Proceedings of the 4th International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/2026.32361

A Comparative Study and Backtesting of Machine Learning–Based Quantitative Stock Selection Models in China's A-Share Market

Bowen Yang

SWUFE-UD Institute of Data Science, Southwestern University of Finance and Economics, Chengdu, China
42355048@smail.swufe.edu.cn

Abstract. Based on machine learning methods, this paper systematically studies the cross-sectional excess returns of A-share stocks. Taking 5,648 stocks in China's A-share market from 2020 to 2024 as a sample, a multi-dimensional factor system covering fundamentals and macro variables was constructed, and the out-of-sample performance of multiple models was compared under the rolling forecast framework of Expanding Window. The empirical results show that the nonlinear machine learning model continues to obtain positive information coefficients in most quarters, and its prediction signals can be converted into stable long-short portfolio returns, while the prediction ability and investment performance of traditional linear models decay significantly with time. In terms of risk control, LightGBM shows a relatively better trade-off between return level and drawdown magnitude. The research in this paper confirms that the machine learning method that combines a nonlinear structure and cross-sectional heterogeneity can effectively improve the return prediction ability and investment performance of the A-share market and provides practical empirical evidence for quantitative stock selection.

Keywords: Machine learning, stock return, rolling forecast, Long-Short portfolio.

1. Introduction

Under the dual requirements of profound adjustment of the global economic landscape and high-quality development of the domestic economy, China's A-share market, as the core hub for resource allocation, has attracted much attention due to its operational trends and efficiency. In recent years, with the deepening of the registration system reform, the change of investor structure, and the frequent emergence of macro policies and external shocks, the generation mechanism of stock returns has shown more significant non-linear and time-varying characteristics, and the traditional forecasting methods relying on stable linear relationships are facing more and more challenges.

Classical asset pricing models, such as the Capital Asset Pricing Model (CAPM) and the Fama–French Multifactor Model, provide an important theoretical basis for understanding the sources of stock returns. However, such models are usually based on linear structures and relatively static risk exposure assumptions, making it difficult to fully describe the complex return patterns caused by

DOI: 10.54254/2754-1169/2026.32361

investor heterogeneity, information friction, and changes in the macro environment in the real market [1]. In this context, how to improve the accuracy and stability of stock return forecasting in a high-dimensional, noisy and nonlinear environment has become an important question that needs to be answered urgently in the current asset pricing and investment practice.

With the improvement of computing power and the improvement of data availability, machine learning methods have gradually been introduced into the field of financial forecasting. Machine learning models have natural advantages over traditional linear models in processing high-dimensional data, characterizing nonlinear interactions, and capturing complex patterns. Recent studies have shown that nonlinear methods such as ensemble tree models demonstrate strong out-of-sample prediction capabilities and potential economic value in stock cross-sectional prediction tasks [2]. However, in the context of China's A-share market, there are still relatively limited studies on systematically comparing different machine learning models and testing their predictive ability and investment performance under a unified rolling forecasting framework.

Based on the above research background, this paper aims to construct a comprehensive factor system that integrates company fundamentals and macroeconomic information and systematically compares a variety of representative models under a strict out-of-sample rolling forecast framework of Expanding Window. The research contribution of this paper is mainly reflected in three aspects. First, at the data level, this article covers 5,648 A-share stocks from 2020 to 2024 and constructs a quarterly dataset containing 18 company fundamental factors and 7 macroeconomic variables. Second, at the methodological level, this paper makes a horizontal comparison of linear models, nonlinear ensemble learning models and structured mixed models under a unified rolling prediction framework to avoid prospective bias and highlight the comparability between models. Third, at the application level, this paper not only evaluates the cross-sectional prediction ability of the model from a statistical perspective, but also systematically analyzes the risk-return characteristics of the model in real investment scenarios through cumulative net compounding value and long-short strategy back testing, so as to provide a more realistic reference for model selection and asset allocation decisions..

