Article

Application of LightGBM in the Chinese Stock Market

Jie Yang *

Posted Date: 6 January 2025

doi: 10.20944/preprints202501.0303.v1

Keywords: LightGBM; Machine Learning; Feature Importance; stock market

Article

Application of LightGBM in the Chinese Stock Market

Jie Yang

School of Management and Economics, The Chinese University of Hong Kong, Shenzhen, 518172, China; jieyang1@link.cuhk.edu.cn

Abstract: This study employs LightGBM, a gradient boosting decision tree model, to predict stock returns and identify key pricing factors in the Chinese A-share market. The empirical analysis yields two main findings. First, LightGBM demonstrates superior predictive performance, achieving a monthly out-of-sample R^2 of 2.13%, more than doubling the 0.95% of traditional OLS regression. This advantage translates into significant economic gains: a LightGBM-based long-short strategy generates a monthly return of 2.63% with a Sharpe ratio of 1.77, substantially outperforming both the OLS-based strategy and the market benchmark. Second, through feature importance analysis, this study finds that liquidity and volatility-related characteristics are the most influential predictors of stock returns in China, consistent with recent literature suggesting the predominant role of market microstructure factors in emerging markets. These findings highlight the potential of machine learning techniques in asset pricing and provide new insights into return prediction in the Chinese stock market.

CCS Concepts: Applied computing; Operations research; Forecasting

Keywords: LightGBM; machine learning; feature importance; stock market

1. Introduction

In recent years, machine learning methods have achieved remarkable success across diverse domains. Among various machine learning approaches, Tree-based Ensemble Models have garnered significant attention from both academia and industry due to their distinctive advantages. These models not only effectively handle non-linear relationships but also perform automatic feature selection while maintaining robust performance in the presence of outliers. More significantly, compared to black-box models such as neural networks, tree models offer superior interpretability, providing them with a distinct advantage in applications requiring decision explanations.

Tree-based models have established themselves as fundamental tools for solving complex problems in finance. In their study of U.S. financial institution bankruptcy prediction, Petropoulos et al. (2020) demonstrated that random forest models exhibit superior performance in both out-of-sample and out-of-time predictions, while effectively identifying key predictive factors within the CAMELS evaluation framework. Through their comprehensive benchmark study of credit scoring algorithms, Lessmann et al. (2015) systematically evaluated contemporary classification algorithms including tree models, providing crucial insights into technical advancements in credit scoring practice. Research by Carmona et al. (2019) revealed that tree models perform exceptionally well in predicting U.S. bank failures, particularly during extreme market conditions such as financial crises.

Tree models have also found extensive applications in asset pricing, primarily addressing two crucial challenges: improving return prediction accuracy and identifying key pricing factors. Nti et al. (2019) applied random forest methodology to stock market prediction, effectively identifying key macroeconomic variables through feature selection, demonstrating superior accuracy compared to traditional methods. Gu et al. (2020) showed that tree models excel in both return prediction and feature contribution quantification. Lin et al. (2022), in their comprehensive survey, revealed tree models' significant advantages in handling non-linear relationships and high-dimensional data. For feature importance identification, Gu et al. (2020) and Leippold et al. (2021) assessed feature importance through variable removal impact, while Ma et al. (2023) employed SHAP methodology. However, these approaches generally overlook the dynamic characteristics of time series data.

Despite their achievements in asset pricing and feature identification, traditional tree models exhibit several limitations in processing large-scale financial data. These models often face challenges in computational efficiency and memory utilization when handling high-dimensional datasets with hundreds of thousands of records (Ferrouhi and Bouabdallaoui, 2024; Siswara et al., 2024). LightGBM (Ke et al., 2017) addresses these challenges through its histogram-based algorithm and leaf-wise growth strategy, enhancing both computational efficiency and memory utilization for large-scale asset pricing research.

Building on these advantages of LightGBM, this study aims to apply it to asset return prediction and feature importance identification. The methodology not only promises to enhance predictive accuracy but also provides new empirical insights into asset pricing theory. Through comparative analysis against traditional econometric methods and other machine learning algorithms, this paper comprehensively evaluates its applicability in asset pricing. Additionally, the feature importance analysis offered by LightGBM helps uncover key factors influencing asset returns, which is crucial for constructing more robust investment strategies and deepening our understanding of market dynamics.The remainder of this paper is organized as follows: Section 2 presents the research design. Section 3 describes the dataset used in this study. Section 4 presents the empirical results. Section 5 concludes the paper.

