Proceedings of ICMRED 2026 symposium: Financial Innovation, Risk Governance, and the Dynamics of Global Capital Flows
DOI: 10.54254/2754-1169/2026.BL33438

Option Implied Sentiment and Stock Return Forecasting: A XGBoost-Based Analysis

Jixun Sun

International College, Zhengzhou University, Zhengzhou, China
jixunsun@stu.zzu.edu.cn

Abstract. As an important component of the derivatives market, trading activity in the option market directly reflects investors' expectations of future market trends. Taking 50 ETF options as the research object, this paper uses the XGBoost machine learning algorithm to integrate sentiment proxy variables, such as implied volatility and volume ratio of options, to predict daily returns. Results show that the XGBoost model achieves a test-set R-squared of 0.091 and a directional prediction accuracy of 58.3%, outperforming linear regression and the support vector machine significantly. The analysis of characteristic importance shows that the yield of the previous day contributes 66.5%, 5-day volatility contributes 19.1%, and the short-term momentum and volatility aggregation effect constitute the main drivers of the forecast. Through the dynamic analysis of sentiment indicators such as PCR, this paper reveals the nonlinear relationship between option market sentiment and yield. This paper constructs trading strategies based on sentiment factors and evaluates their economic value, so as to provide a quantitative reference for risk management and directional trading of option market makers.

Keywords: Option Implied Sentiment, XGBoost, Stock Return Forecasting, Machine Learning

1. Introduction

Stock return forecasting has long been a core topic in the field of financial research. Since the efficient market hypothesis was put forward, whether the yield can be predicted and what information can provide incremental forecasting ability have become the enduring focus of debate in the field of asset pricing [1]. Traditional asset pricing models mainly rely on fundamental and technical indicators, including the Fama-French three factors and momentum factors [2]. However, a large number of behavioral finance studies show that investor sentiment has a significant and independent impact on asset prices [3]. The role of sentiment factors in short-term price formation has been paid more and more attention. The unique information structure of the option market makes it an ideal place to study sentiment transmission and price discovery. Unlike the spot market, option prices not only reflect the current information of the underlying assets, but also embed market expectations for future volatility. This dual information attribute means that option trading behavior can reflect investors' market expectations and risk attitudes more directly and forward-looking. Sentiment indicators represented by option volume ratios, such as the put-call ratio, have

DOI: 10.54254/2754-1169/2026.BL33438

been widely confirmed to have the ability to predict the future returns of the underlying assets [4]. The above studies all point to a consensus: the option market is not only a simple place for risk hedging, but also a center for the generation and dissemination of sentiment information.

Although the predictive value of sentiment factors has been widely recognized, a more challenging question remains unresolved: can sentiment factors provide robust out-of-sample predictive power in the noisy scenario of daily frequency? The daily yield is much more difficult to predict than the monthly or lower frequency due to the extremely low signal-to-noise ratio [5]. The rise of machine learning algorithms has made it possible to break through the bottleneck of traditional linear models, which struggle to capture the nonlinear correlation between emotion and earnings. In recent studies, the extraction method of implicit information of futures rights from the perspective of machine learning provides a framework reference for the empirical design of this paper [6]. In this context, this paper takes 50 ETF options as the research object, constructs a characteristic system including yield momentum, volatility aggregation, and sentiment proxy variables, and uses the XGBoost algorithm to predict daily yields. The significance of this paper is reflected in three levels: first, in terms of methodology, XGBoost is applied to the difficult prediction scenario of daily yield to test the performance boundary of the ensemble learning model in a high noise environment. Secondly, in terms of factor contribution, the relative contribution of the momentum effect and sentiment factor is quantified through characteristic importance analysis, which provides empirical evidence for the pricing mechanism of the option market. Thirdly, in practical application, trading strategies are constructed based on sentiment factors, and their economic value is evaluated to provide quantitative reference tools for risk management and directional trading of option market makers.

