Dean&Francis

ISSN 2959-6157

Multi-factor Selection and Stock Picking Strategy Based on Random Forest and LightGBM Algorithms

Pengjie Yuan

Sino-European School of Technology, Shanghai University, Shanghai, 200444, China
294275073@shu.edu.cn

Abstract:

This study employs Random Forest (RF) and Light Gradient Boosting Machine (LightGBM) algorithms to construct a multi-factor stock selection strategy. Focusing on China Securities Index 300 (CSI 300) constituent stocks, it integrates fundamental and technical factors to conduct factor validity testing and weight optimization through machine learning methods. The study first constructs and preprocesses the factor system. Subsequently, a Random Forest model is employed for binary classification prediction of stock returns, achieving an accuracy of 81.7% and an Area Under the Receiver Operating Characteristic Curve (AUC) value of 0.88 on the test set, demonstrating strong classification capability. Further, a Light Gradient Boosting Machine regression model is used to predict stock market capitalization, yielding a goodness-of-fit of 0.91, which is then utilized to screen for undervalued stock portfolios. During strategy backtesting, four high-importance factors were weighted and synthesized. Over the backtesting period from August 2022 to August 2025, the strategy achieved a total return of 28.68% and an annualized return of 9.07%, generating an excess return of 29.65% relative to the benchmark index. The maximum drawdown was 18.39%, with a Sharpe Ratio of 0.310. The results demonstrate that quantitative strategies integrating multi-factor and machine learning approaches exhibit certain effectiveness and practicality in the A-share market.

Keywords: Random Forest; LightGBM; Multi-factor Stock Selection; CSI 300 Index; Backtesting Analysis

1. Introduction

With the rapid development of financial research theories and artificial intelligence technologies, quantitative investing has increasingly gained prominence in financial markets. Compared to traditional investing, quantitative investing offers advantages in quantification and precision [1]. In the field of quantitative in-

ISSN 2959-6157

vesting, traditional single-factor stock selection strategies are no longer sufficient to adapt to complex and volatile market conditions. Multi-factor stock selection strategies can capture most market information, effectively enhancing the scientific rigor and stability of stock selection. As such, they represent one of the most widely applied methods in stock selection strategies [2].

The three-factor model proposed by Fama and French pioneered multi-factor research, demonstrating the explanatory power of factors such as market capitalization and book-to-market ratio on stock returns [3]. Since then, scholars have continuously expanded the factor system, incorporating financial indicators, technical indicators, and other elements into the analytical framework [1]. In the A-share market, China Securities Index 300 (CSI 300) constituents, as representatives of large-cap blue chips, exhibit price volatility that combines market representativeness with investment security, making them a crucial target pool for quantitative strategy research [4].

Existing research indicates that single-factor strategies are susceptible to market cycles, whereas multi-factor portfolios can reduce strategy volatility through the complementary nature of factors. Ashley Lester's empirical findings indicated that an integrated portfolio of four orthogonal factors generated twice the factor exposure of the corresponding set of single-factor portfolios, twice the outperformance, and a 40% higher information ratio [5]. Fundamental factors such as P/E ratio and ROE reflect a company's intrinsic value, while technical factors like momentum and trading volume capture market sentiment and trends. Combining both provides a relatively comprehensive assessment of a stock's investment value. Peng Huang utilized fundamental factor data and technical factor data as backtesting data for the strategy. By integrating machine learning algorithms, he developed an effective multi-factor stock selection approach. Backtesting results demonstrated that the strategy achieved a cumulative return of 140.61% and an annualized return of 35.57% [6]. Meanwhile, the introduction of machine learning algorithms has opened new avenues for factor weight optimization. Models such as Random Forest and Light Gradient Boosting Machine (LightGBM) can effectively handle nonlinear relationships and higher-order interactions among factors, thereby enhancing stock selection accuracy [7].

This study examines the CSI 300 Index constituents by introducing multidimensional indicators encompassing fundamental and technical factors. Combining machine learning methods such as Random Forest, it tests factor validity and quantifies weights to construct a stock selection strategy capable of generating stable excess returns. Through strategy backtesting and performance evaluation, the research not only validates the applicability of multi-factor models in China's A-share market but also offers further insights and references for quantitative investment practices.