2. Data and Methodology

This study obtains daily stock return data for all A-shares listed on the Shanghai and Shenzhen Stock Exchanges from the Wind database and converts it to monthly frequency. After applying exclusion criteria—removing financial industry stocks, those with less than 12 months of listing history or fewer than 15 monthly trading days, ST-designated stocks, those with negative net assets—the final sample includes 4885 A-share stocks from the Main Board, ChiNext Board, and STAR Market traded from January 2000 to December 2023. Quarterly financial statement data is obtained from CSMAR, which also provides the one-year Chinese government bond yield as the risk-free rate proxy. Based on historical literature, I construct 50 stock characteristics, detailed in Table A1 of the Appendix A.

The analysis employs a generalized prediction error model to describe the relationship between stock excess returns and stock characteristics. The basic form of the model can be expressed as:

$$r_{i,t+1} = E_t(r_{i,t+1}) + \epsilon_{i,t+1}, \tag{1}$$

where $r_{i,t+1}$ represents the excess return of stock i at time $t+1$, $E_t(r_{i,t+1})$ is the conditional expectation based on information at time t, and $\epsilon_{i,t+1}$ is the prediction error term. Furthermore, the conditional expectation of stock excess returns is assumed to be a function of stock characteristics:

$$E_t(r_{i,t+1}) = f(\boldsymbol{X}_{i,t-\tau:t}; \theta), \tag{2}$$

where $X_{i,t-\tau:t} \in R^{(\tau+1)\times m}$ is a matrix containing stock characteristics of stock i from time $t-\tau$ to t. τ represents the historical data window length, m is the number of predictive features (50 in this paper), θ represents the model hyperparameters, and $f(\cdot)$ denotes the LightGBM function.

This study employs the LightGBM model to characterize the relationship between stock returns and firm characteristics. LightGBM is a gradient boosting framework based on decision trees that differs from traditional Gradient Boosting Decision Tree (GBDT) algorithms in its growth strategy. While GBDT uses a level-wise growth strategy, LightGBM implements a leaf-wise growth strategy. Under this strategy, the model iteratively selects the leaf node with maximum loss reduction for splitting, which can achieve lower loss compared to level-wise algorithms given the same number of splits. In LightGBM, the function $f(\cdot)$ is modeled as a weighted sum of T decision trees:

$$f(X_i) = \sum_{t=1}^{T} \beta_t h_t(X_i), \tag{3}$$

where $h_t(X_i)$ represents the t-th decision tree and β_t is its corresponding weight coefficient. Each decision tree is trained to minimize the following objective function:

$$\mathcal{L} = \sum_{i=1}^{n} l(r_i, f(X_i)) + \lambda \sum_{t=1}^{T} \Omega(h_t) , \tag{4}$$

where l denotes the loss function, $\Omega(h_t)$ is the regularization term that penalizes model complexity, and λ is the regularization coefficient. In contrast to traditional GBDT algorithms that employ level-wise growth strategy, LightGBM adopts a leaf-wise growth strategy, where the model iteratively selects the leaf node with maximum loss reduction for splitting, achieving lower loss compared to level-wise algorithms with the same number of splits.

For feature importance identification, LightGBM employs a split-gain-based approach to evaluate feature importance by calculating the total gain generated by each feature during the decision tree splitting process. The larger the gain produced by a feature during splitting, the more significant its contribution to improving the model's predictive capability. For feature k, its importance score can be expressed as:

$$gain_k = \sum_{splits\ on\ k} (l_{before} - l_{after}), \tag{5}$$

where l_{before} and l_{after} represent the loss function values before and after splitting using the feature, respectively. Theoretically, a larger split gain indicates that the split is more effective in reducing the model's prediction error, thereby indicating the feature's greater importance for prediction.

Finally, this study adopts the same sample splitting scheme as Leippold et al. (2021). The dataset is divided into three non-overlapping time periods: training set (2000-2010), validation set (2011-2013), and test set (2014-2023). When the model begins a new round of training, the training set incorporates the next year's data while retaining its original data; the validation set and the one-year test set move forward accordingly to include the next 12 months of data. Model training is conducted annually in January rather than monthly.

3. Results

This section presents the empirical results. The analysis begins with a comparison of LightGBM and traditional Ordinary Least Squares (OLS) regression in predicting excess returns in Chinese A-shares. This is followed by an examination of feature importance and its economic implications.

