Investor Sentiment and Stock Return Prediction: Evidence from Machine Learning and VAR Models

Zixuan Yu

Department of Economics, Guizhou University, Guiyang 550025, China
DelogeJohnstone@gmail.com

Abstract. In the context of rapid information technology development and improved data acquisition capabilities, traditional linear-assumption-based asset pricing models gradually face challenges in short-term return forecasting. As an important behavioral factor affecting stock prices, investor sentiment is of increasing concern about its role in irrational volatility. Taking China's A-share market as the research object, this paper constructs a multi-dimensional factor system covering fundamentals, trading characteristics, and investor sentiment within a unified data framework, and applies a multifactor linear model, a random forest model, and an XGBoost model to predict stock returns separately. At the same time, the vector autoregressive model is introduced to characterize the dynamic relationship between investor sentiment and stock returns. Empirical results show that machine learning models outperform traditional linear models in out-of-sample prediction and portfolio performance, with the XGBoost model performing the best. After the inclusion of sentiment factors, the predictive power and risk-adjusted returns of the models are significantly improved. Further analysis shows that investor sentiment has a significant but short-term impact on stock returns. The results of this paper show that investor sentiment has a certain explanatory power in short-term earnings forecasting and improves the performance of the model to a certain extent.

Keywords: Investor Sentiment, Machine Learning, Stock Return Prediction, Vector Autoregression, Sentiment Dynamics Analysis.

1 Introduction

In recent years, with the continuous improvement of data acquisition ability and computing technology, the application of quantitative investment in capital market has been deepening. Traditional asset pricing research is mostly grounded in the Fama-French multifactor framework, which provides a solid theoretical foundation for explaining long-term returns. However, such research relies essentially on linear assumptions and presents certain limitations when facing complex market structures. In the context of a significant increase in the efficiency of information dissemination, stock prices not only reflect fundamental changes but also are affected to a large extent by investor behavior and sentiment fluctuations. This has been verified in early

© The Author(s) 2026
L. C. Yoong (ed.), *Proceedings of the 2026 2nd International Conference on Financial Innovation and Marketing Management (FIMM 2026)*, Advances in Economics, Business and Management Research 408,
https://doi.org/10.2991/978-94-6239-801-6_47

studies on the relationship between media sentiment and market performance [1]. Therefore, relying solely on traditional factor models can hardly meet the demands of short-term return forecasting. In recent years, machine learning methods have been gradually introduced into asset pricing research. Its advantages in dealing with nonlinear relationships and high-dimensional data make it more adaptable in earnings forecasting [2].

Focusing on the relationship between investor sentiment and stock returns, relevant research has been continuously expanded from two aspects: index construction and modeling methods. On the one hand, from the perspective of index construction, emotional indicators based on text information can reflect the expectations of market participants to a certain extent and have an impact on short-term returns [3]. Studies on the construction of sentiment indicators from the stock level have found that sentiment variables often contain certain non-fundamental information, which may push prices away from their intrinsic value in the short term [4]. On the other hand, from the perspective of the modeling method, the impact of sentiment on the market has periodic characteristics, showing the role of travel alienation in different cycles, which has been verified in multi-time scale analysis [5]. Machine learning models have been gradually used to characterize the nonlinear relationship between emotion and earnings. Some of them have built prediction models based on ensemble learning methods and achieved more stable results [6]. Further, the deep learning method gradually shows strong fitting ability in processing time series data [7].

Although some progress has been made in emotional measurement and prediction methods, most studies focus on a single model framework and lack a systematic comparison of the predictive capabilities of different methods under a unified data system. Existing studies have begun to analyze the relationship between emotion and return from a dynamic perspective, and found that there is a certain linkage between the two, but the relevant conclusions are still different in different market environments. Therefore, it is necessary to systematically test the relationship between emotion and return under a unified framework [8]. Based on this, this paper takes China's A-share market as the research object, constructs a multi-dimensional factor system including fundamentals, trading behavior, and investor sentiment, and uses a multi-factor linear model, a random forest model, and an XGBoost model to predict returns. At the same time, the vector autoregressive model is introduced to describe the dynamic relationship between sentiment and returns. Compared with the existing research, the main contribution of this paper is to systematically compare the prediction capabilities of different models under a unified data framework and to analyze the relationship between investor sentiment and stock returns from a dynamic perspective.

