

A Study on the Incremental Value of Technical Indicator Features in Predicting the Next-Day Direction of the CSI 300 Index

Wei Zhang*

College of Economics and Management, Shanghai University of Electric Power, Shanghai 201306, China

**Corresponding author: Wei Zhang.*

Abstract

Whether technical indicators can provide incremental information for predicting the direction of stock price movements is a classic topic in financial market research. Using the CSI 300 Index as our research setting, we select daily data from 1,073 valid trading days from July 2020 to December 2024, construct a dataset containing 19 features such as historical returns, moving averages, RSI, MACD, Bollinger Bands, KDJ, and trading volume, and use Logistic regression and random forest models to rigorously evaluate the incremental value of technical indicators in prediction. We adopt a rigorous time-series split. In the main experiment, July 2020 to December 2023 is used as the training set (831 samples), the whole year of 2024 is used as the test set (242 samples), and five rolling windows are used to test the temporal robustness of the conclusions. We find that: (1) the best model, the random forest model using only historical return features, has a test-set accuracy of 56.20%. The Bootstrap test shows that this accuracy is higher than the random-guessing baseline at the 95% confidence level, but its difference from the naive baseline is not statistically significant (one-sided $p = 0.123$), indicating that the incremental predictive value of technical indicators is statistically fragile; (2) after technical indicators are added, the training-set accuracy of the random forest increases from 0.7052 to 0.7858, but the test-set accuracy decreases from 0.5620 to 0.5372, and the generalization gap expands from 14.32 percentage points to 24.86 percentage points. This paradox of feature increase, training improvement, and generalization deterioration is our core finding. The redundancy-removal experiment and the regularization control experiment further show that feature redundancy and limited samples jointly lead to the decline in generalization performance; (3) model prediction shows significant and stable asymmetry between up days and down days. The recall rate for down days (78.05%) is much higher than that for up days (33.61%), the predicted proportion of up days is only 27.69%, and this bearish bias persists in all five rolling windows. Market-state analysis shows that the weak market environment in the training period, with a cumulative return of -26.92%, is an important cause of the bearish bias.

Keywords

technical indicators, incremental value, CSI 300 Index, generalization performance, asymmetry between up and down days, rolling-window test

1. Introduction

Stock price prediction is a classic topic in finance research. The efficient market hypothesis states that, in a weak-form efficient market, historical price information has already been fully reflected in current prices, and technical analysis based on historical data cannot obtain excess returns [1]. However, the development of behavioral finance and the rise of machine learning have prompted scholars to re-examine this proposition: are there short-term price patterns in the market that can be mined? As mathematical derivatives of historical prices and trading volume, can technical indicators provide incremental information for future price trends? These questions bear not only on the theoretical judgment of market efficiency, but also directly on investors' strategy choices.

The CSI 300 Index consists of the 300 largest and most liquid stocks on the Shanghai and Shenzhen stock exchanges. It is a core benchmark index reflecting the overall trend of the A-share market. Predicting the short-term direction of this index helps to understand the level of information efficiency in the A-share market and also provides empirical reference for index investment and market-timing strategies. From 2020 to 2024, the A-share market experienced multiple changes, including the impact of COVID-19, economic recovery, and policy adjustment. The complex and changing market environment offers a relatively rich sample period for testing the predictive ability of technical indicators.

In recent years, domestic and foreign scholars have accumulated many research results in stock price prediction, but the following gaps persist in the literature. First, many studies use individual stocks as their research objects, which are greatly affected by firm-level information and make it difficult to purely test the predictive effectiveness of technical indicators. Second, some studies use random splits rather than time-series splits, creating a risk of future information leakage. Third, most studies report only model accuracy and lack statistical significance tests, making it impossible to determine whether accuracy above 50% comes from sampling fluctuation. Fourth, few studies use systematic ablation experiments and generalization performance analysis to discuss the internal mechanism of performance changes from the perspectives of feature correlation and regularization. Fifth, research on prediction asymmetry and its stability across different time windows remains insufficient.

To address these gaps, we use the CSI 300 Index as our research setting, adopts a rigorous time-series split framework, and comprehensively uses ablation experiments, redundancy-removal experiments, feature correlation analysis, regularization control experiments, Bootstrap significance tests, rolling-window out-of-sample tests, permutation importance analysis, and market-state analysis to systematically evaluate the incremental predictive value of technical indicators. We focus on three research questions. First, do technical indicators have statistically reliable incremental value for predicting the next-day direction of the CSI 300 Index? Second, why does the model's generalization performance decline rather than improve after technical indicators are added, and what role does feature redundancy play in this process? Third, does the model show asymmetry in predicting upward and downward directions, what causes it, and does it remain stable across market states?

2. Literature Review

2.1 The Debate Between Technical Analysis and the Efficient Market Hypothesis

The core assumption of technical analysis is that historical price trends contain exploitable patterns. Murphy systematically explains the theoretical basis of technical indicators in *Technical Analysis of the Financial Markets*, including core concepts such as trends, support and resistance, and volume confirmation [2]. In contrast, the efficient market hypothesis proposed by Fama argues that, in a weak-form efficient market, all historical price information has already been reflected in current prices, and technical analysis cannot obtain excess returns [1]. This theoretical debate forms the starting point of stock price prediction research.

In early strict tests of technical analysis, Sweeney systematically tested filter rules, proposed a testing method based on the X-statistic, and found that some filter rules could still obtain slight returns after transaction costs were considered, although the results were controversial in both statistical significance and economic meaning [3]. Brock et al. used the bootstrap method to test 26 technical trading rules and found that moving-average and channel-breakout strategies obtained statistically significant positive returns on the Dow Jones

Index. This finding challenged the EMH [4]. However, Sullivan et al. introduced a correction method for data-snooping bias and found, after retesting, that the conclusions of Brock et al. were no longer statistically significant [5]. Park and Irwin used White's reality check method to test technical trading rules in the U.S. futures markets and found that, after correction for data-snooping bias, most rules did not have stable profitability [6]. Bajgrowicz and Scaillet further used a false discovery rate method and found that a large proportion of significant results in technical trading rules were false discoveries [7]. Harvey et al. further pointed out that, under multiple testing in financial research, the traditional significance standard of $|t| > 2$ is no longer sufficient to exclude the possibility of data mining, and many published factor studies may fail out of sample [8]. These studies jointly show that the predictive ability of technical analysis requires stricter out-of-sample tests and statistical significance judgments.