2. Data and Methods

2.1 Data Sources and Processing

JoinQuant is one of the leading emerging quantitative investment platforms in China, offering researchers free backtesting access to fundamental data covering A-shares, funds, futures, indices, and more [8]. This study obtained daily stock market data for the past two months up to August 15, 2025, through the JoinQuant platform. This paper utilizes the CSI 300 index constituents as the stock pool and extracts several fundamental factors, including: market capitalization, net operating capital, net debt, debt-to-equity ratio, non-current assets ratio, shareholders' equity ratio, revenue growth rate, P/E ratio, P/B ratio, P/S ratio, and ROE. Several technical factors include: momentum lines, trading volume, cumulative energy lines, average deviation, exponential moving averages, simple moving averages, deviation rates, and others. These elements form a multidimensional factor system integrating fundamental and technical analysis. Regarding technical factors with the cutoff date set as two days prior (i.e., August 13, 2025), the calculation period is uniformly 10 days. Data is sourced from daily charts. Any missing values in the obtained data are filled with zeros to prevent subsequent analysis errors. The complete data processing flowchart is shown in Fig. 1.

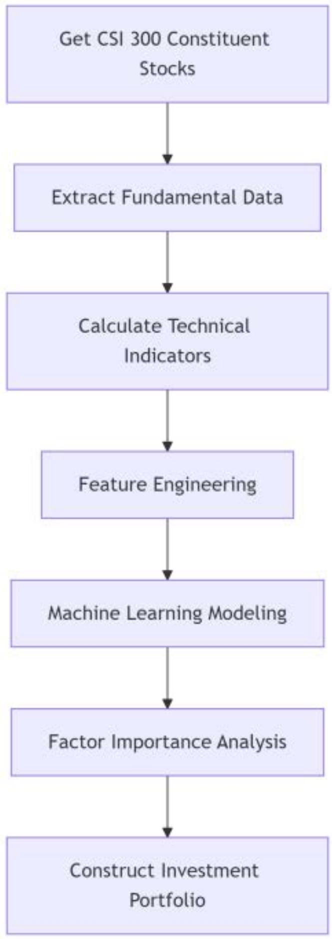

Fig. 1 Complete data processing flowchart (Photo/Picture credit: Original).

2.2 The Construction of Machine Learning Models

2.2.1 Setting objectives and training the Random Forest model

First, locate the historical closing price of the stock (e.g., 50 days prior). Then, calculate the stock‘s return over the past 50 days by dividing the previous day‘s closing price by the closing price 50 days prior and subtracting 1. Next, identify stocks with returns exceeding the average level and mark them as 1, while marking the rest as 0 to serve as the model‘s classification labels (see Table 1). Close1, Close2, and return refer to the closing price one day prior, the closing price fifty days prior, and the return within the specified time period, respectively. The Signal is the classification label used for training the model.

Table 1. Historical returns and classification tags

Ticker symbol	…	Bias Rate	Close1	Close2	return	signal
000001.XSHE	…	0.08	12.40	11.84	0.05	0
000002.XSHE	…	-1.30	6.45	6.29	0.03	0
000063.XSHE	…	-1.56	33.78	30.96	0.09	0
000100.XSHE	…	-1.20	4.42	4.30	0.03	0
000157.XSHE	…	1.46	7.53	6.79	0.11	1

During the model training phase, the model first extracts the feature matrix X and label vector y from the preprocessed dataset. The feature matrix excludes columns directly related to price—specifically closing prices (close1, close2), returns (return), and the signal column to be predicted—retaining only factor values as input features. The signal column is treated as the classification label y. Subsequently, the entire dataset was randomly split into training and test sets at an 8:2 ratio. This study employed a Random Forest classification model, setting the number of decision trees to 5000 and fixing the random seed to ensure experimental reproducibility. After model fitting on the training set, its classification accuracy was evaluated on both the training and test sets to assess the model‘s learning capability and generalization performance.