2 Data Processing and Variable Construction

2.1 Factor Construction and Investor Sentiment Indicators

In quantitative investment research, factor selection is a key step for model construction. In order to capture the factors affecting stock returns as

comprehensively as possible, this paper constructs an initial factor pool from three dimensions: company fundamentals, trading characteristics, and investor sentiment. In terms of fundamental factors, this paper selects variables that can reflect the company's profitability, growth ability, and capital structure from the financial index table of the CSMAR database, including ROE, ROA, asset liability ratio, operating income growth rate, and net profit growth rate. Relevant studies show that such financial indicators have strong economic implications in explaining the long-term returns of stocks, so they are widely used in multi-factor model research. In terms of trading characteristic factors, this paper uses daily stock trading data to construct a set of indicators that reflect market trading behavior, including daily stock returns, trading volume, turnover rate, and momentum indicators for different time windows (5-day, 10-day, and 20-day cumulative returns). Compared with financial data, such variables are updated more frequently and can reflect short-term market fluctuations and capital flows in a timelier manner. In addition to the above traditional factors, this paper further introduces investor sentiment indicators. In the market environment dominated by retail investors, market sentiment often has a significant impact on asset prices in the short term, so incorporating sentiment information into quantitative models helps to improve earnings forecasting ability. Compared with the traditional characterization method based on a single sentiment index, in recent years, some studies have begun to use deep learning methods to decompose investor sentiment in multiple dimensions, so as to capture the impact of different sentiment sources on market behavior in more detail [9].

In order to characterize the overall sentiment of the market, this paper selects variables such as turnover rate, trading volume, price deviation, and momentum indicators that can reflect the activity of market transactions and the characteristics of investor behavior, and constructs a comprehensive sentiment indicator after standardization. Specifically, first of all, each index is standardized by Z-score to eliminate dimensional differences. On this basis, this paper uses principal component analysis to reduce the dimension of the above variables, and selects the first principal component as the investor sentiment index. Compared with the simple weighting method, this method can more objectively reflect the comprehensive information of each index, while reducing the multicollinearity problem between variables on the premise of retaining the main information, so as to improve the representativeness and stability of the index. This indicator can comprehensively reflect the activity of market transactions and the change of investors' risk preference. When the sentiment index rises, it usually means that the market sentiment is more optimistic, while a lower sentiment index may correspond to a decline in risk appetite. After completing factor construction, this paper further screens the variables. Firstly, variables with high missing rates are excluded, and then the correlation coefficient matrix between factors is calculated to identify potential multicollinearity problems. The results show that the correlation between most factors is in a reasonable range, but there is a high correlation between some momentum indicators. Therefore, variables with higher correlation with future returns are retained in the subsequent modeling process to reduce redundant information in the model and improve the stability of the model.

2.2 Data Pre-Processing

Before the model training, the original data needs to be processed systematically. The first is sample screening. This paper selects all A-share listed companies from 2021 to 2025 as the research object, and excludes ST and *ST stocks and companies listed for less than one year to reduce the impact of abnormal samples. In data frequency processing, since financial data is usually disclosed quarterly and transaction data is daily data, this paper uses the forward filling method to continue the latest financial data to all trading days before the next financial disclosure. This method is more common in empirical financial research and can avoid future information leakage problems.

In terms of outlier processing, this paper performs 1% and 99% quantile tailing processing on all continuous variables to reduce the interference of extreme values on model training and maintain the stability of the overall distribution characteristics of variables. After the abnormal value processing is completed, all variables are standardized to make their mean value 0 and standard deviation 1. This processing helps to improve the stability of model training and avoid the impact of different variable scales on the algorithm.

