International Business & Economics Studies
ISSN 2640-9852 (Print) ISSN 2640-9860 (Online)
Vol. 7, No. 5, 2025
www.scholink.org/ojs/index.php/ibes

Original Paper

Profit Prediction and Mechanism Analysis of Stock Price-Volume Patterns Based on Random Forest

Ji Hanlin[1] & Luo Yuxin[1]

[1] School of Management, University of Shanghai for Science and Technology, Shanghai 200093, China

Received: August 22, 2025 Accepted: October 08, 2025 Online Published: October 31, 2025

doi:10.22158/ibes.v7n5p121 URL: http://dx.doi.org/10.22158/ibes.v7n5p11

Abstract

This paper collects 25,598 data entries of stocks that experienced a certain magnitude of price decline, followed by volume contraction and subsequent volume expansion, starting from 2008. It investigates the impact of various influencing factors on the subsequent profitability of stocks after volume expansion. First, an indicator system covering trading volume, market index, K-line patterns, and financial metrics is constructed. A random forest model is then established to identify the top 11 important features among 52 features. These top features include the 1-2 day market index and market profit index, the 0-2 day price change rate, the price decline relative to the highest price in the previous 60 trading days, the 1-2 day K-line entity ratio, and the volume ratio. Second, the forward selection method is used to calculate the AUC_ROC values as the number of features increases from small to large. It is found that the AUC_ROC reaches the first peak when the number of features is 3-4 core technical indicators. After that, the AUC_ROC decreases due to the addition of noise information, and then rises to the second peak with the introduction of financial indicators. Finally, the SHAP method is applied to analyze the marginal contribution of the features included in the second peak. The results show that within a certain range, the marginal contributions of K-line entity, volume ratio, price decline, earnings cash coverage multiple, and interest coverage multiple to the future return rate are positive; however, when these indicators are too high or too low, their marginal contributions turn negative.

Keywords

price-volume pattern, machine learning, stock price prediction, profit mechanism

1. Introduction

As one of the core hubs of the modern economic system, the stock market's growth and stability profoundly affect the efficiency of capital allocation. Stock market fluctuations impact the macroeconomic market, enterprises, and investors in different dimensions. For individual investors, manipulation of the stock market by major institutional funds can prevent them from predicting future market movements through technical indicators or corporate panel data, resulting in losses and a flow of funds towards institutional investors. For enterprises, financial indicators help assess investor confidence and capital recovery in the market, which aids management in making targeted improvements in asset liquidity, inventory turnover, and cash flow, thereby enhancing the company's financial health in the stock market and enabling a new round of financing. At the macroeconomic level, stock market fluctuations not only reflect fundamental conditions but also provide a profound feedback effect on the real economy. When the stock market is healthy and steadily rising, it can stimulate consumer spending through the wealth effect and promote further enterprise investment via the Tobin’s q effect, achieving a virtuous cycle of capital market development and real economic prosperity.

Traditional technical analysis often focuses on the influence of technical indicators and price-volume data, making it difficult to capture the interactive effects of factors such as manipulation by major funds and financial risks. Single technical indicators are easily affected by noise, and the lack of fundamental data such as financial factors further reduces the robustness of predictive models. With increasing market complexity, how to construct an analytical framework that integrates multidimensional information while balancing practicality and interpretability has become key to unlocking the “ack box” of price-volume relationships.

Numerous studies have examined stock market performance in the context of the relationship between price and volume from various perspectives. In terms of research periods, previous studies have mainly concentrated on the normal market conditions of a specific economic cycle. For example, Sun, Chen, and Liu (2019) introduced foreign exchange market trading volume into the existing stock market theory to analyze its impact on stock prices and found that incorporating foreign exchange market volume information helps predict stock price fluctuations. However, their linear regression model for predicting stock price fluctuations was based on the normal market period from 2015 to 2017 and did not provide cross-cycle factor analysis. Regarding influencing factors, although Zheng, Tan, and Chen (2018) analyzed the long-term relationship between financial indicators and the stock market, their focus was on a long-term perspective. Due to short-term speculative sentiment and herd behavior, short-term fluctuations in technical indicators can increase the likelihood of extreme investor emotions, thereby affecting the long-term health of the stock market and corporate financing. Thus, the combined effect of technical indicators and financial factors in the short term is crucial. Yang (2020), in her study on the price volatility of agricultural listed companies, analyzed the causes of agricultural stock price fluctuations, but her focus was on macro factors, such as the level of support the capital market provides to listed companies. Zhang, Wu, Cai et al. (2021) primarily analyzed stock price trends based on information channels such as media and analysts, while trading volume to a certain extent reflects social attention and overall investor sentiment. Technical indicators of the stock itself reflect its historical volatility, market indices reflect the overall market health, and financial indicators reflect the management level of the company's executives and the operation capability of internal funds. Combining stock patterns, market indices, trading volume, and financial conditions allows for the prediction of future market health and the selection of quality stocks from multiple dimensions, including the macroeconomic level, social attention and investor psychology, and the company's actual operating capability.