2. Literature review

Stock earnings forecasting has long relied primarily on traditional econometric models based on linear assumptions, including multi-factor models and time series methods. Such models have strong economic intuition and interpretability in explaining sources of revenue, but their predictive power is often limited in the face of high-noise, non-linear and structurally changing financial markets. Empirical studies on the A-share market also show that the performance of linear models in out-of-sample predictions is unstable, and the economic excess returns that can be generated are relatively limited [3].

With the successful application of machine learning methods in other fields, more and more research has begun to introduce them into financial market forecasting. Compared with traditional models, machine learning methods can automatically identify complex nonlinear patterns from a large number of features with fewer structural assumptions. Previous studies have found that ensemble learning models such as random forests and gradient boosting trees are significantly better than linear benchmark models in stock cross-sectional prediction, especially in high-dimensional feature environments [2]. These results suggest that nonlinear structures may play an important role in the formation of stock returns.

However, existing research also points out that a single complex machine learning model is not a one-size-fits-all solution. On the one hand, the model performance is highly sensitive to the sample

DOI: 10.54254/2754-1169/2026.32361

interval and market environment, and the off-sample stability is often insufficient. On the other hand, the highly complex model structure reduces the interpretability of the results, limiting its application in actual investment decisions. In addition, there may be significant structural heterogeneity in the stock cross-section, with stocks of different industries, sizes, or styles being dominated by different earnings-driven mechanisms, and a unified global model may mask these differences, limiting the effectiveness of forecasting.

In order to address the above problems, recent studies have begun to focus on structured and group modeling methods. For example, stocks are clustered through unsupervised learning and then predictive models are trained within each subsample to characterize potential heterogeneous structures in cross-sections. The empirical results show that this type of grouping modeling method has certain advantages in prediction accuracy and robustness compared with a single model [4]. This research direction emphasizes that instead of continuously increasing model complexity, it is better to explicitly introduce market structure information into the forecasting framework to achieve a balance between predictive power, robustness, and interpretability.

Based on the existing literature, it can be seen that stock return forecasting research is gradually shifting from a single model and static setting to a comprehensive framework emphasizing out-of-sample testing, model comparison and economic value evaluation. However, in the context of China's A-share market, there is still a lack of research on comparing linear models, nonlinear ensemble learning models and structured hybrid models under a unified rolling forecast framework and systematically evaluating their cross-sectional predictive power and investment performance. It is in this context that this paper is carried out, aiming to provide new empirical evidence for understanding the cross-sectional driving mechanism of A-share market returns.

3. Methodology

This study focuses on the cross-sectional prediction of stock excess returns, aiming to test whether fundamental variables at the company level can effectively predict the relative performance of stock excess returns in the next quarter. Different from traditional time series forecasting, this paper focuses on the model's ability to rank between different stocks in the same period, that is, whether it can effectively distinguish between stocks with better and worse performance in the future, so as to provide actionable investment signals for building long-short portfolios.

In order to ensure the out-of-sample validity of the prediction results, this paper adopts the rolling prediction framework of Expanding Window. In each forecast quarter, the model is trained using only all previous historical data and cross-sectional forecasts are made for the current quarter's stocks, so as to strictly avoid forward-looking bias. Specifically, the minimum training period is set for the first 8 quarters, with out-of-sample forecast starting from the 9th quarter. Over time, the training set expands over time, and the test set is always the data for the current forecast quarter. In each round of training, the last quarter in the training sample is divided into a validation set for model parameter selection and early stop control, and the rest of the historical data is used for model estimation. After completing training, the model generates forecasts for all stocks in the next quarter and calculates the ranking correlation between the forecasted values and true earnings.

Under a unified rolling prediction framework, this paper constructs and compares a variety of prediction models of different complexity, including linear contraction models (LASSO, Ridge), nonlinear ensemble models (Random Forest), gradient boosting models (LightGBM), and hybrid models combined with unsupervised learning (KMeans + LightGBM). By comparing the performance of different models in cross-sectional ranking ability and portfolio performance, the impact of model complexity on predictive power is systematically evaluated.