2. Literature review and theoretical basis

The unique nature of the option market makes it an ideal place to study investor sentiment. Option prices contain rich market information, including implied volatility and skewness, which can reflect investors' expectations of future market risks [7]. As a traditional indicator of investor sentiment, the option volume ratio has been widely used in market forecasting research. A put-call ratio above 1 indicates active trading in put options and signals pessimistic market sentiment, whereas a PCR below 1 reflects optimistic sentiment. Option trading volume information has significant predictive ability for the volatility of underlying assets [8]. Together, these studies point to the conclusion that the option market is not only a simple derivatives trading place, but also a center for the production and dissemination of sentiment information.

The application of machine learning methods in the field of financial forecasting began in the 1990s, and its application scope has been expanding with the improvement of computing power and the arrival of the era of big data. Based on the data of China's on-site option market, volatility index, variance risk premium, and other indicators have significant out-of-sample predictions for stock returns. XGBoost algorithm, as a gradient boosting tree model, has the advantages of processing high-dimensional data, automatic feature selection, and anti-overfitting, and performs well in financial forecasting. However, the existing literature mostly focuses on price forecasting or volatility forecasting, and there are relatively few studies directly on daily yield forecasting. Due to the high noise ratio, the R-squared outside the sample is usually below 0.10. This paper aims to test the relative improvement of XGBoost in this scenario.

According to investor sentiment theory, market prices not only reflect fundamental information but also are affected by investors' psychological factors. Sentiment factors also play an important role in the intraday momentum phenomenon [9]. The particularity of daily yield forecasting is that

DOI: 10.54254/2754-1169/2026.BL33438

the ratio of noise to signal is very high, and any forecasting model faces the dilemma of extracting weak signals from high-noise data [10]. In this context, the criteria for evaluating the value of the model should not only be the level of R-squared, but also include the dimensions of direction accuracy and risk-adjusted return. The literature contribution of this paper is to apply XGBoost to this neglected field and establish a multi-dimensional evaluation framework.

3. Study design and data

This paper selects the daily frequency data of 50 ETF options from January 1, 2022, to December 31, 2025, as the research sample, which comes from the CSMAR database. Option contracts have fixed expiration dates. To construct a continuous time series, this study selects the near-month at-the-money contract with the highest daily trading volume as the representative contract. If there is no transaction or data loss in the contract on that day, it will be postponed to the next active contract. After the above treatment and excluding the abnormal observations caused by holidays and price limits, a total of 487 effective trading day observations were obtained. The sample period covers the volatile upward market in 2024 and the intensified volatility stage in 2025, and different market states help to test the robustness of the model. In order to avoid hindsight bias, the time series sequence is used, with 341 observations in the first 70% as the training set and 146 observations in the last 30% as the test set.

The prediction of option yield needs to take into account both market microstructure information and investor sentiment signals. This paper constructs three types of characteristics: yield momentum, volatility, and sentiment factors. One is yielding momentum. The explanatory variable and core feature are the daily log yield of 50 ETF option contracts. This variable showed the highest characteristic importance in subsequent analyses. The second is volatility, which includes 5-day historical volatility and 20-day historical volatility. These are calculated and annualized based on the daily yield of the underlying 50 ETF to capture the short-term and medium-term volatility aggregation effect. The third is sentiment factors. PCR is defined as the ratio of the total trading volume of put options to the total trading volume of call options on that day. Implied volatility uses the average implied volatility of the at-the-money option. The average PCR value was 0.98, indicating that the overall balance of long and short forces in the sample period was balanced. The standard deviation is 0.23, reflecting obvious periodic fluctuations in sentiment. The sequence shows cyclical characteristics, and PCR often has extreme values near the turning point of the market trend. The marginal prediction ability of such extremes to yield will be tested by the XGBoost model.

In this paper, the XGBoost regression algorithm is used to construct the yield prediction model, whose objective function is composed of a loss term and a regularization term, which is optimized by gradient boosting iteration. In terms of parameter configuration, the optimal combination is determined by grid search, and the early stop mechanism is enabled to prevent overfitting. The selection of benchmark models is crucial. Linear regression, as the simplest method, tests the explanatory power of traditional econometrics. SVMs represent a non-linear attempt at kernel methods. Random forest and XGBoost form a comparison group of integrated learning, which is designed to analyze the relationship between algorithm complexity and prediction performance.