2.2.2 The construction of the LightGBM regression model

ISSN 2959-6157

LightGBM is a decision tree algorithm based on lightweight gradient boosting, which demonstrates significant advantages when handling large-scale data [9]. This paper employs a LightGBM regression model to forecast the market capitalization of CSI 300 constituent stocks. The feature matrix X is selected from fundamental factor values and technical factor values within the dataset, while the target variable y refers to stock market capitalization. The model uses LGBMRegressor as the regressor, sets the number of trees to 100, and optimizes the objective for regression tasks. To objectively evaluate model performance, a random stratified sampling method was employed to partition the dataset into training and test sets, with the test set comprising 20% of the data. A random seed was set to ensure experimental reproducibility. During model training, both L2 mean squared error and L1 mean absolute error were introduced as evaluation metrics, enabling the model to continuously monitor performance across multiple dimensions throughout the training process. During training, the model establishes a prediction function by fitting the mapping relationship between features X and target labels y. Ultimately, evaluation metrics are computed on the training set to assess the model's fitting performance.

3. Results

3.1 Performance Evaluation of the Random Forest Model

Experimental results show that the Random Forest model achieved 100% accuracy on the training set and 81.7% accuracy on the test set, indicating that the model possesses good generalization capabilities. This study also constructed a confusion matrix by comparing the predicted labels with the true labels of the test set samples, providing an intuitive visualization of the overall distribution of classification results (see Fig. 2). The specific numerical distribution of this matrix indicates that the model correctly predicted 30 true negative examples in negative sample recognition, while misclassifying 6 as positive examples. In positive sample recognition, the model correctly predicted 19 true positive examples, with 5 misclassified as negative examples. This distribution pattern demonstrates the model's strong ability to distinguish between different categories.

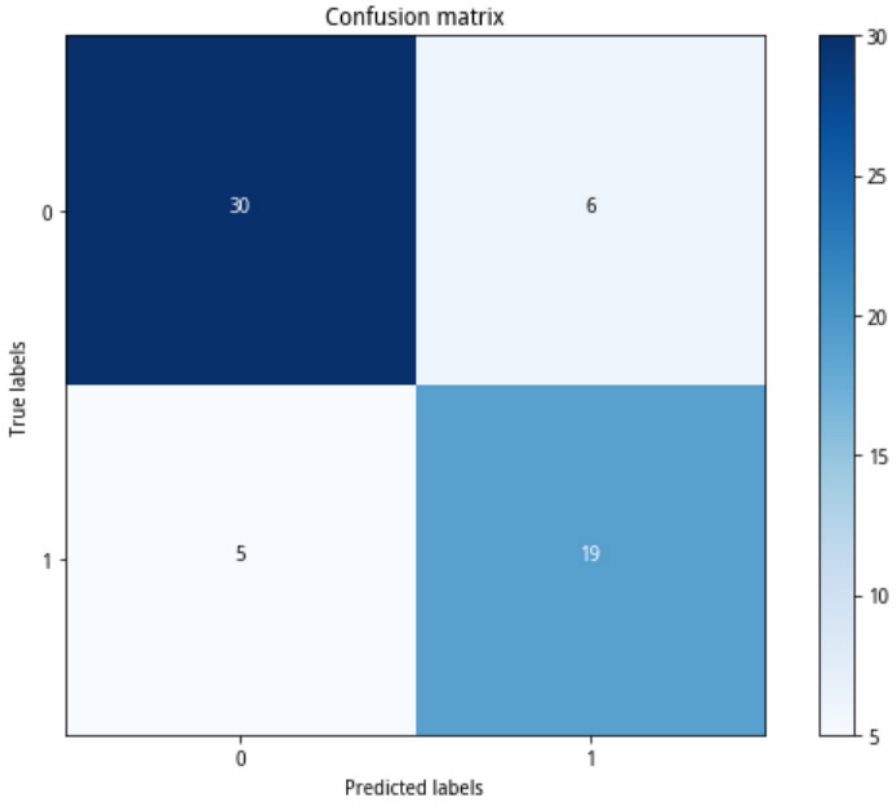

Fig. 2 Confusion matrix visualization (Photo/Picture credit: Original).