DOI: 10.54254/2754-1169/2026.32361

3.1. Data source and sample

The data used in this article is derived from the Cathay Securities database (CSMAR), covering all A-share listed companies on the Shanghai, Shenzhen and Beijing exchanges. The sample screening is based on the stock market classification, excluding financial stocks and special treatment (ST) stocks, and using the Shenyin Wanguo Industry Classification (2021 Revised Edition) for industry division. The final sample contains 5648 stocks from Q4 2019 to Q4 2024.

In terms of data frequency, this paper uses quarterly data as the basic analysis unit. Financial factors and macroeconomic variables at the company level are observable at the end of the quarter and are uniformly lagged forward by one quarter in the empirical evidence to predict the excess return of stocks in the next quarter. The target variable is the quarterly excess return of the stock, which is described as the difference between quarterly return of the stock and the risk-free interest rate for the same period. This lagged operation ensures that the model is based only on known future returns over the current period, meeting the limitations of information availability in a real investment environment.

3.2. Factor construction and variable processing

In terms of factor selection, this paper constructs predictors from two dimensions, namely company fundamentals and macroeconomic factors. Eighteen financial indexes are selected as fundamental factors including valuation level such as price-earnings ratio, profitability, debt repayment level and operational efficiency, which reflect the intrinsic value of the company. On a macro basis, seven indicators are designated to serve as a proxy for changes in the liquidity environment and economic cycle.

For the purpose of mitigating the interference of outliers on model training, the target variables were tailed by 1%–99% quantiles (Winsorization) before training. In addition, for models such as LASSO and ridge regression, they are standardized before estimation to ensure that all transformations are calculated based on training samples, maintaining the rigor of out-of-sample predictions [5].

3.3. Model selection

This paper adopts a multi-dimensional modeling framework from linear benchmarks to nonlinear machine learning models and finally a structured hybrid method, comprehensively evaluating the applicability of different forecasting paradigms in stock cross-sectional return forecasting.

Initially, LASSO regression serves as the linear baseline. By imposing L1 regularization, LASSO automatically selects variables in a high-dimensional factor environment, yielding a sparse and interpretable predictive model. In implementation, LassoCV with 5-fold cross-validation is employed to autonomously determine the optimal regularization strength, with the maximum number of iterations set to 5000 to ensure convergence. At the same time, ridge regression is introduced as a stable control to alleviate the multicollinearity problem between features by selecting the best among 20 logarithmic equidistant α candidates and setting the performance benchmark for subsequent nonlinear models.

Building on this foundation, two mainstream ensemble tree models, which refer to Random Forest and LightGBM, are further adopted. Random forest improves the robustness of the model through Bagging mechanism and random feature selection and effectively captures the nonlinear interaction between variables. The model sets the number of trees to 300, the maximum depth to 6,

DOI: 10.54254/2754-1169/2026.32361

and the minimum number of samples for leaf nodes to 100. In contrast, LightGBM employs a gradient boosting framework to improve fitting accuracy by sequentially correcting prediction errors, often resulting in better prediction performance and computational efficiency in high-dimensional data. The model is configured with an ensemble of 1,000 decision trees, a learning rate of 0.03, a maximum tree depth of 5, and a maximum of 31 leaf nodes per tree. In order to enhance the generalization ability of the model, the subsample sampling rate and feature sampling rate are both set to 0.8.

To further characterize potential structural heterogeneity in the stock cross-section, a hybrid model method combining K-Means and LightGBM is proposed. This method rests on the premise that distinct stock groups may display divergent return-generating logics. Therefore, identifying group structures prior to modeling can probably enhance predictive accuracy and economic interpretability [4]. The process covers two stages. The first phase is to use K-Means to cluster stocks, which runs rapidly and can group stocks with similar structures in the market. And the clustering results also make it easier to understand how the market is divided [6]. In the second step, the LightGBM model is trained separately in each class which is then employed to capture the complex relationship between factors and returns. During the whole experiment, stock was divided into 5 categories, 500 trees were trained in each category, and the learning rate was set at 0.05. When it comes to predicting, focus on which clustering center the test sample is closest to at first, then classify it into that category, and then use the corresponding model to predict. Through the above framework, this paper intends to confirm whether cluster-enhanced modeling is better than the global single model under the condition of heterogeneous market structure, so as to provide empirical evidence for comprehending the return mechanism of different stock groups.