Currently, most articles combine volume and price relationships to predict and analyze stock market fluctuations. Yan, Qin, and Song (2021) selected indicators such as price-to-earnings ratio, price-to-book ratio, and turnover rate, and analyzed the correlations among these indicators to reduce model redundancy. However, they did not analyze the interaction and lag effects between volume and price-related indicators, meaning that a certain interaction pattern between volume and price could significantly increase the likelihood of future stock price rises with a lag, indicating a black box problem between volume and price. Zou and Li (2023) established an RF-SA-GRU hybrid model, significantly improving accuracy in time series predictions, but did not solve the model’s black box problem. Li, Gao, Gao et al. (2023) developed a regression model based on the maximum relevance entropy criterion to predict market indices, taking into account the impact of the macro market on individual stock indices, but did not consider the factors influencing market index fluctuations and their causal relationships, i.e., whether market index fluctuations cause individual stock changes or the good performance of individual stocks boosts the market index. Moreover, market indices are often influenced by key stocks controlled by major funds, and therefore fail to reflect the fluctuation situation of most stocks in the real market.

In terms of model selection and influencing factors, Chi, Yang, and Zhou (2024) used Type_error_II minimization as the criterion for random forest grid search and screened the optimal indicator set through forward selection. Bao, Guo, Xie et al. (2020) adjusted LSTM model parameters using a genetic algorithm to avoid falling into local optima during training, ultimately obtaining the buy or sell signal for the next moment. This study focused on the prediction accuracy of the LSTM model, without providing interpretability analysis of influencing factors, and only predicted stock returns for the next trading day, failing to effectively capture the situation of a significant price increase after volume expansion. Xu and Tian (2021), as well as Wang and Huang (2021), predicted future stock price movements from the perspectives of financial news and investor sentiment, respectively; among them, Xu Xuechen and Tian Kan improved prediction accuracy by combining sentiment features from financial news with stock market trading data. However, they did not address the 'black-box' problem of model predictions, and their conclusions were mainly based on large-cap, highly liquid leading companies. Wang Lianlian and Huang Weiqiang mainly focused on investor sentiment, incorporating it into an early warning indicator system and using the SHAP method to open the black box. This study further analyzes the marginal contribution of features affecting stock return volatility using the SHAP method, opening the black box of factors influencing volume expansion after a period of low volume and a certain range of stock price decline.

It can be seen that although some scholars have predicted and analyzed stock market fluctuations by combining trading volume, price, and financial indicators, they have not deeply explored the interaction and time-lag effects of volume and price fluctuations. The main contributions of this study are: First, it identifies a phenomenon in the stock market: after a period of low trading volume, a sudden surge in volume is often accompanied by a significant rise in stock prices; if the stock price has sharply declined in previous trading days, the likelihood of a price increase in the following few trading days significantly rises. Second, based on the overall market index of gains and losses, to reduce the impact of weight stocks manipulated by major funds, a market earnings index is introduced, which more accurately reflects the daily price movements of all stocks and analyzes the cross-cycle market from 2016 to 2024. Finally, the SHAP method is employed to further analyze the influencing factors of volume surges following low-volume periods, visualizing the linear and nonlinear relationships between indicators and their marginal contributions, ensuring both predictive accuracy and interpretability of the model. The goal is to help individual investors reduce interference from major institutional manipulations when making decisions, assist companies in adjusting management decisions and operational models based on key short-term financial indicators, and enhance the efficiency of capital markets.

2. Model Construction

2.1 Judgment on a Rebound after Hitting a Bottom

A massive trading volume in a stock over a period of time is called a "heavenly volume". When a heavenly volume occurs, a sky-high price often signals a selling opportunity. Conversely, if a stock's trading volume significantly decreases compared to previous periods, it is called a "low volume". Typically, a low volume that appears after a certain drop in stock price may indicate a buying opportunity. By observing a large number of candlestick charts and the relationship between volume and price trends, it has been found that in most cases, if a few days show a low volume pattern, the trading volume and stock price are likely to increase rapidly in the following days. For example, take 600023 Zhejiang Energy Electric Power:

Figure 1. Stock Price Trend of Zhejiang Energy Power from July to September 2024

In Figure 1, the candlestick pattern shows a bearish candle (illustrated in green), indicating that the closing price is higher than the opening price; conversely, a bullish candle (illustrated in red) indicates that the opening price is higher than the closing price. The stock has been in a downtrend since July 15, 2024. Although there were brief periods of upward movement, the overall trend remained sharply downward. The closing price on August 28, 2024, was 6.25, and the closing price on September 23, 2024, was 5.65, a decline of 10.62%. The trading volume on September 23, 2024, was 604,399 lots (one lot equals 100 shares, the same below), and the highest trading volume in the previous five trading days was on September 12, 2024, at 336,831 lots, a rise of 79.44%. This article screens stocks based on trading volumes exceeding those of the previous five trading days. Before September 23, 2024, the stock price had dropped sharply, and the trading volume on that day exceeded the maximum of the previous five trading days. Subsequently, trading volume increased, and the stock price rose significantly.

