown in **Figure 14**, across three major bullish intervals, the strategy tends to follow market trends but only exhibits limited alpha generation. In the first bull phase from 2022-12-30 to 2023-02-16, the strategy outperforms the index by a clear margin, achieving a return of 13.04% compared to the benchmark's 5.73%, with a positive excess return of 7.31%. However, this strong outperformance is not replicated in the other two periods. In the second bull window from 2024-02-29 to 2024-04-12, both the strategy and the benchmark record negative returns with –3.93% and –1.14%, respectively, suggesting a weak or unstable market uptrend. The excess return is slightly negative, which is –2.79%, and the overall volatility remains modest at 15.98%. In the third and longest bull phase from 2024-09-26 to 2024-12-27, the strategy achieves an 11.92% return, closely trailing the index's 12.29%, resulting in a marginal underperformance of –0.36%. Nevertheless, excess return volatility remains controlled throughout, and there is no indication of large drawdowns or destabilizing losses.

During bear market periods, the strategy exhibits unstable performance and is prone to significant drawdowns. As shown in **Figure 15**, in the first bear phase from 2022-09-15 to 2022-11-15, the strategy suffered a substantial decline, with a return of –19.35% while the index fell by only –4.00%. The sharp underperformance was accompanied by extremely high volatility, reaching 56.36%, indicating that the strategy failed to manage risk effectively during sustained downward trends. In the second bear phase from 2023-10-18 to 2023-11-06, the overall market movement was limited. The strategy delivered a small gain of 0.35%, slightly trailing the index which returned 0.61%. Although excess return was negative, the performance gap remained narrow, and volatility was moderate. The third bear phase occurred from 2023-12-13 to 2023-12-27. The strategy declined by 1.31% compared to the index's loss of 0.99%. The excess return again was negative, but the magnitude of both return and volatility was low.

During sideways market conditions, the momentum strategy demonstrates its strongest relative performance. As shown in **Figure 16**, across three representative periods, the strategy not only outperforms the index but also maintains relatively stable volatility levels. In the first sideways period from 2022-07-26 to 2022-09-01, the strategy achieves a strong return of 17.39%, while the benchmark declines by 4.76%. This results in an excess return of 22.15%, the highest observed across all market regimes. Despite moderately elevated volatility at 25.77%, the strategy exhibits a clear upward NAV trajectory. During the second and longest sideways phase, from 2023-02-23 to 2023-10-18, the strategy generates a return of –5.43%, significantly outperforming the index, which declines by –12.02%. The excess return of 6.58% accumulates gradually, with temporary drawdowns that are recovered toward the end. Although strategy volatility is relatively high at 27.02%, its excess return curve remains above zero for most of the period, indicating consistent defensive positioning. In the third period, from 2024-05-15 to 2024-08-114, both the strategy and benchmark decline, but the strategy's loss of –9.96% slightly exceeds the index's –8.74%, resulting in a mild excess return of –1.22%. Even in this less favorable case, the drawdown remains within a controlled range, and volatility is moderate at 20.38%.

In summary, the regime-specific performance analysis reveals that the momentum strategy exhibits distinct behavior under different market conditions. In bull markets, the strategy tends to track the index with relatively low drawdown, but its ability to generate sustained excess returns is limited. In bear markets, the strategy's behavior becomes more volatile and less predictable. While it remains stable in shallow declines, it suffers from significant losses during deep drawdowns. In contrast, sideways markets present the most

favorable setting for the strategy. Across all observed flat market regimes, the model delivers either strong or resilient excess returns, benefiting from cross-sectional momentum patterns even when the overall index lacks direction.

Interestingly, the fact that the strategy performs best in sideways markets appears, at first glance, to contradict conventional wisdom. Traditional momentum strategies are typically designed to thrive in strongly trending markets, where price persistence supports the core logic of trend-following. However, the momentum approach implemented in this thesis deviates from the classical form. Rather than relying on broad market direction, the strategy uses cross-sectional momentum, identifying relative winners and losers within the universe of stocks. Even in a range-bound market where the index shows little movement, individual stocks can diverge significantly in short-term trajectories.