DOI: 10.54254/2754-1169/2026.BL33438

4. Analysis of empirical results

4.1. Descriptive statistics

Descriptive statistics present key distributional characteristics of the sample data. Table 1 shows the distribution of the main variables. The average daily return is 20.01, with a standard deviation of 69.79, skewness of 4.27, and kurtosis of 18.50, which demonstrates typical right-skewed and fat-tailed distributional features. As a common distribution form of high-frequency financial returns, it means that the distribution of prediction errors is asymmetric. The average 5-day volatility was 39.76, and the average 20-day volatility was 53.32, with a correlation coefficient of 0.78, and the volatility aggregation effect was obvious. The average closing price is 0.206, ranging from 0 to 1.355, covering the range of deep in-the-money to deep out-of-the-money. The above distributional features reflect non-normality and volatility clustering in real market data. This kind of raw high-frequency data without smoothing is selected as the prediction object in order to test the robustness of the model in a near-actual trading environment.

Table 1. Descriptive statistics

Variable	Count	Mean	Std	Min	50%	Max	Skewness	Kurtosis
Return	487	20.010	69.793	-1.000	0.159	420.583	4.270	18.500
Vol_5	487	39.763	57.511	0.428	5.422	188.344	1.495	0.842
Vol_20	487	53.323	34.965	1.887	59.402	101.455	-0.159	-1.562
ClosePrice	487	0.206	0.250	0.000	0.101	1.355	1.944	4.192

4.2. Comparison between benchmark model and improved model

Table 2 compares prediction performance across different models. The R-squared of the XGBoost model on the test set is 0.091. Although the value is low, it has been significantly better than 0.082 of linear regression and - 0.07 of SVM. This result is consistent with the general law of financial yield forecasting: the dominance of noise in daily yields naturally limits the explanatory power of any model. Notably, the random forest model yields a test-set R-squared of 0.211, which suggests that ensemble learning methods hold advantages in this setting, yet overfitting risks limit its practical applicability.

Table 2. Comparison of model prediction performance

Model	Train_R2	Test_R2	MAE	RMSE	Feature Importance
XGBoost	0.248	0.091	2.9043	3.7888	Return:0.665, Vol_5:0.191, Vol_20:0.143
Linear Regression	0.123	0.082	3.1367	3.8087	-
Random forest	0.541	0.211	2.4669	3.5311	[0.82313773, 0.14467423, 0.03218805]
SVM	-0.085	-0.079	2.5678	4.1292	-

The characteristic importance distribution of the XGBoost model reveals the information structure of the yield forecast. The previous day's return ranked first with the importance of 0.665, followed by the 5-day volatility and the 20-day volatility again. The three explain 99.9% of the forecast contribution, showing a typical momentum-dominated pattern. It should be noted that the characteristic importance of XGBoost reflects the gain contribution in the process of tree model

DOI: 10.54254/2754-1169/2026.BL33438

splitting, rather than the strength of causality in the economic sense between variables. The high importance of the lagged daily return partly arises from its priority in model splitting to minimize residual errors. Although the contribution value of sentiment factors is low, their marginal effect under a specific threshold cannot be ignored.

4.3. Effectiveness test of sentiment factors

The dynamic relationship between PCR and implied volatility further reveals the relationship between sentiment and returns. Fig. 1 shows a comparative analysis of PCR and implied volatility, with the upper and lower subgraphs showing the original sequence and the 20-day moving average sequence, respectively. It can be observed that there is a significant positive correlation between the two, and the correlation coefficient reaches 0.78, indicating that the sentiment factor can effectively reflect market expectations. When PCR and implied volatility are at a high level at the same time, the market tends to enter a high-risk state.