In order to further analyze the potential for increased trading volume following days of low volume, this study collected trading volumes and stock prices from trading days since 2008, including the five trading days preceding each day. A total of 25,598 data points were selected as research samples, where the recent stock price had dropped more than 10% and the trading volume exceeded the maximum volume of the previous five trading days. Generally, the total of stock trading commissions, stamp taxes, and transfer fees does not exceed 0.4% of the transaction amount. This study focuses on the maximum returns from T1 to T7, calculating the interval statistics by subtracting the 0.4% fee from the maximum possible return over the next seven trading days. Then, the proportion of samples within each return interval relative to the total sample size is calculated:

Table 1. T1 Yield Range Proportion (After Fees)

Yield (after fees)	Count Proportion	Merger Share
< -0.1	0.08%	19.54%
-0.1 - 0	19.45%	
0 - 0.05	65.32%	80.46%
0.05 - 0.1	13.25%	
> 0.1	1.89%	

In Table 1, after deducting a 0.4% handling fee, the number of stocks with a positive maximum T1 return accounts for 80.46% of the total number of stocks. This indicates that after a period of stock price decline and a subsequent period of reduced trading volume, there is a high probability of stock price rising again when trading volume increases.

2.2 Quantification of Influencing Factors

This paper uses a random forest model to consider the impact of multiple factors on stock profitability and builds an indicator system including trading volume, market index, K-line patterns, and financial indicators:

1. Trading Volume

Volume indicators include: current day's trading volume (T0), trading volumes from the previous 1-5 trading days, and the volume ratio. The volume ratio refers to the ratio of the current day's trading volume to the average trading volume of the previous 5 trading days.

2. Market Index

Market index indicators include the market profitability index and market return index for the next 1-2 days. The market return index can be directly obtained from the Tushare library. Considering that high-market-cap blue-chip stocks can easily inflate the market index, causing a situation where most constituent stocks decline while the index rises, this study uses the proportion of rising stocks in the A-share market as the market profitability index to better reflect actual market conditions.

3. Candlestick Patterns

Candlestick patterns encompass price fluctuations and the relationship between stock prices and moving averages. Indicators include: daily price change (T0 change), opening prices for T0-T2, high prices for T0-T2, low prices for T0-T2, closing prices for T0-T2, price decline, highest and lowest closing prices from the previous 60 trading days, price changes for the next 1-2 days, entity sizes for the next 1-2 days, and whether prices are above the 5-day and 10-day moving averages.

4. Financial Indicators

Financial indicators are divided into three dimensions: operating capacity, profitability, and solvency. Operating capacity indicators include: inventory turnover ratio, accounts receivable turnover ratio, operating cycle, current assets turnover ratio, and total assets turnover ratio. Profitability indicators include: operating profit margin, cost-to-profit ratio, return on total assets, return on equity, return on capital, and cash-to-earnings coverage ratio. Solvency is divided into short-term and long-term: short-term solvency indicators include current ratio, quick ratio, and cash ratio; long-term solvency indicators include asset-liability ratio, equity multiplier, and interest coverage ratio.

Table 2. Definition of Feature Variables

Primary Indicator	Secondary indicator	Indicator Explanation
Trading volume	Volume_0	
	Volume_1-Volume_5	Volume_0/Avg(Vloume_1-5)
	Volume_Ratio	
Stock market index	Volume_1_day, Volume_2_day	
	Market_Return_1,	
	Market_Return_2	
	Profit_Index_1, Profit_Index_2	
	Return_0	
K-line pattern	Open_0-Open_2	
	High_0- High_2	
	Low_0-Low_2	
	Close_0- Close_2	
	Drop	Prev60_High/Close_0-1
	Prev60_High	
	Prev60_Low	
	Return_1	Close_1/Close_0-1
	Body_Size_1	abs(Open_1-Close_1)/Close_0
	Return_2	Close_2/Close_1-1
	Body_Size_2	
	Above_MA5	
	Above_MA10	
	Current_Ratio	Current Assets / Current Liabilities
	Quick_Ratio	(Current Assets - Inventory) / Current Liabilities
	Cash_Ratio	(Cash + Trading Financial Assets) / Current Liabilities
	Debt_Ratio	Total Liabilities / Total Assets
Financial indicators	Equity_Multiplier	Total Assets / Shareholders' Equity
	Inventory_Turnover	Cost of Sales / Average Inventory