Table 1 presents a comparison of out-of-sample predictive performance between LightGBM and OLS regression, measured by the R-squared (R^2) statistic. The empirical results indicate that LightGBM achieves a monthly out-of-sample R^2 of 2.13%, representing a substantial improvement over the traditional OLS approach which yields an R^2 of 0.95%. This marked enhancement in predictive accuracy suggests that LightGBM's non-linear framework better captures the inherent complexities in stock return patterns that elude linear regression methods.

Table 1. Monthly out-of-sample predictive R^2 in percentage.

Algorithm	OLS	LightGBM
R^2_{oos}	0.95	2.13

Notes: This table reports monthly out-of-sample predictive R^2 of forecast models. All the numbers are expressed as a percentage.

Table 2 explores the economic significance of predictive performance through portfolio analysis. For the long-only strategy, LightGBM achieves a higher average monthly return of 2.54% compared to 1.83% for OLS, with an improved Sharpe ratio of 1.34 versus 1.01. The long-short strategy demonstrates even stronger performance under LightGBM, yielding a monthly return of 2.63% and a Sharpe ratio of 1.77, substantially outperforming the OLS approach which generates a return of 1.82% and a Sharpe ratio of 1.11. Notably, LightGBM also exhibits better downside protection, with less severe maximum drawdowns in both strategies (-22.55% versus -24.13% for long-only, and -16.60% versus -21.09% for long-short).

Table 2. Performance of portfolios.

	OLS	LightGBM
Long-Only		
Mean	1.83	2.54
SR	1.01	1.34
Max DD	-24.13	-22.55
Long-Short		
Mean	1.82	2.63
SR	1.11	1.77
Max DD	-21.09	-16.60

Notes: This table reports the performance of long-short and unilateral long investment portfolios. The stocks within each portfolio are weighted by market capitalization. "Mean" represents the average monthly return (%). "SR" represents the annualized Sharpe ratio. "Max DD" represents the maximum drawdown of the portfolio (%).

Figure 1 depicts the cumulative performance of different investment strategies relative to the CSI 300 index, which is a capitalization-weighted stock market index designed to track the performance of 300 most representative stocks listed on the Shanghai and Shenzhen Stock Exchanges. Panel A demonstrates that the LightGBM-based long-only strategy consistently outperforms both the OLS approach and the CSI 300 benchmark, with the outperformance becoming particularly pronounced after 2020. The superior performance of LightGBM is even more evident in Panel B's long-short strategy, where it achieves a cumulative log return of approximately 3.0 by 2024, substantially exceeding both the OLS strategy and the market benchmark.

Beyond predictive performance, understanding the relative importance of different features in explaining stock returns is another crucial aspect of asset pricing research. Based on the split-gain calculation specified in Equation (5), Figure 2 presents the feature importance scores derived from the LightGBM model. Consistent with Leippold et al. (2021), the results suggest that both liquidity and volatility-related characteristics are the most influential predictors in the Chinese A-share market. Specifically, liquidity measures, including abnormal turnover (abturn), trading volume (dtv), market capitalization (size), and trading volume variation (cvturn), rank among the top predictors. Volatility indicators such as maximum daily return (maxret) and historical price deviation (bias5) also demonstrate significant predictive power.