4. Empirical analysis

After completing data cleaning, feature construction, and model training, this paper systematically evaluates the predictive power and investment performance of each model through empirical testing. In terms of predictive effectiveness, the Spearman's rank correlation coefficient (Rank IC) is employed to measure cross-sectional ranking capability, calculated quarterly as the rank correlation between predicted value and actual returns [7]. Statistical significance is assessed using metrics such as the average IC and Information Ratio (ICIR). For investment performance, a zero-cost long-short portfolio is constructed by sorting stocks into deciles based on predicted values, taking a long position in the top decile and a short position in the bottom decile. Long-term performance is evaluated using quarterly long-short return spreads and cumulative net compounded value, while risk is measured by maximum drawdown. By integrating these two categories of metrics, the practical applicability of each model in stock return prediction is systematically examined.

4.1. Model prediction performance evaluation

The empirical results show that the average IC values of all models during the testing period were mostly positive. Among them, the LightGBM model performed best with an average IC of approximately 0.031 while that of the Random Forest and KMeans + LightGBM models also ranges around 0.02. To further examine the statistical significance of these prediction signals, this paper uses t-tests to determine whether the average quarterly IC values of each model are significantly positive. While the average IC values for LightGBM, Random Forest, and the hybrid model are all positive, suggesting some directional signal, t-test results indicate that none are statistically significant at the 5% level. This is likely due to the limited number of quarterly observations (around

DOI: 10.54254/2754-1169/2026.32361

12) instead of any fundamental flaw in the models themselves. Put differently, the positive mean IC suggests the models do pick up on certain cross-sectional patterns in earnings, even if the evidence isn't strong enough to pass conventional significance thresholds given the short sample period.

In this study, the cross-sectional predictive ability of different models concerning future excess returns of stocks is first assessed using the quarterly Rank IC (Rank Information Coefficient). As shown in Figure 1, the proportion of random forests with a positive quarterly IC reaches 75%, which is significantly higher than the 50% for linear models. Furthermore, nonlinear models maintain positive Rank IC values in most quarters and these values exhibit relatively controlled variations, indicating that these models possess certain predictive capabilities at the stock ranking level. By comparing the average IC values of different models, it can be noticed that the average quarterly IC value for the LightGBM model is 0.031, significantly higher than -0.003 for the Ridge model. This demonstrates that nonlinear models are more appropriate for capturing the complicated relationships between factors and returns in a multi-factor environment.