	Receivables_Turnover	Operating Revenue / Average Accounts Receivable
	Operating_Cycle	(Inventory + Accounts Receivable) Turnover Days
	Current_Asset_Turnover	Operating Revenue / Average Current Assets
	Total_Asset_Turnover	Operating Revenue / Average Total Assets
	Operating_Profit_Ratio	Operating Profit / Operating Revenue
	Cost_Profit_Ratio	Total Profit / Total Cost
	Cash_Coverage_Ratio	Net Cash Flow from Operating Activities / Net Profit
	ROA	EBIT / Average Total Assets
	ROE	Net Profit / Average Shareholders' Equity
	ROC	EBIT / (Total Assets - Current Liabilities)
	ICR	EBIT / Interest Expense

Subsequent research will be based on this feature system to construct a random forest model to analyze the impact mechanism of profitability after a stock price decline followed by a period of low trading volume and then increased volume.

2.3 Random Forest Model Construction

1. Label Level Code

To reduce the risk of model overfitting, this study uses a 0-9 grading coding method. First, the distribution characteristics of the indicators are analyzed through skewness and kurtosis. Skewness measures the extent to which the data distribution deviates from symmetry, while kurtosis measures the likelihood of extreme values in the data, as shown in Figure 2.

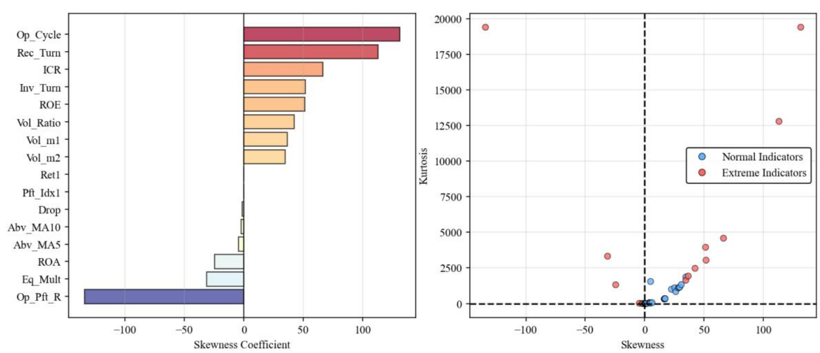

Figure 2. Distribution Characteristics of Indicator Skewness and Kurtosis

In Figure 2, most indicators exhibit an extremely skewed normal distribution, with the overall right skewness of the indicators being significantly higher than the left skewness. Among them, Operating_Cycle and Receivables_Turnover show a pronounced right skew; Operating_Profit_Ratio shows an extreme left-skewed distribution. Traditional fixed-interval grade coding can lead to large differences in the number of categories for coding, making it difficult for the model to learn features from categories with very few samples. Therefore, in this study, a threshold of 2,600 is used to divide the grades, sequentially selecting the 2,600 × n-th smallest data point from a 25,598-dimensional vector as the basis for division, splitting the data into 10 grades ranging from 0 to 9. This study selects a relatively high daily return (2%) as the threshold. If the maximum stock price return on days T1-T7 exceeds 2% (including a 0.04% transaction fee), it indicates that the trading day meets the expected stock price patterns of low volume followed by increased volume, and it is coded as 1; otherwise, it is coded as 0.

2. Construction of a Yield Prediction Model

(1) Meaning of Type_error_I and Type_error_II

Table 3. Definition of Type_error_I and Type_error_II

	Loss	Profit
Loss	True positive (TP)	False negative(FN)
Profit	False positive(FP)	True negative(TN)

Type_error_I is the number of profitable stocks predicted as losing stocks (FP); Type_error_II is the number of losing stocks predicted as profitable stocks (FN). The main goal of stock market investment is to achieve the maximum possible return under given risk conditions. For investors, mistakenly buying losing stocks directly reduces the principal, while missing out on profitable stocks only incurs a limited opportunity cost. That is, predicting future losing stocks as profitable stocks results in much greater losses compared to predicting profitable stocks as losing stocks. Therefore, this paper takes the Type_error_II obtained from predicting the maximum possible returns of T1-T7 as the benchmark, uses grid search for parameter tuning to find the grid parameters that minimize Type_error_II, and if the same Type_error_II occurs, selects the grid parameters with the highest AUC_ROC.

(2) Meaning and Selection of Grid Parameters

The names, meanings, and values of the random forest parameters are shown in Table 4. These parameters are used to predict the returns of samples T1-T7 through 720 parameter grids formed by the Cartesian product.

