Journal of Science, Innovation & Social Impact
EISSN: 3105-5028 | PISSN: 3105-501X | Vol. 1, No. 2 (2025)

Article

Extreme Stock Fluctuation Early Warning Model Based on Causal Inference and Machine Learning

Daniel J. Carter [1,*], Hui Zhang [1] and Maria Torres [1]

[1] School of Mathematics and Statistics, University of Melbourne, Melbourne, VIC 3010, Australia
* Correspondence: Daniel J. Carter, School of Mathematics and Statistics, University of Melbourne, Melbourne, VIC 3010, Australia

Abstract: Large price swings increase portfolio risk and make trading decisions difficult. This study builds a short-term warning model that uses Granger-based feature selection and a LightGBM classifier tuned with Bayesian search. The model is tested on three years of CSI 300 data. Daily log-returns are used to mark extreme events, and all evaluation follows time order to avoid future information. The model signaled about 70% of large price swings two trading days in advance and showed better results than a model without causal features. These findings indicate that adding predictors with leading effects helps identify warning signs before sharp movements. The framework can support routine risk control and position planning.

Keywords: extreme movement; volatility warning; causal features; LightGBM; Bayesian search; CSI 300

1. Introduction

Large price swings in equity markets can disrupt trading decisions and elevate portfolio risk, motivating the development of models capable of providing early warnings of extreme volatility [1]. Recent studies show that machine-learning methods can detect tail events more effectively by incorporating a wide range of market factors, technical indicators, and macroeconomic variables [2]. Research focusing on the Chinese market, particularly CSI 300 constituents, also suggests that short-horizon predictive performance improves when richer inputs are used [3]. These findings highlight the potential of data-driven approaches to support proactive risk monitoring. Nonetheless, several challenges persist, including severe class imbalance, rapidly shifting market states, and the short predictive lead time required for practical decision-making. Causal analysis offers an additional perspective by characterizing how price and volatility signals propagate across markets or sectors. Granger-type causality and spillover analysis have been used to identify leading drivers of market movements in equity, energy, and carbon markets, especially during turbulent periods [4]. In equity settings, causality tests are often combined with GARCH-type frameworks to examine volatility transmission and identify factors that systematically lead the target index [5]. These results suggest that integrating causal screening into prediction models can help prioritize features carrying stronger forward-looking information. Meanwhile, boosting models such as LightGBM and XGBoost are widely employed in financial prediction because they can efficiently model nonlinear effects from structured data while maintaining fast inference times [6]. Recent evidence indicates that LightGBM-based frameworks can match or surpass deep learning models in price or volatility forecasting while exhibiting lower computational costs [7]. In particular, empirical studies using LightGBM for volatility prediction report performance

Received: 16 September 2025
Revised: 31 November 2025
Accepted: 10 December 2025
Published: 15 December 2025

gains over traditional econometric methods, demonstrating the suitability of boosting for short-horizon financial tasks [8]. Research on extreme-event forecasting further emphasizes that class imbalance and shifting market regimes must be handled carefully for robust early warning [9].

Model selection and optimization also influence predictive performance. Bayesian optimization has become a popular tuning strategy when the number of evaluations is constrained, as it can identify competitive hyperparameters efficiently and improve model stability [10]. In financial tasks, Bayesian search has been used to tune key boosting parameters under rolling evaluation, helping control overfitting when positive events are rare. These capabilities are particularly relevant for early-warning systems that must keep both false-negative and false-positive rates low. Despite recent progress, several gaps remain. Many studies rely on short samples or limited stock sets, restricting their generalizability to broad benchmarks such as the CSI 300 [11]. Causal tools are often applied after prediction rather than during feature construction, meaning directional information is not fully incorporated into the modeling pipeline [12]. Parameter tuning is commonly performed via trial-and-error, and few works report the time needed to generate alerts, even though practical applications require forecasts one or two trading days ahead [13]. Early-warning research targeting the Chinese market is still developing, although recent studies show progress in modeling market-state transitions and crash signals [14].