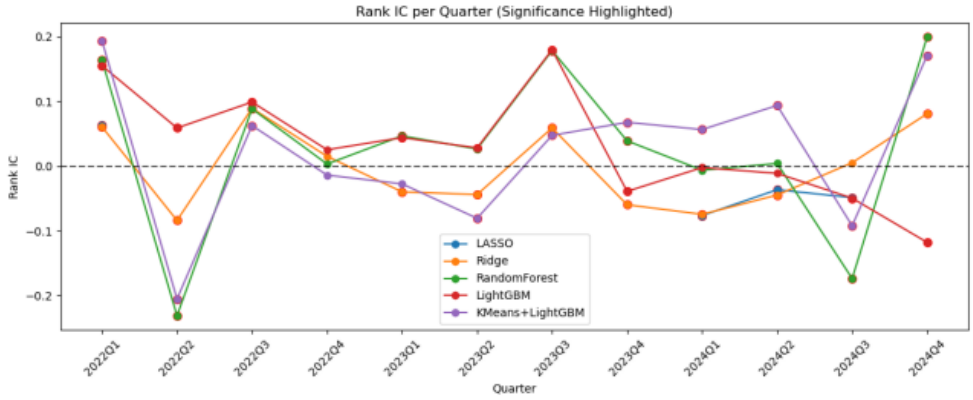

Figure 1. Seasonal variation of rank information coefficient

Apart from predictive power, the strategy of long-short combination can further examine whether the model has an economic alpha [8]. A long-short portfolio is when investors buy long stocks that are expected to rise in value (undervalued) and sell stocks that are expected to depreciate (overvalued stocks) [9]. Whether this investment can obtain the excess return that exceeds the reasonable return corresponding to the risk it takes is what investors are concerned about. Figure 2 shows that random forests, LightGBM and the hybrid model all achieved an overall increase in net value from Q2 2022 to Q4 2023, while the traditional linear model experienced either constant level or fluctuations over the two years. It is also found that a maximum return of 8% can be earned from stock investment based on long-short strategy and the LightGBM model between 2022 and 2023, yet no gain for the LASSO regression. From this perspective, nonlinear models do perform better in terms of sustainable profitability. However, it seems that the net value of all the models almost fell in the first half year of 2024 and then rose back up to their original level.

DOI: 10.54254/2754-1169/2026.32361

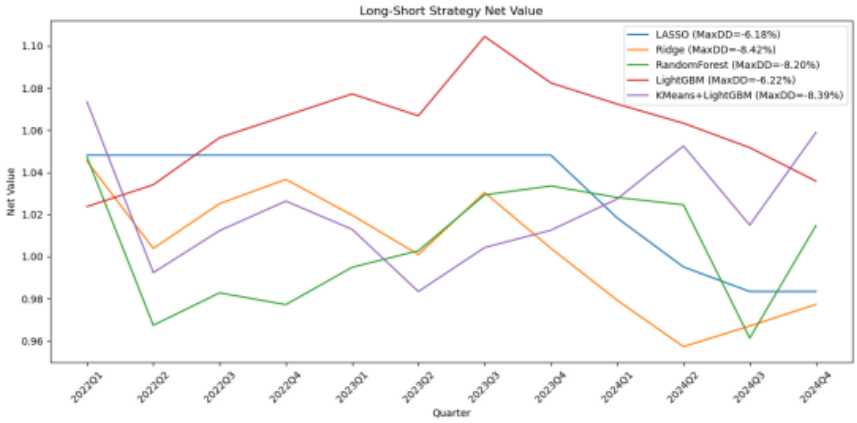

Figure 2. Long-short strategy net value and maximum drawdown

Apart from the level of return, the risk characteristics of the model are also a key consideration in investment decisions. The maximum drawdown reflects the extent to which a strategy could sustain losses in the most adverse scenario in history. From Figure 2 above, LightGBM and LASSO experienced similar maximum drawdowns, -6.18% and -6.22% respectively. And the remaining models went through a maximum drawdown near -8.30%. Although the hybrid model K-means + LightGBM had a high net cumulative return (over 4%) in the years of 2023 and 2024, it was accompanied by higher risk than revenue due to the maximum drawdown of -8.39%. In contrast, LightGBM has been more robust in controlling drawdowns, showing potential for risk-averse or institutional investors. Therefore, model selection needs to weigh the benefits and risks rather than just maximizing returns in business applications [10,11].

The above empirical results indicate significant differences among the models in terms of predictive power, investment returns, and risk characteristics. Nonlinear models such as Random Forest and LightGBM demonstrate robust performance, capturing the complex relationships between features and excess returns, which translates into sustained long-short portfolio returns. In contrast, signals from linear models tend to decay more rapidly. In addition, LightGBM achieves a more balanced trade-off between return generation and drawdown control. These findings underscore the necessity of comprehensive trade-offs in model evaluation, and this study provides an empirical foundation for subsequent tests for robustness and cross-market applicability.