2.2 Application of Machine Learning in Stock Price Prediction

With the improvement of computing power, machine learning methods have been widely applied to stock price prediction. Atsalakis and Valavanis reviewed related papers on stock market forecasting and found that most studies used artificial neural networks and that prediction accuracy varied greatly. However, the studies differed greatly in data selection, evaluation methods, and benchmark settings, making direct comparison difficult [9]. Kara et al. compared artificial neural networks and support vector machines in predicting the direction of the Istanbul stock index and found that both methods could obtain a certain predictive ability, but model performance was significantly affected by feature selection and market environment [10]. Kumar and Thenmozhi further compared support vector machines and random forests in stock index direction prediction and found that random forests outperformed SVMs on some indicators [11]. Patel et al. studied the effect of data preprocessing methods on machine learning prediction and emphasized that trend deterministic data preparation could improve directional predictive ability [12].

Ballings et al. compared random forests, AdaBoost, support vector machines, neural networks, Logistic regression, K-nearest neighbors, and other models in predicting the direction of European stock prices. They found that random forests performed best in terms of AUC (0.55-0.62), but prediction accuracy varied greatly across industries [13]. The authors particularly pointed out that too many features can easily lead to a decline in the generalization performance of tree models. Timmermann described stock return predictability as elusive in a review and pointed out that predictive relationships that are significant in sample often disappear out of sample, and that predictability changes significantly over time and across market states [14]. Gu et al. used U.S. stock market data from 1957 to 2016 to compare multiple machine learning models and hundreds of features. They found that tree models outperformed linear models in predicting stock returns, but the predictive gains were mainly reflected in ranking ability rather than absolute accuracy, and model performance showed significant fluctuations over time [15].

2.3 Related Research on China's A-Share Market

Domestic scholars have also conducted many studies on the efficiency of China's stock market and the predictive ability of technical analysis. Yang et al. conducted an early empirical analysis of weak-form efficiency in the Shanghai stock market and found that the Shanghai stock market in the 1990s had not yet reached weak-form efficiency [16]. Chen and Hong used a new testing method to study weak-form efficiency in China's stock market. Their results also did not support the weak-form efficiency hypothesis, but they pointed out that market efficiency was gradually improving [17]. Zheng studied the impact of XBRL financial reporting on the efficiency of China's stock market from the perspective of information disclosure quality and found that improved information transparency helps improve market efficiency [18]. Wang and Sun examined the returns of technical analysis rules, such as dual moving-average crossover, trading-range breakout, and relative strength index, in the Shanghai and Shenzhen securities markets, and discussed the relationship between the economic returns of technical analysis and market efficiency [19].

In the application of machine learning methods, Dong et al. combined social media text information with technical indicators and used support vector regression for stock index prediction. They found that predictive performance improved after multi-source information was integrated, but the predictive ability of technical indicators used alone was limited [20]. Duan et al. analyzed the information content of Chinese stock online forums and found that investor sentiment in online forums could have a certain effect on future stock prices [21]. Wang introduced the MACD indicator into a stock price predictive model for optimization and discussed

the auxiliary role of a single technical indicator in prediction [22]. Lin et al. used multiple machine learning methods combined with candlestick charts and technical indicators to study the predictive ability of China's stock market and compared differences in model performance [23]. Zhou et al. took the price spread of CSI 300 stock index futures as the research object and used SVM, random forest, and other models for intertemporal arbitrage prediction [24].

In behavioral finance and investor sentiment, Wen et al. studied the influence of investor sentiment characteristics on stock price behavior and found that sentiment fluctuation can significantly affect return distribution and the price formation mechanism [25]. He et al. further pointed out that the joint effect of investor sentiment and limits to arbitrage is an important cause of stock price anomalies, and that the degree of sentiment influence differs across market states [26]. Overall, domestic research has made some progress in using technical indicators to predict the A-share market, but most studies lack systematic ablation experiments on the incremental value of technical indicators, rarely conduct statistical significance tests, and provide insufficient discussion of generalization performance and prediction asymmetry.

2.4 Research Gaps

Existing studies have not fully solved the following problems: (1) systematic ablation experiments for the CSI 300 Index are lacking, making it impossible to accurately judge the reasons for changes in generalization performance; (2) statistical significance tests of prediction accuracy are lacking, and accuracy higher than 50% is not equal to real predictive power; (3) multi-window out-of-sample tests are lacking, and the result from a single test window may be affected by specific market conditions; (4) in-depth analysis of prediction asymmetry and tests of its temporal stability are lacking; and (5) feature importance analysis mostly uses built-in model methods and lacks more robust testing methods such as permutation importance. We design experiments to address these gaps and seeks to provide a more comprehensive and rigorous evaluation.

3. Research Methodology

3.1 Research Object and Data

We select the CSI 300 Index (code: 000300) as the research object. The reasons for choosing an index rather than individual stocks are as follows: index data are not affected by firm-level information of individual companies and can provide a cleaner test of the predictive power of technical indicators; the CSI 300 Index covers about 60% of the market capitalization of the A-share market and has strong representativeness; and index data are easier to obtain and verify, which is conducive to research replication.