Further quantitative analysis was conducted using classification performance metrics, with the results presented in the table below (see Table 2).

Table 2. Classification report

	precision	recall	F1-score	support
0	0.86	0.83	0.85	36
1	0.76	0.79	0.78	24
avg/total	0.82	0.82	0.82	60

The results show that the model achieved a precision of 0.86 and a recall of 0.83 for negative class recognition, with an F1 score of 0.85. For positive class recognition, it attained a precision of 0.76 and a recall of 0.79, yielding an F1 score of 0.78. Overall, the model maintained a consistent classification performance at an F1 score of 0.82, indicating a relatively balanced recognition capability across both positive and negative categories.

To further validate the model's discriminatory ability, this study plotted a ROC curve and calculated the area under the curve (see Fig. 3). The ROC curve analysis results indicate that the model achieved an AUC value of 0.88, which is significantly higher than the baseline level of 0.5 for random classifiers, this demonstrates the model's excellent classification capability [10]. The shape of the ROC curve further indicates that the model maintains a high true positive rate and relatively low false positive rate across different decision thresholds. This characteristic holds significant importance for risk control in practical applications. Based on a comprehensive evaluation of all metrics, the constructed predictive model demonstrates robust and reliable performance in the stock return classification task, providing solid data support for subsequent practical applications.