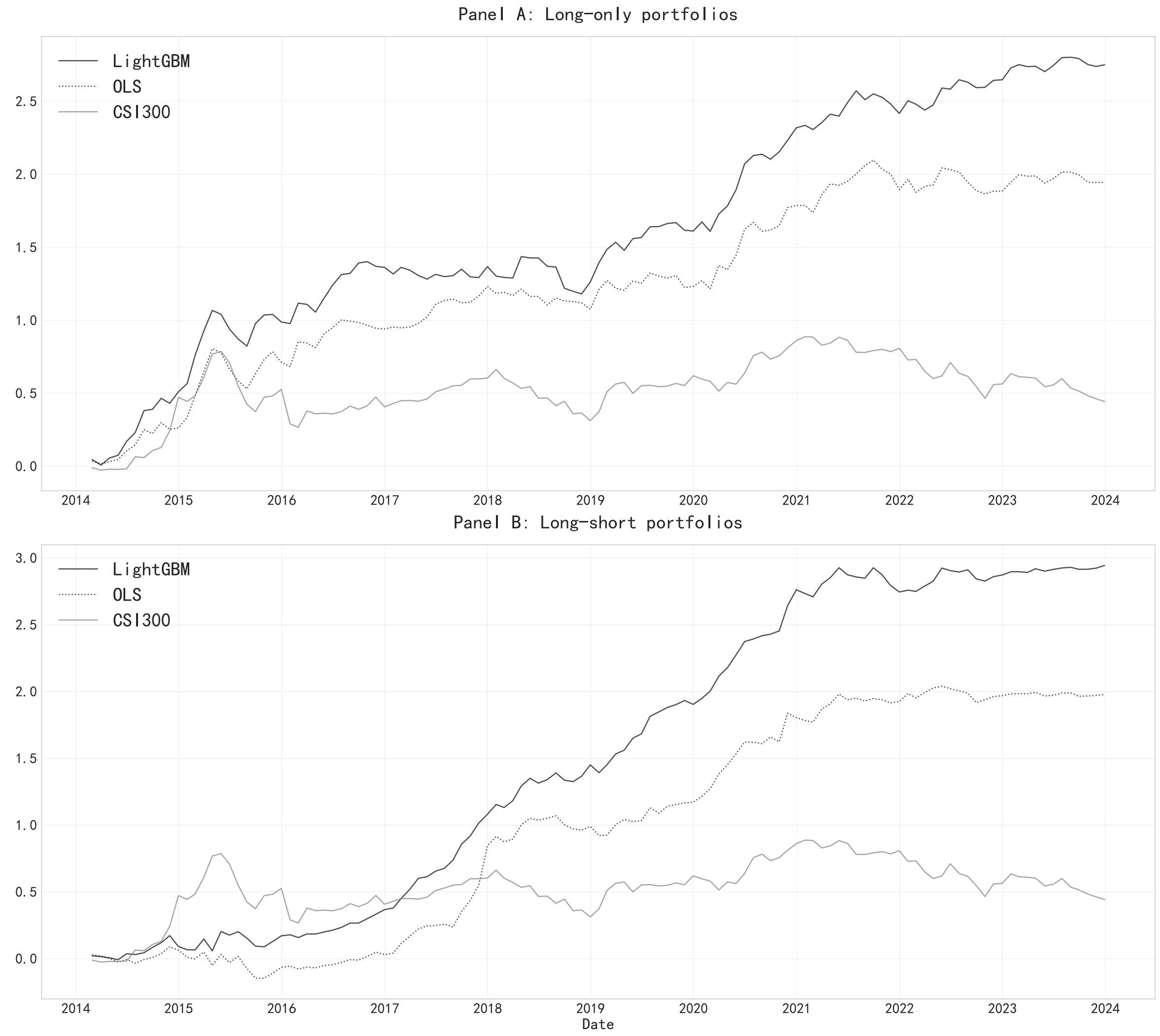

Figure 1. Cumulative log return of portfolios. Notes: This figure shows the cumulative log returns of all portfolios and the CSI 300 market index.