Table 4. Parameter Grid

Name	Meaning	Value
n_estimators	Number of decision trees	100, 300, 500, 700, 1000

max_depth	Maximum depth of a single decision tree	None, 10, 20, 30
min_samples_split	Minimum number of samples contained in a leaf node	2, 5, 10, 20
min_samples_leaf	Minimum number of samples required for a decision tree split	1, 5, 10
max_features	The number of indicators randomly selected by the decision tree	Sqrt, log2, 0.5

This study selected the grid parameters with the highest AUC_ROC based on the minimum Type_error II, obtaining the optimal parameter grid as shown in Table 5:

Table 5. Optimal Parameter Grid for T1-T7

Date	FP	Accuracy	n_estimators	max_depth	min_samples_split	min_samples_leaf
T+1	594	0.694212	1000	None	20	10
T+2	1278	0.685857	100	None	2	1
T+3	1380	0.675596	100	None	2	1
T+4	1344	0.676808	100	30	2	1
T+5	1249	0.680696	100	None	2	1
T+6	1177	0.680512	100	30	2	1
T+7	1100	0.681072	100	30	2	1

Subsequently, optimal parameter grids for T1-T7 were used to build random forests to further explore the influencing mechanisms.

(3) Gini Coefficient Calculation

The Gini coefficient is an important metric for determining the ability of a non-leaf node to predict the future rise or fall of stocks. It assesses the importance of a feature to the overall decision tree based on the purity after classification by that node's feature. The formula is:

$$Gini = 1 - \left(\frac{N_1}{N}\right)^2 - \left(1 - \frac{N_1}{N}\right)^2 \tag{1}$$

In equation (1), represents the proportion of stocks whose prices have increased out of the total number of stocks.

For the first indicator of a decision tree with N samples, take the value of the first sample as the threshold for splitting. If the subsequent sample values are less than this threshold, they are split into the left child node; otherwise, they are split into the right child node. The Gini coefficients of the two child nodes are then calculated based on the number of stocks whose prices go up or down, and subsequently the coefficient of this node is calculated. In this way, the Gini coefficient for N samples can be calculated, and the Gini coefficient forx_1 is:

$$Gini_m = min\left\{Gini_{m1}, Gini_{m2}, \cdots, \ Gini_{mn}\right\} \tag{2}$$

Similarly, the Gini coefficients for other features can be obtained. Since a smaller Gini coefficient indicates a higher purity of the indicator in selecting rising or falling stocks, this paper selects the indicator with the smallest Gini coefficient among the 52 features as the first splitting node. Stocks with values below this threshold are placed in the left child node, while those above are placed in the right child node. In this way, a complete decision tree is constructed. The random forest determines the final judgment of stock price movement through voting. If most decision trees predict that a stock will "rise" on the next trading day, the random forest will also predict that the stock will "rise".

(4) Judgment of the Optimal Indicator Combination from T1 to T7

Due to the collinearity among the 52 features affecting model performance, this study designed a method to optimize model performance as much as possible, starting with the calculation of feature importance. Random Forest normalizes the sum of feature importances evaluated by individual decision trees:

$$FI_m = \frac{\sum_{i=1}^{op-tree} FI_{mi}}{\sum_{i=1}^{Ind} FI_{mi}} \tag{3}$$

In Equation (3), op-tree represents the optimal decision tree obtained through grid search; Ind indicates the number of times a certain feature is sampled when generating decision trees, and FI_{mi}represents the feature importance score of that feature in the random forest. Additionally, this study ranks the features from highest to lowest importance, designates the most important feature as Set 1; includes both the most important and the second most important features as Set 2, and continues in this manner to obtain feature values for 52 indicators. Then, the AUC_ROC values for each feature set are compared to determine the optimal feature set.

2.4 Feature Interpretability Analysis Based on the SHAP Method

To further investigate the impact of changes in key feature values on future returns, this study introduces the SHAP method. For example, if the stock "600023" does not have a volumn_ratio on 2024/12/14, the new return can be predicted using other feature data. The marginal contribution of that volumn_ratio to the predicted return of stock "600023" on 2024/12/14 is then calculated as the difference between the original return and the new return. This method reflects the magnitude and direction of the impact of the same feature value on stock returns by illustrating the relationship between feature values and their marginal contribution rates.

$$\varnothing_i^t = \sum_{S \subseteq F\setminus\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}\left[f(S\cup\{i\}) - f(S)\right] \tag{4}$$

In equation (4), F represents the set of all features; S represents a subset that does not include feature i; f(S) represents the predicted values of returns from T1 to T7 using the subset S in the decision tree;$\frac{|S|!(|F|-|S|-1)!}{|F|!}$ represents the probability of S occurring in all possible arrangement.