The data source is the public API of Sina Finance, from which daily K-line data for the CSI 300 Index are obtained, including six fields: date, opening price, highest price, lowest price, closing price, and trading volume. The sample period is from July 3, 2020 to December 31, 2024, with a total of 1,093 trading days. This period is selected because it covers multiple market environments, including the impact of the pandemic, economic recovery, and policy adjustment, allowing the predictive ability of technical indicators to be tested under relatively rich market states. At the same time, the time span of about four and a half years ensures a sufficient sample size while avoiding comparability problems caused by long-term changes in market structure. After technical indicators are calculated, missing values from the initial period are removed, and 1,073 valid samples are obtained.

3.2 Variable Definitions

Dependent variable: the next-day direction of the CSI 300 Index (target), defined as $\text{target}_t = 1$ if $\text{close}_{(t+1)} > \text{close}_t$, and 0 otherwise. Direction prediction is used rather than return regression because direction prediction avoids the sharp peak and fat tail distribution problem of return series and directly corresponds to actual investment decisions.

Independent variables (19 features in total):

Time split: the training set is from July 31, 2020 to December 29, 2023 (831 samples), and the test set is from January 2, 2024 to December 31, 2024 (242 samples). The proportion of up days is 48.13% in the training set and 49.17% in the test set, showing similar distributions. In addition, five rolling windows are designed, one per quarter, with the previous 12 months used for training and the following three months used for testing,

to examine the temporal robustness of the conclusions. The last three months of the training set (60 samples) are separated as a validation set for selecting the L1 regularization parameter.

- Historical return features (4): cumulative returns over the previous 1, 2, 5, and 10 days.
- Moving average features (3): deviations of the closing price from the 5-day and 20-day moving averages, and the crossover signal between the 5-day and 20-day moving averages.
- RSI indicator (1): 14-day relative strength index.
- MACD indicators (3): DIF line, DEA line, and MACD histogram.
- Bollinger Band indicator (1): the relative position of the price within the Bollinger Band channel.
- KDJ indicators (3): K value, D value, and J value.
- Trading volume features (2): volume ratio and volume change rate.
- Volatility features (2): standard deviations of 5-day and 20-day returns.

3.3 Model Selection

We select the following models for comparison:

- (1) Naive baseline: uses yesterday's direction to predict today's direction.
- (2) Logistic regression (L2 regularization): linear classification model with default L2 regularization, set as the baseline model.
- (3) Logistic regression (L1 regularization): L1 regularization has a feature selection function, and the regularization strength C is selected through the validation set.
- (4) Random forest (max_depth=5): ensemble learning method capable of capturing nonlinear relationships. n_estimators=100, random_state=42.
- (5) Random forest (max_depth=3): used as a regularization control, with limited tree depth and increased minimum samples per leaf node.

The above models are trained using two groups of inputs: historical return features only and all 19 features. L1 regularization and restricted RF use only the full-feature setting. All features are standardized by StandardScaler before training, with fit conducted on the training set.

3.4 Evaluation Metrics and Statistical Tests

Three complementary metrics are used to evaluate model performance: accuracy, F1 score, and AUC. AUC reflects the model's ranking ability for positive and negative samples, is not affected by thresholds or class proportions, and is a relatively robust evaluation metric in financial prediction [13].

To determine whether accuracy above 50% has statistical significance, the Bootstrap method is used to estimate the 95% confidence interval of each model's accuracy [27]. Specifically, the test set is sampled with replacement 2,000 times, accuracy is calculated after each sampling, and the 2.5% and 97.5% quantiles are taken as the lower and upper bounds of the confidence interval. If the lower bound of the confidence interval is greater than 50%, the model accuracy is considered higher than random guessing at the 95% confidence level. In addition, paired Bootstrap tests are conducted between the best model and the naive baseline, and between the historical-feature-only and full-feature models, to judge the statistical significance of performance differences.

3.5 FEATURE Importance and Correlation Analysis

We use two methods to analyze and compare feature importance: (1) the random forest built-in importance, namely MDI, which is based on impurity reduction and is easy to calculate but is prone to bias when features are highly correlated; and (2) permutation importance, which measures importance by randomly shuffling the value of a feature and observing the decline in model accuracy. It is calculated on the test set and reflects the actual contribution of the feature to generalization performance. Importance with a mean greater than two times the standard deviation is used as the criterion for significance.

Feature correlation analysis calculates the Pearson correlation coefficient matrix of all 19 features on the training set and identifies highly correlated feature pairs ($|r| > 0.7$) to evaluate the degree of feature redundancy.

4. Experimental Results

4.1 Descriptive Statistics

After data cleaning, 1,073 valid samples are retained, including 831 samples in the training set (July 31, 2020 to December 29, 2023) and 242 samples in the test set (January 2, 2024 to December 31, 2024). There are 400 up days (48.13%) and 431 down days (51.87%) in the training set, and 119 up days (49.17%) and 123 down days (50.83%) in the test set. The up-down distributions of the two sets are similar, and the classes are relatively balanced.

4.2 FEATURE Correlation Analysis

The collinearity problem among features is quite serious. Among the 19 features, 43 feature pairs have absolute correlation coefficients greater than 0.7. The correlations within the same type of indicator are especially high: the average correlation coefficient among the three KDJ indicators is 0.818, and the correlation coefficient between MACD DIF and DEA is as high as 0.953. Cross-category correlations are also common. The correlation coefficients between RSI, KDJ, Bollinger Band position, and 10-day return are all above 0.7. This high feature redundancy means that many technical indicators, although different in name, essentially capture similar price patterns.