In terms of data set division, this paper adopts the time sequence division method. Specifically, from January 2021 to June 2024 is used as a training set, from July 2024 to June 2025 is used as a validation set, and from July 2025 to December 2025 is used as a test set to ensure that the model does not use future information during the training phase, thereby improving the reliability of out-of-sample test results.

3 Model Setting and Emotional Analysis

3.1 Model Construction and Parameter Optimization

To compare the performance of different methods in stock return forecasting, this paper constructs three types of models for analysis: the traditional multifactor linear model, the random forest model and the XGBoost model. The traditional model incorporates the sentiment factor constructed in this paper into the Fama-French multifactor framework and predicts the subsequent excess stock return via ordinary least squares regression. The structure of the model is relatively simple, but it has good economic explanatory power. The random forest model improves the stability of prediction results by building multiple decision trees and integrating the results. In the process of model training, the verification set is used to optimize the key parameters. The final model parameters are set as follows: the number of decision trees is 500, the maximum depth is 10, and the minimum number of sample splits is 5. The XGBoost model is based on a gradient boosting framework, which has good performance in dealing with the nonlinear relationship and interaction effects between variables. In this paper, the learning rate is set to be 0.05, the maximum tree depth is 6, and the regularization parameters are determined through the verification set to avoid overfitting.

In the actual prediction process, this paper uses a rolling training method. The model is retrained every month with the latest historical data and predicts the next stage of earnings. Compared with one-time training, rolling training can better adapt to the changes of the market environment, thus improving the stability of the model in the stage outside the sample. After obtaining the forecast return, the stocks are sorted according to the forecast value, and the portfolio is constructed. The specific method is to select the 20% stock with the highest forecast return to build an equal-weight long portfolio, and rebalance the portfolio every day.

3.2 Dynamic Analysis of Investor Sentiment

In addition to the above prediction model, in order to further characterize the dynamic relationship between investor sentiment and stock returns, this paper introduces a vector autoregressive model. VAR model can describe the interaction between multiple time series variables at the same time, so it is widely used in financial time series research. Before model estimation, the stationarity test of variables is carried out first. The results of the ADF unit root test show that both the sentiment index and stock return series reach a steady state after the first order difference, so they can be used for VAR Modeling. In terms of lag order selection, this paper determines that the optimal lag order is second order according to the AIC and BIC criteria. In order to analyze the dynamic impact of investor sentiment shocks on stock returns, this paper calculates the impulse response function based on a VAR model to characterize the dynamic transmission path between variables.

The impulse response results of the VAR model show that when investor sentiment is positively impacted, stock returns show a significant positive response in the short term, and reach a peak in the first to second periods. Subsequently, the impact gradually weakened and stabilized after phase 5, indicating that sentiment has a significant but relatively short-term impact on stock returns. In addition, the results of variance decomposition show that investor sentiment variables can explain part of the proportion of stock return volatility, indicating that sentiment factors have a certain explanatory power in the formation of market volatility. This result shows that investor sentiment shocks may trigger market overreaction in the short term, but gradually return to fundamentals over time, thus reflecting certain mean reversion characteristics.

4 Empirical Results and Risk-Return Analysis

4.1 Back Testing and Portfolio Construction

In order to evaluate the performance of the model in actual investment, this paper constructs a portfolio based on forecast returns and carries out historical back testing. The research sample covers all A-share non-ST or *ST stocks from 2021 to 2025, with an overall sample size of more than 4000 stocks. The process of portfolio construction is as follows: first, the model is used to predict the next earnings of all stocks, and the forecast value is sorted from high to low. Subsequently, 20% of the

stocks with the highest predicted returns are selected to build a weighted bull portfolio. The holding period of the portfolio is one trading day, and the position will be adjusted at the opening of the next trading day.