3. Model Solving

3.1 Analysis of the Impact Characteristics of T1-T7 Yields

To improve the generalization ability of the model, this study uses 5-fold cross-validation to calculate the feature importance of for T1-T7. Specifically, the dataset is divided into 5 equal parts, and in each iteration, one subset is used as the validation set while the remaining 4 subsets are used to train the model. Finally, the top 11 features in terms of impact are obtained, as shown in Table 6:

Table 6. Feature Importance Ranking of T1-T7

fature	Profit_Index_2	Market_Return_2	Profit_Index_1	Market_Return_1	Body_Size_2	Return_2
importance	0.136839	0.123081	0.097344	0.06863	0.027526	0.026033
fature	Return_0	Drop	Return_1	Body_Size_1	Volume_Ratio	other_index
importance	0.037816	0.015189	0.01379	0.013673	0.009845	0.029215

(1) Profit_index_2 and Market_return_2 are the most important influencing factors because an increase in Market_return_2 reflects overall market optimism, which in turn drives individual stock prices up. Compared to Market_return_2, Profit_index_2 better explains the price rise from T1 to T7 because it results from the price increase of a majority of ordinary stocks, indicating more sustained future profits.
(2) The impact of Profit_index_2 and Market_return_2 is stronger than that of Profit_index_1 and Market_return_1 because institutional investors typically need 1-2 days to complete research, risk control, and trading. Continuous increases in Profit_index_2 and Market_return_2 at T2 are considered genuinely strong trading opportunities rather than random fluctuations, leading to continued accumulation over the next 5-7 trading days. A period of low volume followed by high volume significantly increases the likelihood of a price rise. Retail investors also often require two consecutive trading days of index gains to confirm market enthusiasm. Two days of continuous optimism strengthen retail expectations and attract substantial capital inflows.
(3) Body_size_2 has a greater impact on T1-T7 stock price fluctuations compared to Body_size_1 because an increase in Body_size_1 alone may represent the following scenario: major investors buy large amounts of stock in the few minutes before the 15:00 close, pushing up prices and forming a large bullish candlestick. If major investors hope to maintain that price the next day, they would need to absorb selling pressure throughout the trading session, significantly increasing costs compared to boosting prices at the close. As a result, the next day major investors may abandon maintaining high prices, causing a price drop and forming a small bearish candlestick. If prices continue to rise with high volume the next day, it likely indicates strong genuine market demand, greatly increasing the probability of T1-T7 stock price surges.
(4) Compared to Return_1, which reflects short-term profitability, Drop filters out short-term noise and illustrates the medium- to long-term price trend; on the other hand, emotional fluctuations caused by

the decline from medium- to long-term price peaks are far greater than those from short-term price movements. A price bottoming may attract substantial bottom-fishing funds, whereas a single-day profit or loss may be considered merely a short-term signal. This judgment results in Drop having a greater impact than Return_1.

3.2 Optimal Indicator Combination Assessment

Based on the observation that the overall impact of the 11 features gradually decreases over time, it can be inferred that the overall influence of financial indicators generally shows an upward trend over time. Since financial indicators mainly provide fundamental information, their effects are usually lagged and persistent. Therefore, this paper visualizes the changes in AUC_ROC with the number of indicator sets:

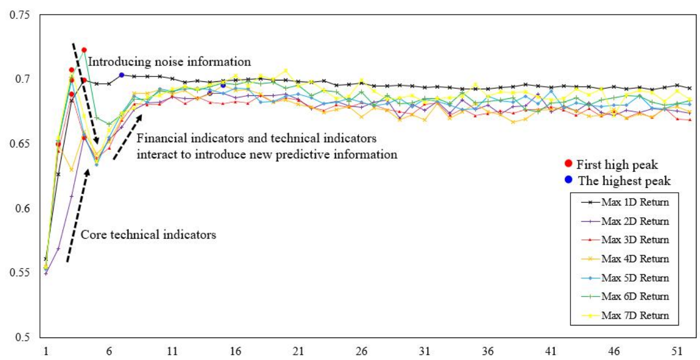

Figure 3. Optimal Indicator set of Random Forest for T1-T7

When the number of indicators is 8, T1 reaches the maximum AUC_ROC value; when the number of indicators is 13, T2 has the highest AUC_ROC; when the number of indicators is 15, T4 reaches the maximum AUC_ROC value. When the number of indicators is 3 or 4, the maximum possible AUC_ROC for other time periods is highest, then it drops sharply and slowly rises to another peak. This is because when the number of indicators is 3 or 4, the most important features are selected by the random forest, and these technical indicators have strong predictive value. By retaining these core features, the random forest can achieve high generalization ability; adding 1-2 more indicators introduces noise information, such as Inventory_turnover and MA5, and the addition of these features dilutes the importance of the original features, leading to decreased model classification performance.

As the number of indicators continues to increase, AUC_ROC gradually rises, due to the introduction of new predictive information from financial indicators. For example, Cash_Ratio and Dept_Ratio measure a company’s borrowing situation, and their combination can capture the company's risk resistance under credit tightening; meanwhile, the synergy between Inventory_turnover and Receivables_turnover reveals the company's operational capability in high-growth periods. The random forest, by aggregating local decisions from multiple decision trees, indirectly incorporates nonlinear relationships between features, leading to AUC_ROC reaching a second peak.

3.3 Visualization of Feature Marginal Contribution Based on SHAP

Draw a SHAP plot based on the indicator corresponding to the second peak of the T 7 stock return AUC_ROC, with the horizontal axis representing the indicator value and the vertical axis representing the SHAP value of the indicator.