Table 1: Highly correlated feature pairs ($|r| > 0.7$; selected examples)

Category	Feature pair	Correlation coefficient $ r $
Returns and moving averages	return_2d - ma5_diff	0.901
return_10d - ma20_diff	0.903	
return_10d - boll_pos	0.833	
Within moving-average features	ma20_diff - ma_cross	0.881
ma20_diff - boll_pos	0.921	
Within MACD	macd_dif - macd_dea	0.953
Within KDJ	kdj_k - kdj_d	0.916
kdj_k - kdj_j	0.896	
Cross-category	rsi_14 - kdj_d	0.841

4.3 Comparison of Model Prediction Performance

Several points can be observed from Table 2. First, the best model is the random forest model using only historical features (depth = 5), with an accuracy of 56.20% and an AUC of 0.5890. Second, technical indicators have a positive effect on Logistic regression (L2), as accuracy increases from 49.59% to 54.55% after technical indicators are added. However, for the random forest, accuracy decreases from 56.20% to 53.72% after technical indicators are added. This contrast indicates that different models have different sensitivities to the number of features. Third, when the tree depth of the random forest is reduced from 5 to 3, the accuracy of the full-feature model rebounds from 53.72% to 54.96%, providing evidence that regularization can alleviate the decline in performance.

Table 2: Test-Set Prediction Performance of Each Model

Model	Number of features	Accuracy	F1	AUC
Naive baseline (yesterday's direction)	—	0.5041	0.4958	0.5040
Logistic regression (historical only, L2)	4	0.4959	0.3711	0.4940
Logistic regression (full features, L2)	19	0.5455	0.4444	0.4869
Logistic regression (full features, L1)	1*	0.4835	0.5136	0.4700
Random forest (historical only, depth = 5)	4	0.5620	0.4301	0.5890
Random forest (full features, depth = 5)	19	0.5372	0.4167	0.5214
Random forest (full features, depth = 3)	19	0.5496	0.3842	0.5336

Note: The L1-regularized Logistic regression selects the optimal $C = 0.05$ through the validation set, retaining only 1 non-zero feature (macd_hist), marked with an asterisk (*).

4.4 Generalization Performance Analysis

To judge why performance declines after technical indicators are added, Table 3 compares the accuracy of each model on the training set, validation set, and test set.

Table 3 reveals several important facts. First, the generalization gap of the random forest is much larger than that of Logistic regression, indicating that tree models are more prone to fitting noise in the training set. Second, for the random forest (depth = 5), after technical indicators are added, the training-set accuracy increases from 0.7052 to 0.7858 (+8.06 percentage points), but the test-set accuracy decreases from 0.5620 to 0.5372, and the generalization gap expands from 14.32 percentage points to 24.86 percentage points. Third, the regularization control experiment provides further evidence: when the tree depth of the random forest is limited from 5 to 3, the generalization gap of the full-feature model narrows substantially from 24.86 percentage points to 5.93 percentage points. Fourth, for Logistic regression (L2), the generalization gap actually narrows after technical indicators are added, indicating that for linear models, technical indicators do provide some incremental information, and due to limited model capacity, deterioration in generalization performance is less likely to occur.

Based on three pieces of evidence, namely (1) the gradient pattern in which the generalization gap expands monotonically as the number of features increases, (2) the significant narrowing of the generalization gap after the regularization strength is increased, and (3) the serious collinearity among features (43 pairs with $|r| > 0.7$), the evidence strongly suggests that feature redundancy and limited samples jointly lead to the decline in model generalization performance. We use the more cautious phrasing "decline in generalization performance" rather than directly characterizing it as overfitting.

Table 3: Accuracy Comparison of Each Model on the Training, Validation, and Test Sets

Model	Training Acc	Validation Acc	Test Acc	Generalization gap (train-test)
Logistic regression (historical only, L2)	0.5367	0.5833	0.4959	0.0408
Logistic regression (full features, L2)	0.5740	0.6167	0.5455	0.0286
Random forest (historical only, depth = 5)	0.7052	0.6833	0.5620	0.1432
Random forest (full features, depth = 5)	0.7858	0.8167	0.5372	0.2486
Random forest (full features, depth = 3)	0.6089	0.6500	0.5496	0.0593

4.5 Ablation Experiment

To examine the marginal contribution of each group of technical indicators in more detail, the random forest model (depth = 5) is used as the benchmark model, and technical indicator features are added group by group starting from historical returns only.

The ablation experiment shows that, as the number of features increases, training-set accuracy generally shows an upward trend, while test-set accuracy shows an overall fluctuating downward trend, and the generalization gap generally expands. The addition of Bollinger Bands is an exception. After it is added, test-set accuracy rebounds from 0.5372 to 0.5702, possibly because Bollinger Bands can describe the relative position of prices. However, the addition of KDJ causes accuracy to fall sharply to 0.5124, indicating that interactions among features are complex. Overall, for the random forest, technical indicator features have a negative overall effect: the more features are included, the better the training-set fitting becomes, but the worse the test-set generalization performance becomes. This is consistent with the finding of Ballings et al. [13].

Table 4: Ablation Experiment Results (Random Forest, max depth = 5)

Feature group	Number of features	Training Acc	Test Acc	Generalization gap	AUC
Historical returns only	4	0.7052	0.5620	0.1432	0.5890
+ Moving averages	7	0.7329	0.5661	0.1667	0.5600
+ RSI	8	0.7521	0.5413	0.2108	0.5563
+ MACD	11	0.7605	0.5372	0.2233	0.5387
+ Bollinger Bands	12	0.7521	0.5702	0.1819	0.5811
+ KDJ	15	0.7690	0.5124	0.2566	0.5255
+ Volume features	17	0.7906	0.5413	0.2493	0.5585
All features	19	0.7858	0.5372	0.2486	0.5214

4.6 Bootstrap Statistical Significance Test

Among the seven models, only the random forest using historical features only has a lower bound of the 95% confidence interval (0.5041) greater than 0.5, indicating that its accuracy is statistically higher than random guessing. The confidence intervals of the other models all include 0.5, so the null hypothesis that accuracy is equal to 50% cannot be rejected at the 95% confidence level. The paired Bootstrap test shows that the mean accuracy difference between RF-historical-only and the naive baseline is 5.71 percentage points, with a 95% confidence interval of $[-0.0372, 0.1529]$ and a one-sided p-value of 0.123, which does not reach the 5% significance level. Similarly, the difference between the historical-feature-only and full-feature random forests is also not statistically significant (one-sided $p = 0.244$). This finding reminds us that even if model accuracy is numerically higher than 50%, it may still come from sampling fluctuation in the test sample. Harvey et al. pointed out that, because multiple comparison problems exist in financial research, the traditional significance standard of $|t| > 2$ is no longer sufficient to exclude the possibility of data mining [8].