doi:10.20944/preprints202501.0303.v1

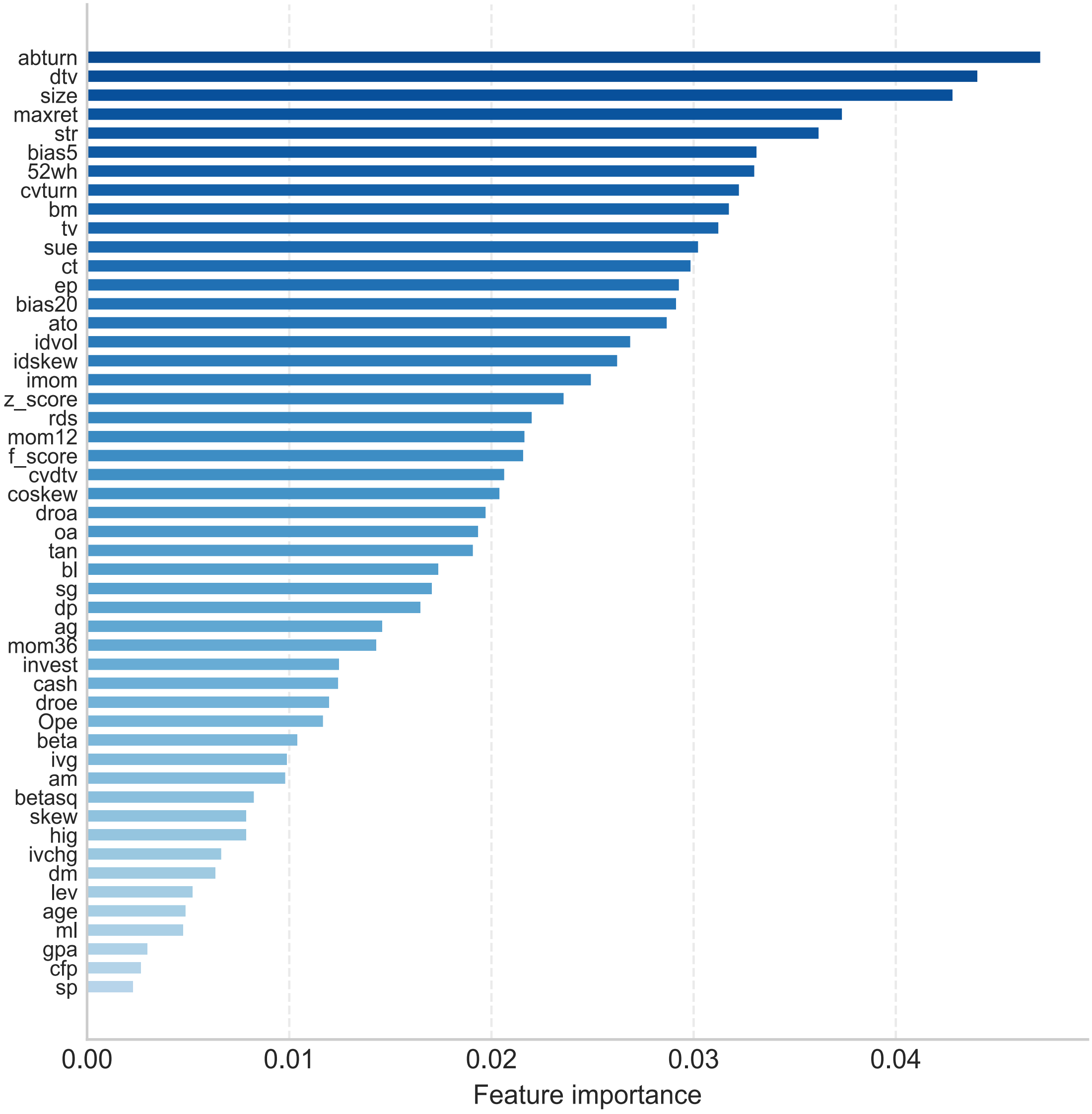

Figure 2. Feature importance.

4. Conclusions

This study demonstrates the effectiveness of LightGBM in predicting stock returns and identifying key pricing factors in the Chinese A-share market. The empirical results reveal two main findings. First, LightGBM significantly outperforms traditional OLS regression in return prediction, achieving a monthly out-of-sample R² of 2.13% compared to 0.95% for OLS. This superior predictive power translates into substantial economic gains, with the LightGBM-based long-short strategy generating a monthly return of 2.63% and a Sharpe ratio of 1.77, while maintaining better downside protection.

Second, the feature importance analysis provides new insights into the return-generating process in China's stock market. Consistent with Leippold et al. (2021), liquidity and volatility-related characteristics emerge as the dominant predictors, with measures such as abnormal turnover, trading volume, and price volatility demonstrating the strongest predictive power.

These findings have important implications for both academic research and investment practice. From an academic perspective, they highlight the value of machine learning approaches in capturing complex, non-linear relationships in asset pricing. For practitioners, the results suggest that incorporating advanced machine learning techniques and focusing on market microstructure factors could enhance investment strategies in the Chinese stock market.

Appendix A

Table A1. Details on stock characteristics.