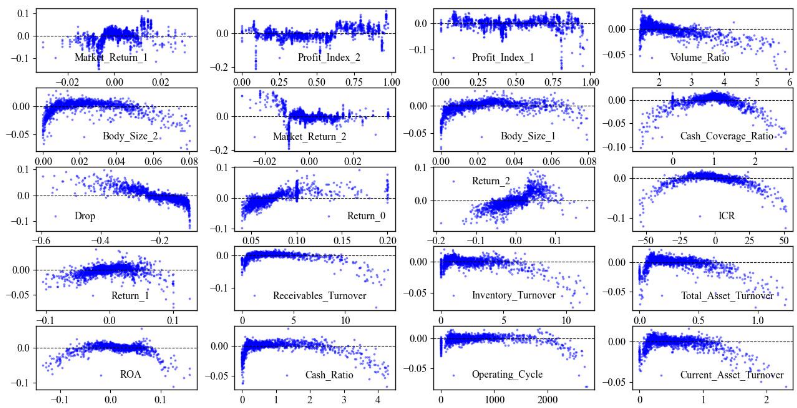

Figure 4. T7 Optimal Indicator Set SHAP Fluctuation Trend Chart

The sizes of Body_size_1 and Body_size_2 are concentrated between 0-0.04. There is a nonlinear relationship between Body_size and its SHAP value. When Body_size < 0.01, most SHAP values are negative, meaning that when the entity is small on that day, the market maker considers it not the best time to sell at a high price and therefore continues to hold the funds awaiting a better selling opportunity. As Body_size increases, two situations may occur: first, the market maker starts selling the stocks held, and if there is sufficient buying demand in the market, it could trigger positive market sentiment, driving the stock market upward; second, the stock price on that day is experiencing a significant increase.

Therefore, the likelihood of a substantial future price rise is much higher for stocks with a larger Body_size than for those with a smaller Body_size. When the Drop value is between -0.2 and -0.1, the impact is negative. During this time, investors exhibit a wait-and-see attitude under slight market declines, and trading volume further decreases as the market digests negative news. When Drop < -0.3, the impact turns positive. The stock price drops sharply, and once it falls below the threshold set by institutional investors, automatic sell orders are triggered. Market supply exceeds demand, causing the stock price to decline further, which attracts contrarian investors to buy the dip, thereby promoting a recovery in price and volume.

When Volumn_Ratio < 2 or > 3, SHAP values are mostly negative, whereas when Volumn_Ratio ∈ (2,3), SHAP values are positive. Volumn_Ratio < 2 reflects strong wait-and-see sentiment among investors, and under limited trading volume, price fluctuations lack directionality; price increases in this scenario are likely considered as a “low-volume rebound.” Volumn_Ratio > 3 usually indicates strong disagreement among investors about future stock price movements. Sellers predict price drops and sell large amounts, causing a significant increase in trading volume. Supply exceeds demand in the stock market, and stock prices face the risk of falling.

When predicting stock returns from T2 to T7, some financial indicators are gradually added into the random forest model. The curves of ICR, Cash_coverage_ratio, and their SHAP values exhibit an inverted U-shape. When ICR ∈ (-10, 10), reducing the interest coverage ratio has a positive impact on SHAP values. When >10, overly conservative financial policies may trigger agency costs, adversely affecting the company. When Cash_coverage_ratio ∈ (0, 1.5), its marginal contribution to T2-T7 is positive. Values >1.5 indicate redundant cash, reflecting management's inability to identify investment opportunities; excessively low values suggest potential financing risks for the company, indirectly leading to a decline in T2-T7 stock prices.

4. Conclusions and Implications

4.1 Conclusion

This study constructs a cross-cycle, multi-dimensional analytical framework through theoretical analysis and machine learning methods, revealing the mechanism by which a sudden surge in trading volume following a period of extremely low volume during a stock price decline can lead to future stock price increases.

Firstly, the optimal random forest feature ranking indicates that Market_Return and Profit_index are the core influencing factors of this phenomenon. Two consecutive days of index gains can significantly increase the probability of individual stock price rises. Body_size_2 has greater explanatory power compared to Body_size_1. This is because continued trading volume the next day indicates strong real market demand, significantly boosting the probability of individual stock price increases. When Drop exceeds a threshold, it triggers a market bottom-fishing signal, whose impact is greater than Return_1, which may be regarded as a short-term signal.

Secondly, most dates require only 3-4 technical indicators to reach the first peak value, after which the AUC_ROC first decreases and then rises, forming a second peak. This occurs because initial technical indicators already cover the key information; adding new features introduces noise and weakens model performance; but as more features are added, financial indicators bring new predictive dimensions, enhancing the model’s performance again.