Table 5: Bootstrap 95% Confidence Intervals for Model Accuracy

Model	Mean	95% lower bound	95% upper bound	Lower bound > 50%
Naive baseline	0.5046	0.4380	0.5661	No
Logistic regression (historical only, L2)	0.4948	0.4339	0.5579	No
Logistic regression (full features, L2)	0.5461	0.4835	0.6074	No
Random forest (historical only, depth = 5)	0.5617	0.5041	0.6240	Yes
Random forest (full features, depth = 5)	0.5381	0.4793	0.6033	No

4.7 Up-Down Asymmetry Analysis

To analyze in depth the difference in model performance on up days and down days, Table 6 reports detailed metrics for each model in the two classes of up and down.

Table 6 clearly shows asymmetry between up and down days. Except for the L1 regularized model, the recall rates for down days of all machine learning models (68%-80%) are much higher than their recall rates for up days (30%-37%). The naive baseline performs almost symmetrically for up and down days, indicating that this asymmetry is not caused by class imbalance in the data itself, but by systematic model bias. Taking the best model, the random forest using historical features only, as an example, its confusion matrix shows that the model makes 175 down predictions and 67 up predictions, so the number of down predictions is 2.6 times the number of up predictions. Among the true 119 up days, the model correctly predicts only 40 (recall rate 33.61%), while the other 79 are incorrectly predicted as down days. The model's bearish bias is very significant.

Table 6: Detailed Metrics for Upward and Downward Direction by Model

Model	Up Prec	Up Rec	Up F1	Down Prec	Down Rec	Down F1
Naive baseline	0.4958	0.4958	0.4958	0.5122	0.5122	0.5122
Logistic-historical only (L2)	0.4800	0.3025	0.3711	0.5030	0.6829	0.5793
Logistic-full features (L2)	0.5570	0.3697	0.4444	0.5399	0.7154	0.6154
RF-historical-only (depth = 5)	0.5970	0.3361	0.4301	0.5486	0.7805	0.6443
RF-full features (depth = 5)	0.5479	0.3361	0.4167	0.5325	0.7317	0.6164

4.8 FEATURE Importance Analysis

Using the `feature_importances_` attribute of the random forest (the MDI method), `return_1d` ranks first, followed by `kdj_j` and `rsi_14`. However, the MDI method has two known limitations. First, it measures the contribution to training-set fitting rather than the contribution to generalization performance. Second, when features are strongly correlated, MDI results are disturbed.

To compensate for the shortcomings of the MDI method, we calculate permutation importance on the test set. The most notable finding is that, among the 19 features, none has permutation importance statistically greater than 0, meaning that the mean importance of every feature is less than two times its standard deviation. This means that randomly shuffling the value of any single feature does not cause a statistically significant decline in model accuracy. This result forms a sharp contrast with the picture presented by MDI, in which every feature appears to make a certain contribution. Combined with feature correlation analysis, the following explanation can be given: MDI reflects the role of features in training-set fitting. Because there are many spurious correlations between technical indicators and price trends, and because the features are highly

redundant, the model can use these correlations in the training set to improve fitting. However, permutation importance reflects the real contribution of features to generalization performance. On the test set, these spurious correlations no longer exist, and because features can substitute for one another, separately shuffling any single feature does not significantly affect model performance.

4.9 Rolling-Window Out-of-Sample Test

The results from a single test window may be affected by specific market conditions. To test the robustness of our findings, we design five rolling windows.

The performance of each model fluctuates greatly across different windows. Taking RF-historical-only as an example, accuracy ranges from 49.18% (Window 5) to 68.75% (Window 4), a difference of nearly 20 percentage points. The conclusion that RF-historical-only is the best holds in only three of the five windows, indicating that the ranking of model performance is unstable and depends on the choice of test window. From the summary statistics, RF-historical-only has the highest mean (58.18%) but also a relatively large standard deviation (7.42%), forming a mean-variance tradeoff: the better the model performs, the more unstable its results are.

The bearish bias exists in all five windows. The recall rate for down days is always higher than that for up days, with differences ranging from 18 percentage points to 70 percentage points. This indicates that asymmetry between up and down days is not an accidental phenomenon in a particular test window, but a systematic feature shown by the model under different training data.

Table 7: Rolling-Window Out-of-Sample Test Results (Accuracy)

Window	Naive baseline	LR-full features	RF-historical-only	RF-full features
Window 1: 2023Q4	50.00%	50.00%	51.67%	56.67%
Window 2: 2024Q1	53.45%	44.83%	56.90%	41.38%
Window 3: 2024Q2	49.15%	64.41%	64.41%	61.02%
Window 4: 2024Q3	59.38%	56.25%	68.75%	62.50%
Window 5: 2024Q4	39.34%	57.38%	49.18%	45.90%

4.10 Redundancy-Removal FEATURE Experiment

Feature correlation analysis shows that there are 43 highly correlated feature pairs among the 19 features. If redundant features are removed and only representative indicators from each group are retained, can generalization performance improve? We remove highly correlated features within the same class from the 19 features and retains nine representative features. It should be noted that, even after within-class redundancy removal, there are still 10 cross-category highly correlated feature pairs ($|r| > 0.7$) among the nine features. This shows that information redundancy among technical indicators not only exists within the same type of indicator, but is also rooted in different mathematical expressions of the same price information by different indicators.