No	Acronym	Stock Characteristics	Author(s)	Year, Journal
1	abturn	Abnormal turnover	Liu et al.	2016, JFE
2	ag	Asset growth	Cooper et al.	2008, JF
3	age	Firm age	Jiang et al.	2005, RAS
4	am	Asset to market equity	Bhandari	1988, JF
5	ato	Asset turnover	Soliman	2008, TAR
6	beta	Beta	Fama & MacBeth	1973, JPE
7	betasq	Beta squared	Fama & MacBeth	1973, JPE
8	bias5	The 5-day bias	Zhang & Wu	2009, ESA
9	bias20	The 20-day bias	Zhang & Wu	2009, ESA
10	bl	Book leverage	Fama & French	1992, JF
11	bm	Book to market	Rosenberg et al.	1985, JPM
12	cash	Cash holdings	Palazzo	2012, JFE
13	cfp	Cash flow to price	Desai et al.	2004, TAR
14	coskew	Skewness coefficient	Harvey & Siddique	1999, JFQA
15	ct	Capital turnover	Hou et al.	2020, RFS
16	cvturn	Coefficient of Variation of Share Turnover	Chordia et al.	2001, JFE
17	cvdtv	Coefficient of Variation of Trading Volume	Chordia et al.	2001, JFE
18	dm	Debt to market equity	Hou et al.	2020, RFS
19	dp	Dividend to price	Litzenberger et al.	1982, JF
20	dtv	Trading volume	Brennan et al.	1998, JFE
21	droa	Change in return on asset	Hou et al.	2020, RFS
22	droe	Change in return on equity	Hou et al.	2020, RFS
23	ep	Earnings to price	Basu	1977, JF
24	f_score	F score	Piotroski	2000, JAR
25	gpa	Gross profit to asset	Novy-Marx	2013, JFE
26	hig	Employee growth rate	Bazdresch et al.	2014, JPE
27	idvol	Idiosyncratic volatility	Ali et al.	2003, JFE
28	idskew	Idiosyncratic skewness	Boyer et al.	2010, RFS
29	imom	Idiosyncratic momentum	Blitz et al.	2011, JEF
30	invest	Capital expenditures and inventory	Chen & Zhang	2010, JF
31	ivchg	Inventory changes	Hou et al.	2020, RFS
32	ivg	Inventory growth	Hou et al.	2020, RFS
33	lev	Leverage	Bhandari	1988, JF
34	maxret	Maximum daily return	Bali et al.	2011, JFE
35	mom12	12-month momentum	Jegadeesh	1990, JF
36	mom36	36-month momentum	Jegadeesh &Titman	1993, JF
37	ml	Market leverage	Fama & French	1992, JF
38	oa	Operating accruals	Hribar & Collins	2002, JAR
39	Ope	Operating profits to book equity	Fama & French	2015, JFE
40	rds	R&D to sales	Guo et al.	2006, JBFA
41	sp	Sales to price	Barbee et al.	1996, FAJ
42	skew	Skew	Amaya	2015, JFE
43	size	Size	Banz	1981, JFE
44	sue	Standardized unexpected earnings	Foster et al.	1984, TAR
45	str	Short term reversal	Jegadeesh	1990, JF
46	sg	Sales growth	Lakonishok	1994, JF
47	tv	Total volatility	Ang et al.	2010, JF
48	tan	Debt capacity/firm tangibility	Almeida et al.	2007, RFS
49	52wh	The highest return in 52-week	George et al.	2004, JF

50	z_score	Z score	Altman	1968, JF

References

1. Carmona, P., Climent, F., & Momparler, A. **2019**. Predicting failure in the US banking sector: An extreme gradient boosting approach. International Review of Economics & Finance, 61, 304-323.
2. Chen, X., Mao, Z., & Wu, C. **2024**. Multi-class Financial Distress Prediction Based on Feature Selection and Deep Forest Algorithm. Computational Economics, 1-40
3. Deng, S., Zhu, Y., Huang, X., Duan, S., & Fu, Z. **2022**. High-frequency direction forecasting of the futures market using a machine-learning-based method. Future Internet, 14(6), 180.
4. Ferrouhi, E. M., & Bouabdallaoui, I. **2024**. A comparative study of ensemble learning algorithms for high-frequency trading. Scientific African, 24, e02161.
5. Gu, S., Kelly, B., & Xiu, D. **2020**. Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223-2273.
6. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. **2017**. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.
7. Leippold, M., Wang, Q., & Zhou, W. **2021**. Machine learning in the Chinese stock market. Journal of Financial Economics, 145(2), 64-82.
8. Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. **2015**. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124-136.
9. Lin, W., Hu, Y., & Tsai, C. **2022**. Machine learning in financial crisis prediction: A survey. IEEE Transactions on Systems, Man, and Cybernetics, 42(4), 421-436.
10. Lundberg, S. **2017**. A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874.
11. Ma, T., Wang, W., & Chen, Y. **2023**. Attention is all you need: An interpretable transformer-based asset allocation approach. International Review of Financial Analysis, 90, 102876.
12. Nti, K. O., Adekoya, A., & Weyori, B. **2019**. Random forest based feature selection of macroeconomic variables for stock market prediction. American Journal of Applied Sciences, 16(7), 200-212.
13. Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Vlachogiannakis, N. E. **2020**. Predicting bank insolvencies using machine learning techniques. International Journal of Forecasting, 36(3), 1092-1113.
14. Siswara, D., Soleh, A. M., & Wigena, A. H. **2024**. Classification Modeling with RNN-Based, Random Forest, and XGBoost for Imbalanced Data: A Case of Early Crash Detection in ASEAN-5 Stock Markets. arXiv preprint arXiv:2406.07888.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.