Another contributing factor is that, over the full sample period, the overall market shows a mild downward trend, albeit with limited short-term volatility. Since market regimes are defined using a 60-day rolling window, these persistent but gradual declines are not fully captured by the classification rule and are often labeled as sideways. The strategy which uses some long-term momentum factors can still generate alpha through relative momentum, even though it is identified as sideways market.

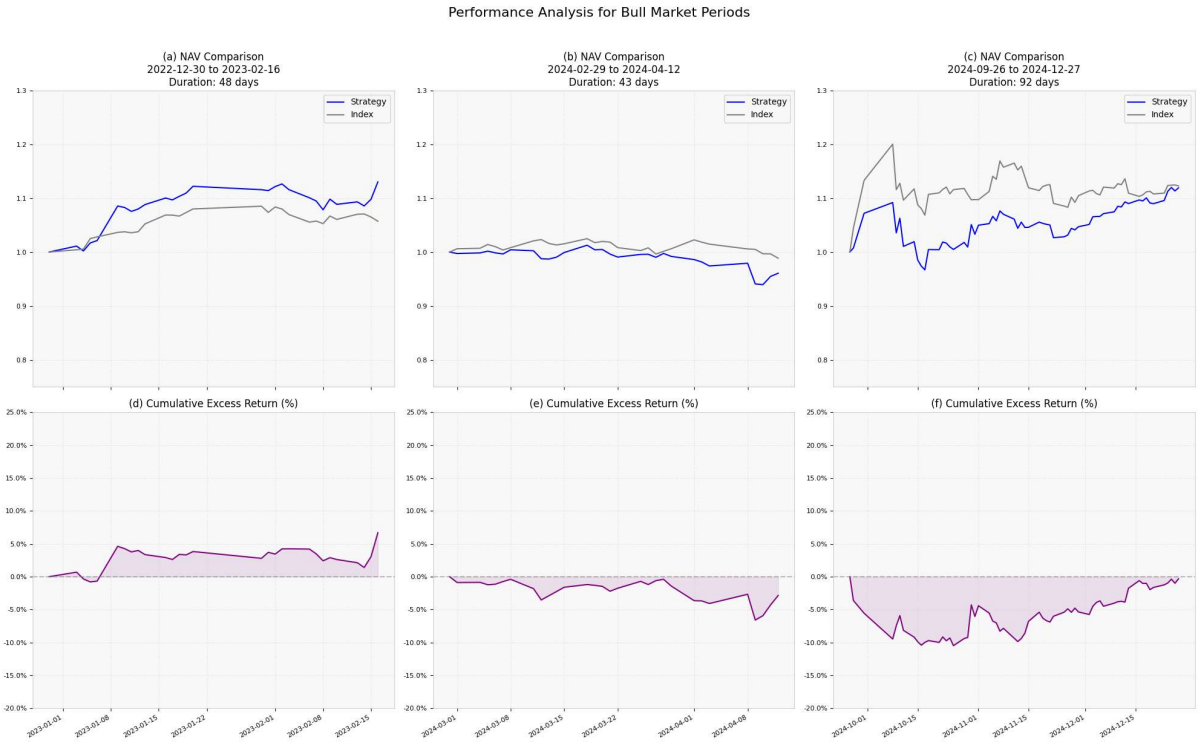

Figure 14: Performance analysis during the three longest bull market periods

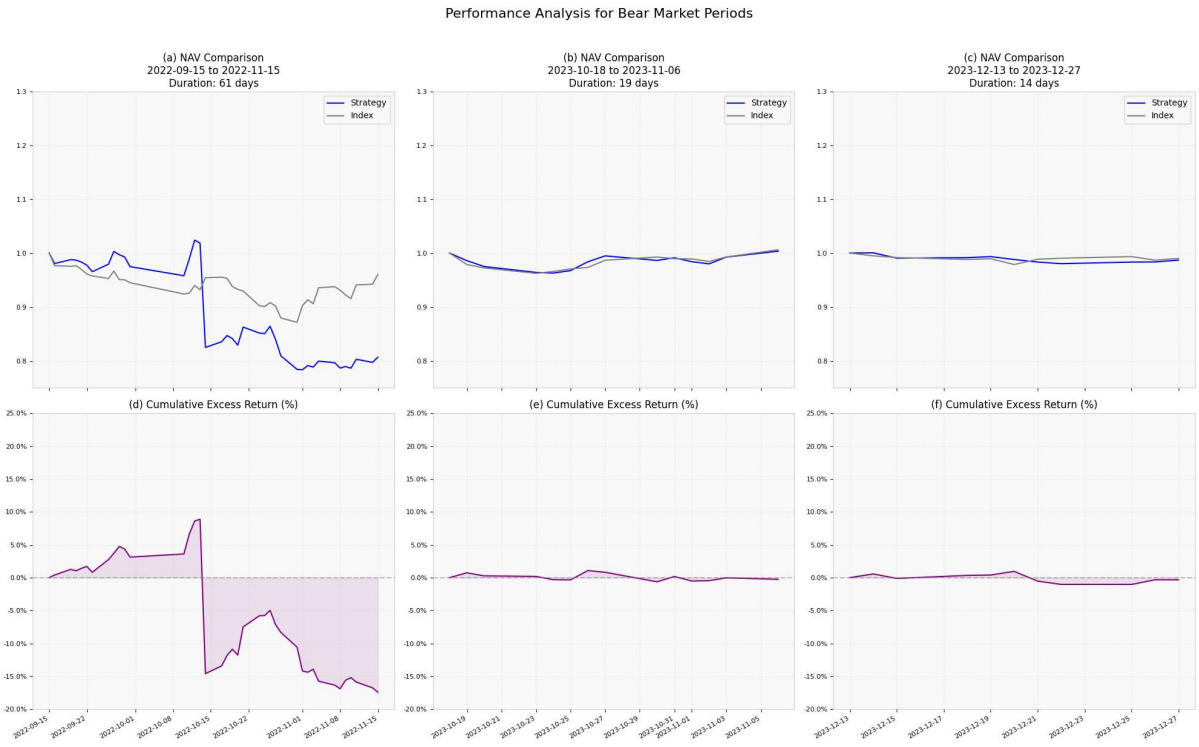

Figure 15: Performance analysis during the three longest bear market periods

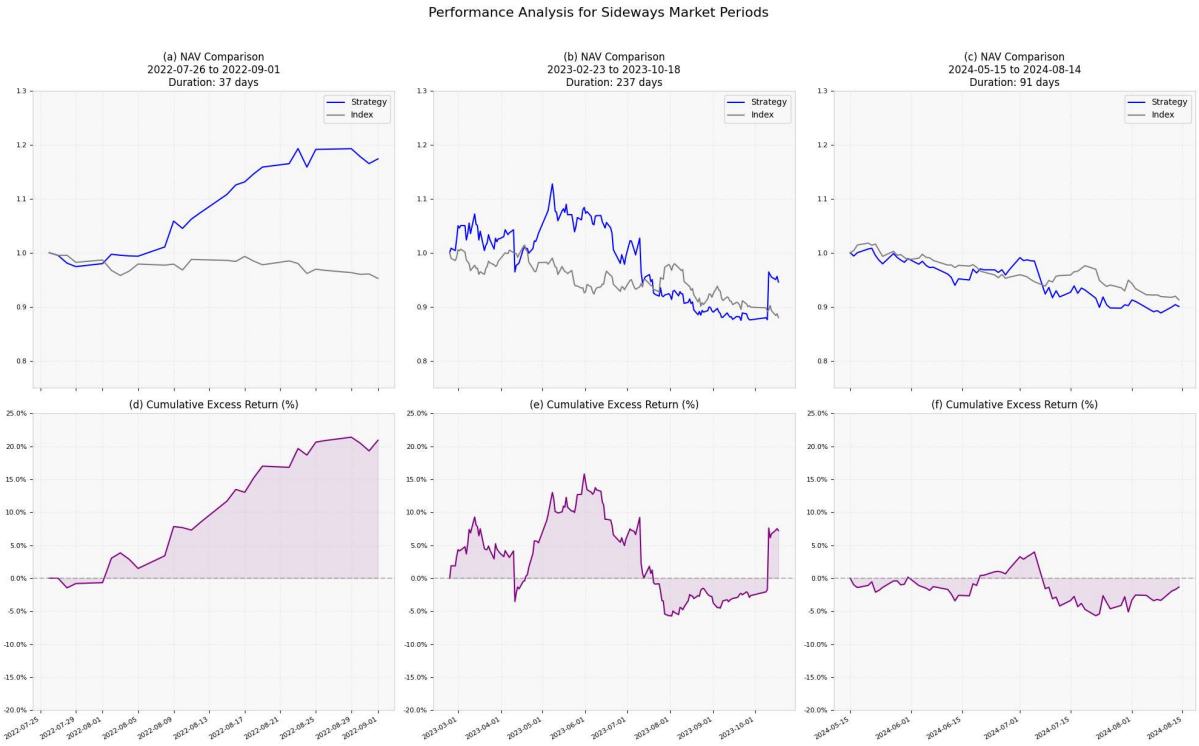

Figure 16: Performance analysis during the three longest sideways market periods

6. Conclusion

This thesis investigates whether a momentum-based stock selection strategy, powered by a random forest classification model, can generate persistent excess returns in the Chinese stock market. Empirical results demonstrate that the proposed strategy not only outperforms a standard market benchmark but also benefits significantly from incorporating risk-based portfolio weighting methods. In particular, assigning portfolio weights by minimizing maximum drawdown leads to the highest risk-adjusted returns, despite introducing greater portfolio turnover. These findings suggest that combining machine learning techniques with price momentum signals and risk control mechanisms offers a promising approach to navigating relatively inefficient emerging markets such as China.