Table 8: Redundancy-Removal Feature Experiment Results (Random Forest, max_depth = 5)

Model	Number of features	Training Acc	Test Acc	Generalization gap	AUC
RF-historical-only	4	0.7052	0.5620	0.1432	0.5890
RF-redundancy-removal	9	0.7617	0.5496	0.2121	0.5502
RF-full features	19	0.7858	0.5372	0.2486	0.5214

Table 8 reveals a key gradient effect: from 4 to 9 to 19 features, the training-set accuracy rises monotonically (0.7052 to 0.7617 to 0.7858), but the test-set accuracy falls monotonically (0.5620 to 0.5496 to 0.5372), and the generalization gap expands monotonically (0.1432 to 0.2121 to 0.2486). This indicates that even after within-class redundant features are removed, cross-category collinearity still leads to a decline in generalization performance. This result further confirms that the paradox of feature increase leading to generalization deterioration does not stem from individual redundant features, but rather from the fact that technical indicators as a whole fail to provide incremental information.

4.11 Market-State Analysis

The previous sections found that the model has a stable bearish bias. To explore its causes, we analyze two aspects: first, the difference in market trends between the training period and the test period; and second, the model's performance under different market states.

The market as a whole is in a downward trend during the training period. The CSI 300 Index falls from 4,695.05 points at the end of July 2020 to 3,431.11 points at the end of 2023, with a cumulative return of -26.92%. The proportion of up days falls from 53.85% in 2020 to 44.21% in 2023, and the proportion of time that the market is above MA20 also falls from 69.23% in 2020 to 35.95% in 2023, with the lowest value of 32.64% in 2022. In contrast, the market cumulative return in the test period (2024) is +16.20%, and the market environment changes from weakness to rebound. This reversal of market trends between the training and test periods provides an intuitive explanation for the model's bearish bias: the model learns a decision boundary in the weak market in which predicting down is safer, and this strategy leads to serious missed predictions of up days during the rebound.

Using the 20-day moving average as the standard, the test period is divided into bull states and bear states. The model's accuracy in the bear state (58.87%) is higher than that in the bull state (52.48%). However, the key issue is that, in the bull state, the model's up-day recall rate is only 26.79%, and the predicted-up proportion is only 21.78%, far lower than the true proportion of up days in the bull state. This means that the model becomes even more conservative when the market is in a bull state. This is consistent with the logic of Fischer and Krauss's finding: patterns learned by a model in one period are difficult to transfer directly to another period [28].

The Markov regime-switching model proposed by Hamilton shows that switching among different states in economic time series is endogenous and nonstationary [29]. The study by Ang and Bekaert further confirms that regime-switching characteristics are common in international stock markets, and that volatility increases significantly and correlations strengthen in bear-market states [30]. van Norden and Schaller find significant regime-switching characteristics in stock returns, and the predictive effect of the price-dividend ratio is significantly asymmetric across different states [31]. Bekaert and Wu point out from the perspective of volatility asymmetry that the risk-return relationship in stock markets has systematic differences across different states [32]. Our findings align with the logic of these studies: the bearish strategy learned by the model in a single weak training period is essentially an adaptation to that specific regime, rather than a universal rule across states. Ang and Bekaert also point out that stock return predictability itself is state-dependent [33].

5. Discussion

5.1 Comparison Between the Results and the Literature

Our experimental results are compared with the main views in the literature review as follows:

(1) The view that technical indicators have weak predictive power is partially supported, but the significance is more limited. After technical indicators are added to Logistic regression, accuracy increases from 49.59% to 54.55% (+4.96 percentage points), roughly consistent with the 3-5 percentage point improvement reported in many studies. However, after technical indicators are added to the random forest, test-set accuracy decreases instead, and the Bootstrap test shows that, except for RF-historical-only, the accuracy of the other models is not significantly higher than the random level. The difference between RF-historical-only and the naive baseline is also not statistically significant (one-sided $p = 0.123$). Therefore, the predictive power of technical indicators is statistically fragile.

(2) The view that too many features damage generalization performance is supported, and multi-level evidence is provided. The ablation experiment combined with training-set/test-set comparison clearly shows that, for the random forest, as the number of features increases, training-set accuracy generally rises but test-set accuracy generally declines, and the generalization gap expands from 14.32 percentage points to 24.86 percentage points. This pattern is consistent with the finding of Ballings et al. [13]. We further provide explanatory evidence from three dimensions: feature correlation analysis finds 43 highly correlated feature pairs; the redundancy-removal experiment shows a gradient effect, indicating that cross-category collinearity is a deeper cause; and the regularization control experiment shows that the generalization gap of the restricted RF decreases substantially.

(3) The view that model performance fluctuates over time is supported. The study by Gu et al. points out that the performance of return predictive models based on machine learning shows significant fluctuations

over time [15]. Our rolling-window test confirms this finding. The window accuracy of RF-historical-only ranges from 49.18% to 68.75%, with a fluctuation of more than 19 percentage points. Timmermann summarizes this phenomenon as elusive return predictability [14]. Guo's study on earnings extrapolation bias further reveals the subtle relationship between investor behavioral bias and return predictability: investors' systematic bias in extrapolating future earnings from announced earnings leads to a special prediction pattern for returns at the beginning of a quarter, and this predictability is not a simple price pattern that technical indicators can capture [34].

(4) The view that random forests outperform linear models is partially supported. Under the historical-feature-only condition, random forest does outperform Logistic regression. However, under the full-feature condition, random forest is slightly inferior to Logistic regression. In the rolling windows, although RF-historical-only has the highest mean across the five windows, it also has the largest standard deviation, indicating that the advantage of random forests comes at the cost of higher instability.