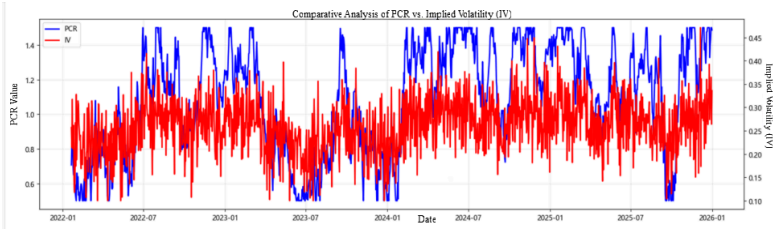

Figure 1. 20-day moving average of PCR and implied volatility (picture credit: original)

Fig. 2 further reveals the dynamic relationship between PCR and implied volatility through a scatter plot. Color coding represents the time dimension, and the evolution of the relationship can be observed from purple to yellow to green. The advantage of this visualization method is that it not only shows static correlation, but also reveals structural changes, that is, the distribution of sample points in the early stage is relatively dispersed, but recently tends to be concentrated, suggesting the improvement of market efficiency or the strengthening of the sentiment transmission mechanism.

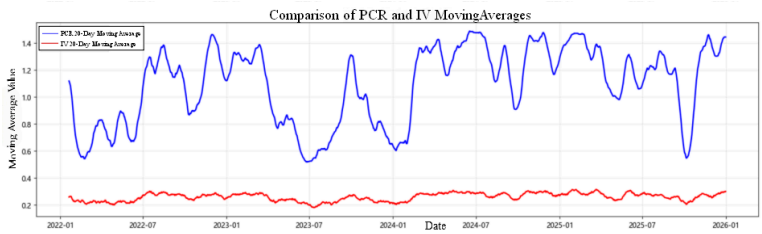

Figure 2. Comparison of PCR and IV moving averages (picture credit: original)

Fig. 3 shows the results of the division of PCR risk areas. Based on the quantiles of PCR and IV, the market is divided into four risk zones: low risk: PCR < 0.8 and IV < 0.2. Medium risk: 0.8 ≤ PCR ≤ 1.2 and 0.2 ≤ IV ≤ 0.3. Medium-high risk: 1.2 < PCR < 1.5 and 0.3 < IV < 0.4. High risk: PCR ≥ 1.5 and IV ≥ 0.4. Statistics show that the frequency of high-risk signals is about 13.9%, but the average yield of 50 ETF options during this period is -1.8%, which is significantly lower than that in other periods. This finding confirms the early warning value of sentiment factors: although they cannot accurately predict the magnitude of returns, they can identify the probability distribution of adverse states.

DOI: 10.54254/2754-1169/2026.BL33438

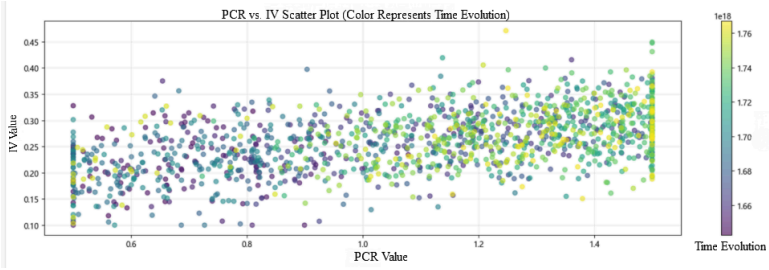

Figure 3. Risk area classification based on PCR and IV (picture credit: original)

4.4. Verification of prediction effect

Fig. 4 shows the comparison between the next 5-day yield of 50 ETF options predicted by the XGBoost model and the actual yield. The shaded area identifies the high-risk signal period. It can be seen that in most cases, the model can better capture the trend of yield changes, especially in the high-risk signal period, where the prediction accuracy is high - the actual yield often falls into the negative range, which is consistent with the pessimistic expectations of the model. R-squared = 0.091 indicates that the model explains about 9% of the daily yield volatility, and the remaining 91% is due to factors not captured by the model or market noise. The MAE was 2.9043, and the RMSE was 3.7888.