Interestingly, the strategy achieves its strongest performance under sideways market conditions, which traditionally pose challenges for trend-following models. This anomaly can be attributed to the cross-sectional nature of the momentum signal and an overall long-term trend, even in the absence of a clear market direction in the short term.

Overall, the findings of this thesis provide empirical support for the notion that the Chinese stock market does not fully satisfy the weak-form efficiency hypothesis. The effectiveness of a technical momentum-based strategy, which is driven purely by past price information and implemented using a machine learning model, suggests that historical return patterns contain exploitable predictive signals. This contradicts the assumption of random walk behavior implied by weak-form efficiency and highlights the existence of persistent return anomalies in the Chinese equity market. Such inefficiencies may be attributed to structural factors such as the dominance of retail investors, regulatory frictions, limited short-selling mechanisms, etc. As a result, data-driven strategies that exploit historical price behavior, like the one proposed in this thesis, can still yield significant abnormal returns in this context.

However, there are also some limitations of this thesis. First, the analysis relies solely on historical price-based technical indicators, without incorporating fundamental data or macroeconomic variables that may provide additional predictive power. Second, while the random forest model performs well, the thesis does not compare it against other state-of-the-art machine learning models such as LightGBM, XGBoost, or attention-based deep learning architectures. Third, the evaluation does not consider realistic liquidity constraints, or market impact, which could affect the practical implementation of the strategy.

Future research could address these limitations by expanding the feature set to include multi-source information including fundamental features, macroeconomic features, liquidity

and other uncommon features such as investors sentiment. Moreover, further studies can also use alternative model techniques and compare these models, which may also shed lights on the enhancement of the performance.

References

Ballings, M., Van den Poel, D., Hespeels, N., Gryp, R., 2015. Evaluating Multiple Classifiers for Stock Price Direction Prediction. Expert Syst. Appl. 42, 7046–7056.

Barberis, N., Shleifer, A., Vishny, R., 1998. A Model of Investor Sentiment. J. Financ. Econ. 49, 307–343.

Belciug, S., Sandita, A., 2017. Business Intelligence: Statistics in Predicting Stock Market. Ann. Univ. Craiova-Math. Comput. Sci. Ser. 44, 292–298.