5.2 Core Findings

Based on the above experimental results, we identify three core findings:

Finding 1 (core): the feature increase, training improvement, and generalization deterioration paradox of technical indicators. After technical indicators are added, the training-set accuracy of the random forest increases from 0.7052 to 0.7858, but test-set accuracy decreases from 0.5620 to 0.5372, and the generalization gap expands from 14.32 percentage points to 24.86 percentage points. The redundancy-removal experiment further reveals a gradient effect: from 4 to 9 to 19 features, training-set accuracy rises monotonically, test-set accuracy falls monotonically, and the generalization gap expands monotonically. The regularization control experiment shows that, after tree depth is limited to 3, the generalization gap decreases to 5.93 percentage points. Permutation importance analysis shows that no single feature has a statistically significant independent contribution to generalization performance. These results jointly point to one conclusion: although technical indicators improve model fitting ability on the training set, they do not provide incremental information under limited samples because there is serious information redundancy among indicators (43 highly correlated feature pairs), resulting in a decline rather than an improvement in generalization performance.

Finding 2: model prediction shows significant and stable asymmetry between up and down days, and the bearish bias originates from the weak market environment during the training period. Except for the L1 regularized model, the recall rates for down days of all machine learning models (68%-80%) are much higher than those for up days (30%-37%), and the predicted-up proportions (24%-33%) are much lower than the true up proportion (49.17%). Market-state analysis provides systematic evidence for the cause of the bearish bias: the weak market with a cumulative return of -26.92% during the training period causes the model to learn a conservative decision boundary. The bearish bias persists in all five rolling windows, although to different degrees, which is consistent with the expectation of regime-switching theory [29, 31].

Finding 3: predictive accuracy is highly unstable across different time intervals and depends on the market state. The rolling-window test shows that the accuracy of RF-historical-only fluctuates between 49.18% and 68.75%, with a standard deviation of 7.42%. The model's accuracy in the bear state (58.87%) is higher than that in the bull state (52.48%), further confirming that predictability is highly dependent on the market state. This shows that technical indicators reflect the continuation of historical patterns. When the market operating pattern changes, predictive methods based on historical data face fundamental challenges [1, 14].

5.3 Relationship with the Efficient Market Hypothesis

Our findings are broadly consistent with the expectations of the weak-form efficient market hypothesis. On the daily scale, predictive methods based on historical prices and technical indicators can obtain only very limited predictive power, and this predictive power is statistically fragile and unstable over time. Domestic studies on the efficiency of China's stock market also provide a reference: early studies did not support the judgment that China's stock market had reached weak-form efficiency, but they pointed out that market efficiency was gradually improving [16, 17]; improvement in information disclosure quality helps improve market efficiency [18]. These studies show that the efficiency level of the A-share market is a dynamically evolving process.

It is important to note, however, that the results of this study do not constitute a direct test of the efficient market hypothesis. The EMH is a general proposition about the overall information efficiency of the market, while we examine only the daily prediction problem of the CSI 300 Index over about four and a half years. Tests of the EMH usually need to combine transaction costs and excess return analysis, while we examine only prediction accuracy. Therefore, our conclusions should be understood as, within the sample and setting of our study, our findings are broadly consistent with the predictions of the weak-form efficient market hypothesis, rather than as proof of the weak-form efficient market hypothesis.

6. Conclusion

Using the CSI 300 Index as our research setting, this study uses daily data from July 2020 to December 2024, constructs a dataset containing 19 features, and systematically evaluates the incremental value of technical indicators for predicting the next-day direction through comparative experiments using Logistic regression and random forest models, ablation experiments, redundancy-removal experiments, feature correlation analysis, regularization control experiments, Bootstrap significance tests, rolling-window out-of-sample tests, permutation importance analysis, asymmetry between up and down days analysis, and market-state analysis. Our study does not claim that technical indicators can effectively predict the CSI 300 Index. In fact, the experimental results show that their incremental value is very limited. Rather, through a rigorous out-of-sample testing framework, we systematically document empirical evidence on the incremental value, feature redundancy, and prediction asymmetry of technical indicators, and provides a reference for understanding the boundary of the role of technical indicators in the A-share market.

The main conclusions of the study are as follows. First, the incremental value of technical indicators for predicting the next-day direction of the CSI 300 Index is limited, and its statistical significance is fragile. The best model, RF-historical-only, has an accuracy of 56.20%. Although it is higher than the random baseline at the 95% confidence level, its difference from the naive baseline is not statistically significant (one-sided $p = 0.123$). After technical indicators are added, the generalization gap of the random forest expands from 14.32 percentage points to 24.86 percentage points, and test-set accuracy decreases instead. The redundancy-removal experiment further reveals that, from 4 to 9 to 19 features, training-set accuracy rises monotonically while test-set accuracy falls monotonically. Even after within-class redundancy is removed, cross-category collinearity still leads to generalization deterioration. The regularization control experiment shows that limiting model capacity can reduce the generalization gap to 5.93 percentage points. Permutation importance analysis shows that no single feature has a statistically significant independent contribution to model generalization performance.

Second, model prediction shows significant and stable asymmetry between up and down days, and the bearish bias originates from the weak market environment during the training period. The recall rate for down days (78.05%) is much higher than that for up days (33.61%), and the predicted-up proportion is only 27.69% (the true proportion is 49.17%). Market-state analysis shows that the weak market with a cumulative return of -26.92% during the training period causes the model to learn a conservative decision boundary, and this bearish bias becomes even more severe in a bull market (the up-day recall rate is only 26.79%). It persists in all five rolling windows, although to different degrees.

Third, prediction accuracy fluctuates greatly across different time intervals and depends on the market state. The rolling-window test shows that the mean accuracy of RF-historical-only is 58.18%, but its standard deviation is as high as 7.42%, and the lowest value is only 49.18%. The model's lower accuracy in the bull state than in the bear state further shows that model performance is highly dependent on the market state.

The above conclusions have two implications for theoretical research and investment practice. In theory, we document empirical evidence that the incremental value of technical indicators is limited through a rigorous out-of-sample testing framework, and our findings are broadly consistent with the predictions of the weak-form efficient market hypothesis. However, it should be emphasized that this study does not constitute a direct test of the EMH. In practice, investors should not over-rely on technical indicators for short-term market timing. On the one hand, the predictive gains brought by technical indicators are limited and statistically fragile. On the other hand, the bearish bias learned by the model during a weak-market training period may lead to serious missed opportunities in rebound markets, and abrupt changes in market states may completely invalidate prediction based on historical patterns.