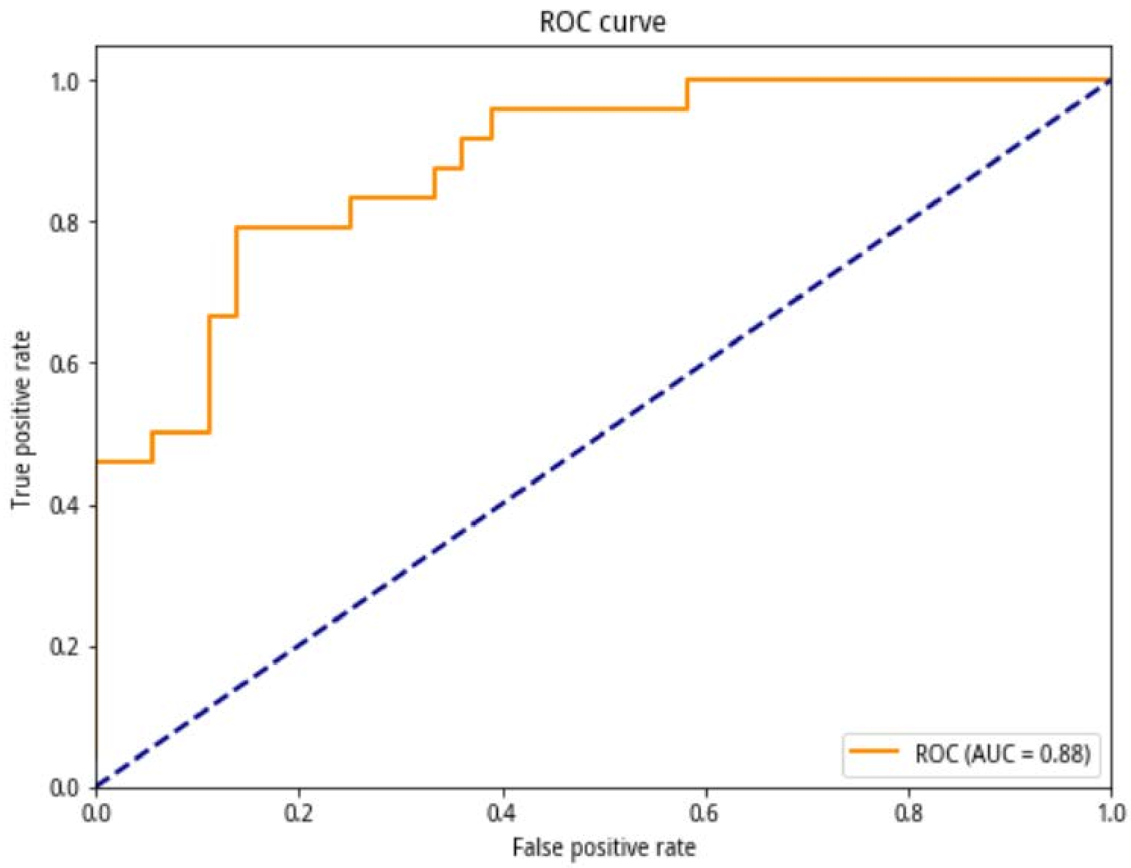

Fig. 3 ROC curve (Photo/Picture credit: Original).

3.2 Performance Evaluation of LightGBM

To comprehensively evaluate the model's predictive accuracy, this paper employs four core metrics: MSE measures the squared deviation between predicted and actual values, MAE reflects the average level of absolute deviation, RMSE—as the square root of MSE—provides a more intuitive representation of error magnitude, and R2 char-

ISSN 2959-6157

acterizes the model's ability to explain return volatility (see Table 3).

Table 3. LightGBM Model Evaluation Metrics

Evaluation metrics	Values
MSE	0.0063
MAE	0.0618
RMSE	0.0796
R^2	0.7565

Based on the above evaluation metrics, the model demonstrates good fitting performance.

3.3 Factor Importance Analysis

In this study, to evaluate the contribution of each feature variable within the model, feature importance values were obtained using the trained classifier. These values were then combined with their corresponding feature names (i.e., factor names) to construct a DataFrame structure. The dataframe is sorted in descending order by feature importance scores to more clearly identify the key factors influencing model decisions. To further visualize the analysis results, the seaborn library is used to plot a horizontal bar chart, where the vertical axis represents factor names and the horizontal axis represents importance levels. The graphic dimensions were set to 6×4 inches at a resolution of 128 dpi to more intuitively illustrate the importance ranking of different factors (see Fig. 4).