5. Conclusion

This paper compares the out-of-sample performance of various models in cross-sectional earnings forecasts of China's A-shares under the rolling forecast framework of Expanding Window. The results show that the nonlinear model is superior overall than the linear model. Taking Rank IC as the core indicator, LightGBM and Random Forest obtain positive rank correlations in most quarters, with an average Rank IC of about 0.02–0.03, which is higher than LASSO and Ridge. This shows that in a market environment where multiple factors and potential non-linear relationships coexist, the tree model is more capable of portraying the complex relationship between factors and future returns. The long-short combination backtest results also show that the cumulative net value of the strategy constructed by the non-linear model shows an overall upward trend, with LightGBM performing relatively balanced between returns and drawdowns. However, although the average IC is positive, its t-statistic and p-value do not reach traditional significance levels, indicating that the forecast signal is directional in an economic sense, but its statistical robustness is still limited. This

DOI: 10.54254/2754-1169/2026.32361

result may be related to the smaller number of out-of-sample quarters and the larger changes in market structure during the study period.

This paper still has several limitations. First, the sample period only spans from 2020 to 2024, a time frame marked by considerable market volatility. As a result, the model may capture more short-term dynamics than long-term patterns. Second, the analysis relies primarily on quarterly fundamental and macroeconomic variables, without incorporating higher-frequency data such as momentum or reversal factors. This may partly explain why the hybrid model proposed in this study does not consistently outperform the standalone nonlinear model. Additionally, the long-short backtest does not explicitly account for transaction costs or short-selling constraints, which means the reported returns could be overly optimistic.

Future research can further assess the robustness and feasibility of the model by extending the sample period, enriching the information dimension, and adding transaction constraints that are closer to reality in backtesting. In conclusion, nonlinear machine learning methods show some potential in A-share cross-section prediction, but their long-term effectiveness still needs to be tested empirically.

References

[1] Balvers, R.J. and Huang, D. (2009) Evaluation of linear asset pricing models by implied portfolio performance. Journal of Banking & Finance, 33(9), 1586–1596.
[2] Du, Q., Wang, Y., Wei, C. and Wei, K.C.J. (2023) Machine learning, anomalies, and the expected market return: Evidence from China. Pacific-Basin Finance Journal, 82, 102168.
[3] Lu, S., Song, B. and Li, G. (2025) Enhancing multi-factor stock selection with transformer networks: A comparative analysis against traditional machine learning models. Procedia Computer Science, 266, 1028–1034.
[4] Ashrafzadeh, M., Sadrani, M. and Zolfani, S.H. (2025) Deep learning and machine learning models for portfolio optimization: Enhancing return prediction with stock clustering. Results in Engineering, 27, 106263.
[5] Melkumova, L.E. and Shatskikh, S.Y. (2017) Comparing Ridge and LASSO estimators for data analysis. Procedia Engineering, 201, 746–755.
[6] Toufighi, S.P., Khani, A.M., Rezasoltani, A., Sahebi, I.G. and Vang, J. (2025) Forecasting stock market anomalies in emerging markets: An OPTUNA-optimized isolation forest and K-means approach. Machine Learning with Applications, 22, 100770.
[7] Shekhovtsov, A. (2021) How strongly do rank similarity coefficients differ used in decision making problems? Procedia Computer Science, 192, 4570–4577.
[8] Eivazlu, R., Bajalan, S. and Mohammadi, M. (2024) Cross-sectional alpha dispersion of investment funds and performance evaluation: Is there a connection? (Evidence from an emerging market). Iranian Journal of Finance, 8(2), 23–46.
[9] Goumatianos, N., Christou, I. and Lindgren, P. (2013) Stock selection system: Building long/short portfolios using intraday patterns. Procedia Economics and Finance, 5, 298–307.
[10] Aritonang, P.K., Wiryono, S.K. and Faturohman, T. (2025) Hidden-layer configurations in reinforcement learning models for stock portfolio optimization. Intelligent Systems with Applications, 25, 200467.
[11] Doeswijk, R. and Swinkels, L. (2026) The risk and reward of investing. Journal of International Money and Finance, 160, 103453.