The limitations of this study are as follows: the sample period is relatively short and may not fully cover different market cycles; the weak market during the training period may cause systematic bias toward bearish predictions; macroeconomic and fundamental variables are not introduced, so the influence of external information cannot be controlled; model hyperparameters mainly use empirical values and are not selected through systematic grid search; only daily data are used, and higher-frequency information is not considered; and the parameters of technical indicators all use traditional default values and are not optimized for the A-share market. Future research can be expanded in the following directions: introducing macro factors and sentiment factors to construct multi-factor models; using deep learning methods to capture long-term dependencies in sequences; using rolling training and adaptive threshold methods to enable models to dynamically adapt to market changes; retraining models over longer sample periods covering complete bull and bear cycles; conducting backtesting based on actual trading strategies and combining transaction costs to transform prediction accuracy into investment return analysis; and testing more market indices and individual stocks to verify the generality of the conclusions.

References

- [1] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383-417.
- [2] Murphy, J. J. (1999). *Technical analysis of the financial markets*. New York Institute of Finance.
- [3] Sweeney, R. J. (1988). Some new filter rule tests: Methods and results. *Journal of Financial and Quantitative Analysis*, 23(3), 285-300.
- [4] Brock, W., Lakonishok, J., & LeBaron, B. (1992). Simple technical trading rules and the stochastic properties of stock returns. *The Journal of Finance*, 47(5), 1731-1764.
- [5] Sullivan, R., Timmermann, A., & White, H. (1999). Data-snooping, technical trading rule performance, and the bootstrap. *The Journal of Finance*, 54(5), 1647-1691.
- [6] Park, C.-H., & Irwin, S. H. (2010). A reality check on technical trading rule profits in the U.S. futures markets. *Journal of Futures Markets*, 30(7), 633-659.
- [7] Bajgrowicz, P., & Scaillet, O. (2012). Technical trading revisited: False discoveries, persistence tests, and transaction costs. *Journal of Financial Economics*, 106(3), 473-491.
- [8] Harvey, C. R., Liu, Y., & Zhu, H. (2016). ... and the Cross-Section of Expected Returns. *The Review of Financial Studies*, 29(1), 5-68.
- [9] Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying stock market forecasting techniques-Part II: Soft computing methods. *Expert Systems with Applications*, 36(3), 5932-5941.
- [10] Kara, Y., Boyacioglu, M. A., & Baykan, O. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert Systems with Applications*, 38(5), 5311-5319.
- [11] Kumar, M., & Thenmozhi, M. (2006, January). Forecasting stock index movement: A comparison of support vector machines and random forest. In *Indian Institute of Capital Markets 9th capital markets conference paper*.
- [12] Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, 42(1), 259-268.
- [13] Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications*, 42(20), 7046-7056.
- [14] Timmermann, A. (2008). Elusive return predictability. *International Journal of Forecasting*, 24(1), 1-18.
- [15] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223-2273.

- [16] Yang, C., Cai, M., & Hong, Y. (1998). Empirical analysis of weak-form efficiency in the Shanghai stock market. *Journal of Shanghai Jiao Tong University*, (3), 65-69.
- [17] Chen, D., & Hong, Y. (2003). Is China's stock market weak-form efficient? An empirical study based on a new method. *China Economic Quarterly*, 3(1), 97-124.
- [18] Zheng, J. (2015). The impact of XBRL financial reporting on the efficiency of China's stock market. *Journal of Financial Research*, (12), 194-206.
- [19] Wang, T., & Sun, Q. (2015). Economic returns of technical analysis and market efficiency. *International Business Research*, (5), 87-96.
- [20] Dong, L., Wang, Z., & Xiong, D. (2017). Stock index prediction based on textual information. *Journal of Peking University (Natural Sciences)*, 53(2), 273-278.
- [21] Duan, J., Liu, H., & Zeng, J. (2017). Information content analysis of Chinese stock online forums. *Journal of Financial Research*, (10), 178-192.
- [22] Wang, X. (2012). Stock Price Prediction Model. *Journal of Longdong University*, (3), 40-43.
- [23] Lin, Y., Liu, S., & Yang, H. (2022). A stock market investment strategy based on machine learning and candlestick charts. *China Journal of Econometrics*, 2(1), 126-140.
- [24] Zhou, L., Chen, C., & Li, N. (2022). Research on intertemporal arbitrage based on machine learning and empirical mode decomposition. *Journal of Southwest University (Natural Science Edition)*, 44(1), 148-159.
- [25] Wen, F., Xiao, J., Huang, C., Chen, X., & Yang, X. (2014). Research on the influence of investor sentiment characteristics on stock price behavior. *Journal of Management Sciences in China*, 17(3), 60-69.
- [26] He, C., Chen, R., Xue, B., & He, M. (2021). Investor sentiment, limits to arbitrage, and stock price anomalies. *Economic Research Journal*, 56(1), 58-73.
- [27] Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall.
- [28] Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654-669.
- [29] Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2), 357-384.
- [30] Ang, A., & Bekaert, G. (2002). International asset allocation with regime shifts. *The Review of Financial Studies*, 15(4), 1137-1187.
- [31] van Norden, S., & Schaller, H. (1997). Regime switching in stock market returns. *Applied Financial Economics*, 7(2), 177-191.
- [32] Bekaert, G., & Wu, G. (2000). Asymmetric volatility and risk in equity markets. *The Review of Financial Studies*, 13(1), 1-42.
- [33] Ang, A., & Bekaert, G. (2007). Stock return predictability: Is it there? *The Review of Financial Studies*, 20(3), 651-707.
- [34] Guo, H. (2025). Earnings extrapolation and predictable stock market returns. *The Review of Financial Studies*, 38(6), 1730-1782.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights

Copyright for this article is retained by the author (s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).