Breiman, L., 2001. Random Forest. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Carhart, M.M., 1997. On Persistence in Mutual Fund Performance. J. Finance 52, 57–82. https://doi.org/10.1111/j.1540-6261.1997.tb03808.x

Chan, L.K.C., Jegadeesh, N., Lakonishok, J., 1996. Momentum Strategies. J. Finance 51, 1681–1713. https://doi.org/10.1111/j.1540-6261.1996.tb05222.x

Chekhlov, A., Uryasev, S., Zabarankin, M., 2005. Drawdown Measure in Portfolio Optimizaiton. Int. J. Theor. Appl. Finance 08, 13–58. https://doi.org/10.1142/S0219024905002767

Chong, E., Han, C., Park, F.C., 2017. Deep Learning Networks for Stock Market Analysis and Prediction: Methodology, Data Representations, and Case Studies. Expert Syst. Appl. 83, 187–205.

Daniel, K., Hirshleifer, D., Subrahmanyam, A., 1998. Investor Psychology and Security Market Under- and Overreactions. J. Finance 53, 1839–1885. https://doi.org/10.1111/0022-1082.00077

DeMiguel, V., Garlappi, L., Uppal, R., 2009. Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy? Rev. Financ. Stud. 22, 1915–1953.

Enke, D., Thawornwong, S., 2005. The Use of Data Mining and Neural Networks for Forecasting Stock Market Returns. Expert Syst. Appl. 29, 927–940.

Fama, E.F., French, K.R., 2015. A Five-factor Asset Pricing Model. J. Financ. Econ. 116, 1–22.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical Asset Pricing via Machine Learning. Rev. Financ. Stud. 33, 2223–2273.

Hong, H., Stein, J.C., 1999. A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets. J. Finance 54, 2143–2184. https://doi.org/10.1111/0022-1082.00184

Jagannathan, R., Ma, T., 2003. Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps. J. Finance 58, 1651–1683. https://doi.org/10.1111/1540-6261.00580

James, G., Witten, D., Hastie, T., Tibshirani, R., Taylor, J., 2023. An Introduction to Statistical Learning: with Applications in Python, Springer Texts in Statistics. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-38747-0

Jegadeesh, N., Titman, S., 1993. Returns to Buying Winners and Selling Losers: Implications for Efficiency. J. Finance 48, 65–91.

Kavin, S., Mohan, S.K., Karthick, V.I., Sudar, K.M., 2018. Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection. Spec. Sect. Surviv. Strateg. Emerg. Wirel. Netw. IEEE.

Krauss, C., Do, X.A., Huck, N., 2017. Deep Neural Networks, Gradient-boosted Trees, Random Forests: Statistical Arbitrage on the S&P 500. Eur. J. Oper. Res. 259, 689–702.

Pagan, A.R., Sossounov, K.A., 2003. A simple framework for analysing bull and bear markets. J. Appl. Econom. 18, 23–46. https://doi.org/10.1002/jae.664

Yu, B., 2024. Is the Chinese Stock Market Efficient? Evidence from a Combined Liquidity Trading Strategy. China Finance Rev. Int.

Yuan, X., Yuan, J., Jiang, T., Ain, Q.U., 2020. Integrated Long-term Stock Selection Models Based on Feature Selection and Machine Learning Algorithms for China Stock Market. Ieee Access 8, 22672–22685.

Zhu, M., Philpotts, D., Sparks, R., Stevenson, M., 2011. A Hybrid Approach to Combining CART and Logistic Regression for Stock Ranking. J. Portf. Manag. 38, 100–109.

Zhu, M., Philpotts, D., Stevenson, M.J., 2012. The Benefits of Tree-based Models for Stock Selection. J. Asset Manag. 13, 437–448. https://doi.org/10.1057/jam.2012.17

AI Usage Note

In accordance with the transparency policy on the use of AI tools, hereby the following instances of AI assistance during the preparation of this thesis is disclosed as follows.

- Grammar checking: ChatGPT is used to identify grammatical errors of academic writing. All content was originally written by the author and subsequently revised with AI assistance.
- Coding support: During the implementation of the Random Forest-based stock selection strategy, AI tools were used to help identify and resolve syntax errors in Python code. The overall design, logic, and structure of the code were developed independently by the author.

All AI usage was limited to technical support, and no part of the thesis was generated or written entirely by AI.