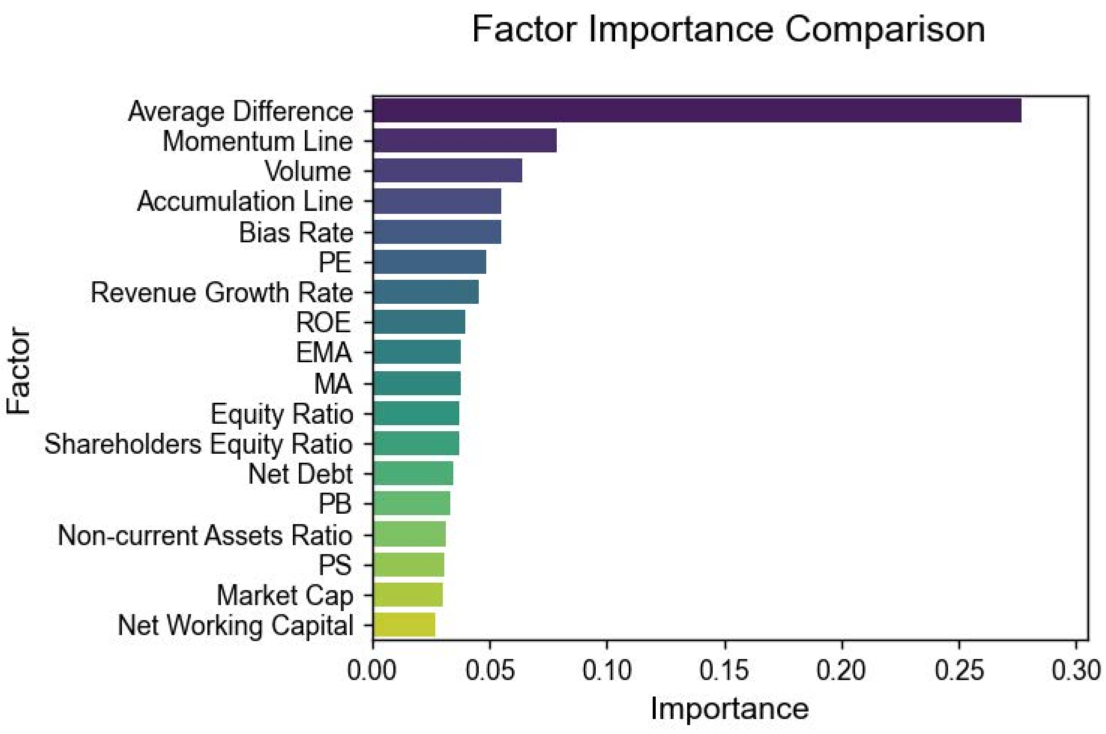

Fig. 4 Factor importance comparison (Photo/Picture credit: Original).

3.4 Stock Selection Strategy

Based on the market capitalization predictions for CSI 300 constituent stocks generated by the LightGBM regression model in Section 1.2.2, it subtracts the predicted values from the actual market capitalization values. These values are then sorted in ascending order to identify the top 10 stocks—those with the highest degree of undervaluation (a larger negative value indicates greater undervaluation). Such stocks can be considered promising prospects (see Table 4).

Table 4. The 10 selected stocks

Ticker symbol	Difference between actual and predicted values
600000. XSHG	-5737.39
601818.XSHG	-4114.70
601088.XSHG	-3844.98
601658.XSHG	-3003.09
601328.XSHG	-2859.04
600036.XSHG	-2284.22
601390.XSHG	-2003.17
600030.XSHG	-1780.23
600016.XSHG	-1707.24
601211.XSHG	-1657.62

4. Strategy Backtesting

During the strategy initialization phase, several parameters were set: rebalancing every 15 days, a historical sample length of 63 days, and a maximum of 20 holdings. Based on the factor importance ranking in Figure 2, the top two fundamental factors (P/E ratio and revenue growth rate) and the top two technical factors (average deviation and momentum line) were selected. These four factors formed the core components of the strategy backtest. Each factor was assigned directional weights based on its characteristics: 1 indicates a preference for lower factor values, while -1 indicates a preference for higher factor values. Consequently, PE, revenue growth rate, average deviation, and momentum were set to 1, -1, 1, and -1, respectively. Before the market opens each trading day, the strategy determines whether it is a rebalancing day. If so, screen the viable stock pool by removing stocks from the CSI 300 Index that are currently suspended or have a suspension record within the past 63 days. Then, retrieve the required financial data, including P/E ratio and revenue growth rate. Additionally, transaction commission rates are adjusted according to different historical periods to simulate real market conditions.

During the trading session, the strategy first calculates the current allocation of capital across positions. It then retrieves and processes factor data, filling in missing values with the mean. Each factor's ranking score is calculated using a weighted method, and a composite score is derived for each stock. Following Bubble Sort, the top 20 stocks are selected as the target holdings. During portfolio rebalancing, first sell holdings not on the target list, then evenly allocate funds to purchase target stocks. After the daily market close, the strategy performs no additional operations. The overall backtesting process simulates investment performance across different time periods by cyclically executing the aforementioned daily operations.

This study conducted strategy backtesting via the JoinQuant platform, covering the period from August 5, 2022, to August 4, 2025. With an initial capital of RMB 2,000,000 and a daily execution cycle, the benchmark was the CSI 300 Index. The strategy implementation yielded the return trend chart shown in Fig. 5. Calculations revealed a strategy return of 28.68%, an annualized return of 9.07%. The benchmark return was -0.75%, resulting in an excess return of 29.65%. The maximum drawdown reached 18.39%, and the Sharpe ratio was 0.310.