Finally, by analyzing the optimal indicator set data and their SHAP values, we obtain the impact of indicator ranges on stock returns: Body_size < 0.01 indicates that most investors are in a wait-and-see state; as Body_size gradually increases, investors begin to follow the buying trend, or the stock price may already be in a substantial upward movement. When Drop < -0.3, on one hand, it can attract contrarian investors to buy the dip; on the other hand, it can trigger pre-set sell orders by institutional investors, thereby promoting a rebound in volume and price. Volumn_Ratio < 2 indicates stock fluctuations lack direction; > 3 indicates market investors have disagreements about the future, posing a risk of stock price decline. Among financial indicators, ICR and Cash_coverage_ratio show an inverted U-shaped relationship. When ICR is between -10 and 10, and Cash_coverage between 0 and 1.5, financial health is moderate and has a positive marginal effect on stock prices; being overly conservative or aggressive may lead to agency costs or financing risks.

4.2 Implication

Firstly, on a personal level, a drop in stock price of more than 30% can be considered an important signal to buy at the bottom. At this point, the probability of a rebound after the stock hits the bottom is much higher than when the drop is less than 30%. When there is a surge in trading volume on the first day, one should observe whether there is also a surge the following day. This can help rule out intentional market manipulation by major investors trying to push the stock price, allowing one to gauge genuine strong market demand. When the volume ratio exceeds 3, it indicates significant market disagreement, and the stock price faces a risk of decline, so one should not blindly chase after high prices. The probability of individual stocks rising significantly increases when both the overall market index and the market earnings index rise continuously.

Secondly, due to the inverted U-shaped effects of the interest coverage ratio and cash coverage ratio, companies should optimize financial and operational management. This means that lower debt is not always better; moderate leverage indicates that a company is actively using financing to grow rather than being overly conservative. Excessive cash reserves, on the other hand, indicate redundant capital and a lack of suitable investment opportunities. Therefore, maintaining moderate financial and operational indicators makes long-term, steady investors more optimistic about a company's high growth potential and high returns.

Finally, from the perspective of the entire economic market, this study introduces a market rise-and-fall index to genuinely reflect the overall market. When the market index is too high, the central bank can use the market rise-and-fall index to better assess the conditions of the capital market and the real economy, thus implementing appropriate monetary policies to regulate the market. Additionally, relevant financial regulatory authorities can monitor manipulative behaviors affecting the market index by constructing explainable models, thereby creating a more accurate and effective capital market and further enhancing the sustainability and inclusivity of the stock market's wealth effect. This provides a microeconomic foundation for the macroeconomic strategy of "expanding domestic demand" and promotes further prosperity of the real economy.

References

Bao, Z. S., Guo, J. N., Xie, Y. et al. (2020). Stock Price Rise and Fall Prediction Model Base d on LSTM-GA. *Computer Science*, *47*(S1), 467-473.

Chi, G. T., Yang, J. Q., & Zhou, Y. (2024). Research on Dynamic Early Warning of Bond De faults Based on Optimal Critical Points. *Journal of Management Science*, *27*(11), 136-158.

Li, Y., Gao, S. H., Gao, Y. T. et al. (2023). Market Index Prediction Based on Maximum Corr elation Entropy Criterion Regression Model. *System Science and Mathematics*, *43*(01), 186-198.

Sun, Y. L., Chen, S. D., & Liu, Y. (2019). Stock Price Volatility Prediction from the Perspective of Stock and Foreign Exchange Market Trading Volume Information. *Systems Engineering Theory and Practice*, *39*(04), 935-945.

Wang, L. L., & Huang, W. Q. (2025). Investor Sentiment Contagion and Systemic Risk Early Warning in the Banking Industry—A Study Based on Explainable Machine Learning Methods. *Systems Engineering*, 1-16.

Xu, X. C., & Tian, K. (2021). A New Method for Stock Index Prediction Based on Financial Text Sentiment Analysis. *Quantitative & Technical Economics Research*, *38*(12), 124-145.

Yan, Z. X., Qin, C., & Song, G. (2021). Stock Price Prediction Using a Random Forest Model Based on Pearson Feature Selection. *Computer Engineering and Applications*, *57*(15), 286-296.

Yang, N. N. (2020). Research on Stock Price Volatility of Agricultural Listed Companies—Empirical Analysis Based on the GARCH Model Family. *Price Theory and Practice*, (10), 96-99.

Zhang, C. X., Wu, H. Q., Cai, G. W. et al. (2021). Research on the Impact of Information Channels on Stock Price Volatility—From the Perspective of Media Reports and Analyst Forecasts.*Southern Economy*, (11), 122-136.

Zheng, Q. H., Tan, W. J., & Chen, T. T. (2018). Research on the Correlation between Financial Indicators and Stock Prices—An Analysis Based on "Internet +" Concept Stocks. *Value Engineering*, *37*(28),273-276.

Zou, J., & Li, L. (2023). Research on Stock Price Prediction Using the RF-SA-GRU Model. *Computer Engineering and Applications*, *59*(15), 300-309.