In terms of back-testing benchmark selection, this paper uses the whole market equal-weight index in the CSMAR database as the market benchmark. Compared with the market value weighted index, the equal weight index can better reflect the overall performance of small and medium-sized market value stocks, so it is more suitable as a comparative benchmark for quantitative strategies. In addition, a unilateral transaction cost of 0.1% is added in the back testing process to simulate the Commission and impact cost in the real trading environment.

4.2 Profitability of Investment Portfolio

In order to more intuitively show the investment performance of different model strategies, this paper draws the cumulative return curve of the strategy portfolio and market benchmark, as shown in Fig. 1. Over the full sample period, the overall performance of portfolios constructed by machine learning models is better than that of the market benchmark. Among them, the cumulative return of the XGBoost model is the highest, and its return curve maintains steady growth in most periods of time. In contrast, although the traditional multi-factor model can obtain a certain excess return, the overall performance is relatively stable. In stages, during the period of intensified market volatility, the machine learning model can still maintain relatively stable earnings growth, indicating that it has certain advantages in capturing complex market structure and nonlinear relations.

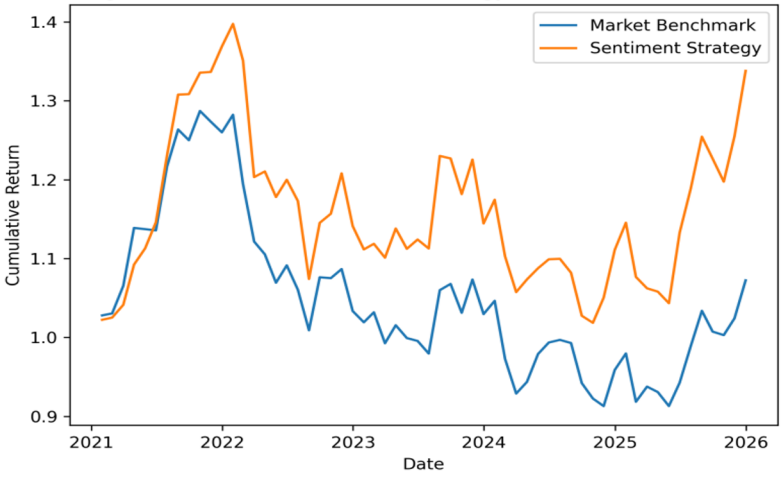


Fig. 1. Cumulative return of strategy and market benchmark (Picture credit: Original).

In order to further evaluate the profitability of the model, this paper calculates the annualized return of each strategy and the excess return relative to the market benchmark. The results show that the annual returns of the traditional multifactor model are about 8%, while the annual returns of the random forest model and

XGBoost model are about 12% and 14%, respectively. In different market environments, the performance of each model is also different. In the rising stage of the market, most models can obtain considerable returns; while in the stage of market adjustment, the machine learning model can still maintain a certain degree of excess returns, indicating that it has strong adaptability in a complex market environment.

After adding investor sentiment factors to the model, the overall forecasting ability has been significantly improved. Comparative experiments show that the annualized yield of the model is about 11% without introducing emotional factors. After adding emotional factors, the earnings were further improved, and the portfolio's Sharpe ratio was also significantly improved. This shows that investor sentiment plays an important role in short-term price fluctuations, and its inclusion in the quantitative model can improve the ability of earnings forecasting to a certain extent. This result confirms, to a certain extent, the research conclusion that deep learning has a stronger ability to depict nonlinearity in multi-factor earnings forecasting in recent years [10].

4.3 Risk-Adjusted Return Analysis

In addition to the level of return, risk control ability is also an important indicator to evaluate investment strategies. This paper further calculates risk indicators such as portfolio volatility, maximum retracement, and Sharpe ratio, and the results are shown in Table 1. Overall, the XGBoost strategy is the most prominent in risk-return performance. This strategy not only achieves the highest annualized yield, but also shows a lower level of volatility and a smaller maximum withdrawal. From the perspective of behavioral finance, investor sentiment may lead to stock prices deviating from fundamental value in the short term. When the model can effectively capture emotional changes, it can obtain certain excess returns in the process of price regression.