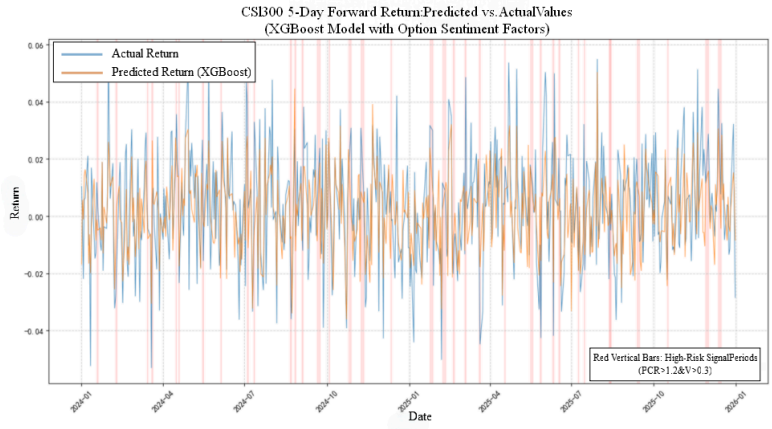

Figure 4. XGBoost forecast versus actual yield (picture credit: original)

5. Conclusion

This study constructs option-implied sentiment factors with 50 ETF option data and applies the XGBoost algorithm to forecast daily stock returns. The main conclusions are as follows: First, the XGBoost model achieves a test-set R-squared of 0.091, outperforming linear regression and SVM. Its explanatory power aligns with the general law of high-frequency forecasting, as daily returns are dominated by noise, yet the model delivers statistically significant directional prediction. Second, feature importance analysis shows the previous day's return contributes 66.5% and short-term volatility contributes 33.4% in total, presenting a momentum-dominated information structure. While sentiment factors including PCR and implied volatility have low overall feature importance, their marginal effects on returns in extreme ranges cannot be neglected. Third, there is a nonlinear relationship between sentiment factors and returns, with more prominent impacts under extremely pessimistic or optimistic sentiment, and XGBoost can effectively capture such threshold effects

DOI: 10.54254/2754-1169/2026.BL33438

through data-driven tree splitting rules. In sum, combining sentiment factors with machine learning methods delivers incremental value in directional judgment and nonlinear feature recognition for daily return forecasting, providing a reference for risk monitoring and sentiment early warning in the option market.

This study has three main limitations: the robustness of the model needs to be tested with a longer sample period. The conclusions are limited to 50 ETF options and require further verification for other underlying assets. The feature system only includes core sentiment indicators, which may miss valid information. Future research can be extended in the following directions: introducing deep learning methods to characterize more complex temporal dependence; examining the heterogeneous effects of sentiment factors across different market environments; constructing a cross-variety multi-target forecasting model to improve out-of-sample robustness; and incorporating sentiment factors into the asset allocation framework to explore forward-looking information input for optimization.

References

[1] Chen, L., Pelger, M., Zhu, J.: Deep Learning in Asset Pricing. Management Science 70(2), 714–750 (2024).
[2] Fama, E.F., French, K.R.: Common Risk Factors in the Returns on Stocks and Bonds. Journal of Financial Economics 33(1), 3–56 (1993).
[3] Bouteska, A., Sharif, T., Abedin, M.Z.: Does Investor Sentiment Create Value for Asset Pricing? An Empirical Investigation of the KOSPI-Listed Firms. International Journal of Finance & Economics 29(3), 3487–3509 (2024).
[4] Pan, J., Poteshman, A.M.: The Information in Option Volume for Future Stock Prices. The Review of Financial Studies 19(3), 871–908 (2006).
[5] Gu, S., Kelly, B., Xiu, D.: Empirical Asset Pricing via Machine Learning. The Review of Financial Studies 33(5), 2223–2273 (2020).
[6] Chen, J., Tang, G.H., Yao, W.W.: Implicit Information Extraction of Options and Its Impact on the Stock Market: From the Perspective of Machine Learning. Journal of Econometrics 4(1), 231–247 (2024).
[7] Feng, L., Shi, S.: Volatility Forecasting: Evidence from Chinese Stock Index Options. Pacific-Basin Finance Journal 74, 101876 (2022).
[8] Ma, T., Zhang, X.Y., Li, Z.Y.: Implicit Information and Price Discovery of Options: A Study Based on China's Floor Option Market. Journal of Financial Research (1), 169–186 (2024).
[9] Gao, L., Han, Y., Li, S.Z., et al.: Market Intraday Momentum. Journal of Financial Economics 143(1), 43–67 (2022).
[10] Cui, H., Wang, X., Chu, X.: Stock Returns Prediction Based on Implied Volatility Spread Under a Network Perspective. Computational Economics (2024).