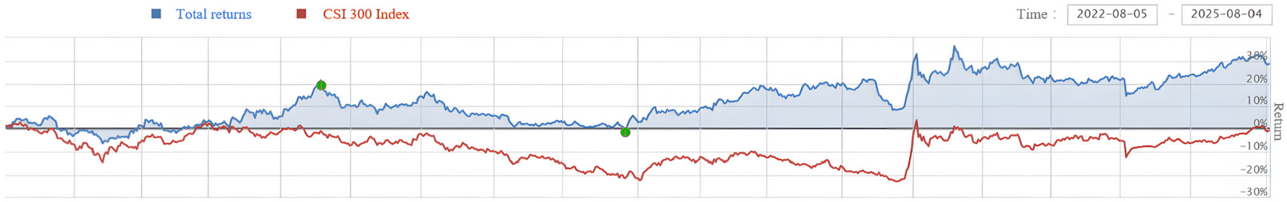

Fig. 5 Total returns trend (Photo/Picture credit: Original).

5. Conclusion

This study constructs a quantitative investment strategy integrating multi-factor models and machine learning algorithms. Empirical backtesting is conducted using

ISSN 2959-6157

CSI 300 constituent stocks as the equity pool to validate whether stable excess returns can be achieved through the combined application of fundamental and technical factors. The strategy achieved a total return of 28.68% during the backtesting period, with an annualized return of 9.07%. Compared to the benchmark return (-0.75%), it delivered an excess return of 29.65%. The maximum drawdown was 18.39%, and the Sharpe ratio was 0.310. This indicates that the strategy generated positive returns while assuming a certain level of risk, demonstrating practical viability in real-world trading.

It is worth noting that while the strategy significantly outperformed the benchmark during the backtesting period, it still experienced a maximum drawdown of 18.39%. This indicates that the strategy remains subject to certain risks during overall market downturns or style shifts. This phenomenon may be related to the stability of the factors across different market environments, suggesting that future research should incorporate risk control modules—such as dynamic position management or stop-loss mechanisms—to enhance the strategy's robustness.

Additionally, this study incorporates machine learning methods for factor importance assessment and composite signal generation during factor synthesis and stock screening, offering greater flexibility and adaptability compared to traditional linear weighting approaches. LightGBM demonstrates strong performance in regression tasks, exhibiting robust stock market capitalization prediction capabilities. This also provides avenues for further optimization in subsequent research, such as integrating more alternative data or natural language processing techniques to enhance factor dimensions and information quality.

In summary, the empirical findings of this study demonstrate that the quantitative stock selection strategy based on multi-factor integration and machine learning assistance exhibits certain effectiveness and practicality in the A-share market. Its return performance significantly outperforms market benchmarks, validating the applicability of factor investing principles in the Chinese market. However, the strategy still carries certain volatility and drawdown risks. Therefore, subsequent research should place greater emphasis on risk control and factor timing to further enhance the strategy's overall performance and practical value.

References

[1] Wang J, Zhuang Z, Feng L. Intelligent optimization-based multi-factor deep learning stock selection model and quantitative trading strategy. Mathematics, 2022, 10(4): 566.

[2] Yang Y Z. Research on multi-factor quantitative stock selection strategy based on machine learning. Xi'an University of Technology, 2024.

[3] Fama E F, French K R. The cross-section of expected stock returns. Journal of Finance, 1992, 47(2): 427-465.

[4] Zhang S X. The CSI 300 Index marks two decades since its launch, establishing itself as the core benchmark for China's A-share market. Securities Times, 2025-07-01(A002).

[5] On the theory and practice of multifactor portfolios. The Journal of Portfolio Management, 2019, 45(3): 87-101.

[6] Huang P. Research on multi-factor stock selection scheme based on knockoff screening factors and CatBoost algorithm prediction. Shanghai Normal University, 2024.

[7] Wang X Y. Construction of multi-factor stock selection model based on random forest: A special topic report on financial engineering. Beijing: Sina.com Technology (China) Co., Ltd., 2025-07-01 [cited 2025-08-20]. Available from: https://finance.sina.com.cn/stock/stockzmt/2025-07-01/doc-infcxncq9134858.shtml.

[8] Yang S L. Application of stock multi-factor strategy based on JoinQuant quantitative investment platform. Zhejiang University, 2018.

[9] Li C H. Multi-factor quantitative stock selection analysis based on LightGBM algorithm and investor sentiment. Money, 2024, (9): 89-90+92.

[10] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters, 2006, 27(8): 861-874.