Table 1. Portfolio performance metrics.

Model	Annual return	Volatility	Max drawdown	Sharpe ratio
Linear Factor	0.081	0.152	-0.263	0.67
Random Forest	0.124	0.138	-0.221	0.98
XGBoost	0.141	0.123	-0.185	1.15

5 Conclusion

Under the unified data framework, this paper systematically analyzes the relationship between investor sentiment and stock returns. The empirical results show that compared with the traditional multi-factor model, the machine learning method has better performance in earnings forecasting and portfolio performance as a whole. Among them, the XGBoost model performs best in terms of cumulative return, annualized return, and risk-adjusted return. Further analysis shows that after the introduction of investor sentiment factors into the model, the prediction ability and

portfolio performance are improved, indicating that sentiment information has a certain explanatory power in short-term price fluctuations. Meanwhile, the dynamic analysis results based on the VAR model show that the impact of investor sentiment shocks on stock returns is mainly concentrated in the short term, showing a positive response and reaching a peak rapidly, and then gradually weakening and stabilizing. To a certain extent, this feature reflects the short-term overreaction of the market and the gradual return to fundamental value in the follow-up stage.

This paper makes a supplementary analysis of the role of investor sentiment in asset pricing from the perspectives of predictive performance and dynamic mechanism. On the one hand, comparing the performance of different models in a unified data system helps to more clearly identify the advantages of machine learning methods in dealing with non-linear relations and high-dimensional information. On the other hand, by introducing sentiment factors and combining time-series analysis, this research not only remains at the prediction level but also reveals the dynamic transmission path through which sentiment affects stock returns to a certain extent. It should be pointed out that there are still some limitations in this paper, such as emotional indicators are mainly based on transaction data, have not yet been included in text or media information, and the types of models are relatively limited. Future research can combine multi-source data and more complex model methods to characterize investor sentiment more carefully, so as to further enhance the ability of earnings forecasting and economic interpretation.

References

1. Tetlock, P.C.: Giving Content to Investor Sentiment: The Role of Media in the Stock Market. Journal of Finance 62(3), 1139–1168 (2007).
2. Bustos, O., Pomares-Quimbaya, A., Stellian, R.: Machine Learning, Stock Market Forecasting, and Market Efficiency: A Comparative Study. International Journal of Data Science and Analytics 20, 6815–6839 (2025).
3. Htun, H.H., Biehl, M., Petkov, N.: Forecasting Relative Returns for S&P 500 Stocks Using Machine Learning. Financial Innovation 10, 118 (2024).
4. Renault, T.: Sentiment Analysis and Machine Learning in Finance: A Comparison of Methods and Models on One Million Messages. Digital Finance 2(1), 1–13 (2020).
5. Kim, J.S., Kim, D.H., Seo, S.W.: Individual Mean-Variance Relation and Stock-Level Investor Sentiment. Journal of Business Economics and Management 18(1), 20–34 (2017).
6. Nguyen, T., Pham, H., Nguyen, D.: Investor Sentiment and Market Returns: A Multi-Horizon Analysis. Research in International Business and Finance 74, 102701 (2025).
7. Cohen G, Aiche A, Eichel R. Artificial intelligence models for predicting stock returns using fundamental, technical, and entropy-based strategies: A semantic-augmented hybrid approach[J]. Entropy, 2025, 27(6): 550.

8. Balasubramanian, P., Chinthan, P., Badarudeen, S., et al.: A Systematic Literature Survey on Recent Trends in Stock Market Prediction. PeerJ Computer Science 10, e1700 (2024).
9. Wang, Z., Guo, J., Wang, G., Shen, X.: Deep Learning-Based Multi-Dimensional Investor Sentiment and Stock Liquidity: Evidence from China. Quantitative Finance and Economics 9(4), 745–779 (2025).
10. Wang, K.: Multifactor Prediction Model for Stock Market Analysis Based on Deep Learning Techniques. Scientific Reports 15(1), 5121 (2025).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