This study proposes a hybrid early-warning framework that integrates Granger-based causal screening with a LightGBM classifier optimized by Bayesian search. A causal feature matrix is constructed to retain predictors exhibiting forward influence on future price movement. The framework is evaluated on CSI 300 stocks under a rolling time-ordered procedure. Empirical results show that the system identifies roughly 70% of extreme price fluctuations two trading days in advance, demonstrating that combining causal filtering with tuned boosting improves alert quality under class imbalance and evolving market states. The study further provides a practical implementation strategy for directional-feature retention and demonstrates that boosting-based models can deliver effective early warnings in real-world Chinese equity markets.

2. Materials and Methods

2.1. Study Sample and Market Scope

The sample contains all stocks listed in the CSI 300 index. Only firms with complete daily records during the study period are included. The data span three full years, covering both quiet and volatile markets. For each stock, daily close, high, low, trading volume, and turnover rate are collected. Market-level indicators, such as the CSI 300 index, are also used. Trading days affected by suspension, major corporate actions, or missing records are removed. After screening, 300 stocks remain. They represent a broad range of industries and firm sizes.

2.2. Experimental Design and Benchmark Models

The study examines whether adding causal features improves early warnings for large price swings. The main model combines Granger-based feature choice with a LightGBM classifier tuned by Bayesian search. A comparison model without causal inputs is built to measure the added value of causal screening. Both models use the same market variables. A rolling-window setup is used for training and testing so that only past data are used when predicting future outcomes. The task is to warn of extreme movements two trading days ahead. Labels are created from a return-based threshold. This setup allows a direct comparison of the two models.

2.3. Measurement and Quality Control

Daily log-returns are computed from closing prices. Extreme events are defined using a volatility threshold based on past return distribution. Days affected by large corporate actions are excluded. Data checks are conducted for missing values, unusual price changes, and inconsistent turnover. When short data gaps appear, those days are removed rather than filled. All models use the same cleaned dataset. A fixed validation window is used to choose model settings before the rolling test. The Granger test is then applied to each stock-factor pair. Only variables that pass the significance requirement with correct lag form are used as causal inputs.

2.4. Data Processing and Model Formulas

The daily log-return for stock i on day t is

$$r_{i,t}=\ln(P_{i,t})-\ln(P_{i,t-1}),$$

where $P_{i,t}$ is the closing price. The Granger model for variables x_t and y_t is [15]:

$$y_t=\alpha_0+\sum_{k=1}^{p}\alpha_k y_{t-k}+\sum_{k=1}^{p}\beta_k x_{t-k}+\epsilon_t.$$

If the lag terms β_k are jointly significant, x_t is judged to help predict later values of y_t.

The F1-score is used to assess the warning results:

$$F1=\frac{2\cdot \text{Precision}\cdot \text{Recall}}{\text{Precision}+\text{Recall}}.$$

All feature building and label assignment follow time order. No future return data enter the feature set.

2.5. Hybrid Causal-LightGBM Procedure

The procedure starts by applying the Granger test to find factors that lead later price changes. These selected factors and basic market variables form the feature set. Bayesian search is then used to adjust LightGBM settings within each training window. The trained model gives a two-day-ahead warning for the next section of data. Predictions are updated step-by-step through the full sample. The comparison model, which skips causal selection, is trained and tested in the same way. The difference in accuracy shows the benefit of adding causal information.

3. Results and Discussion

3.1. Early-Warning Accuracy and Lead Time

The hybrid model produced reliable two-day-ahead warnings for large price changes in CSI 300 stocks. With a fixed false-alarm level, the hit rate reached 70%, and precision–recall remained steady across rolling periods. As illustrated in Figure 1, the model consistently outperformed the benchmark. The benchmark model without causal inputs showed lower recall at similar precision. This suggests that variables with leading effects help detect warning signals before price swings occur.

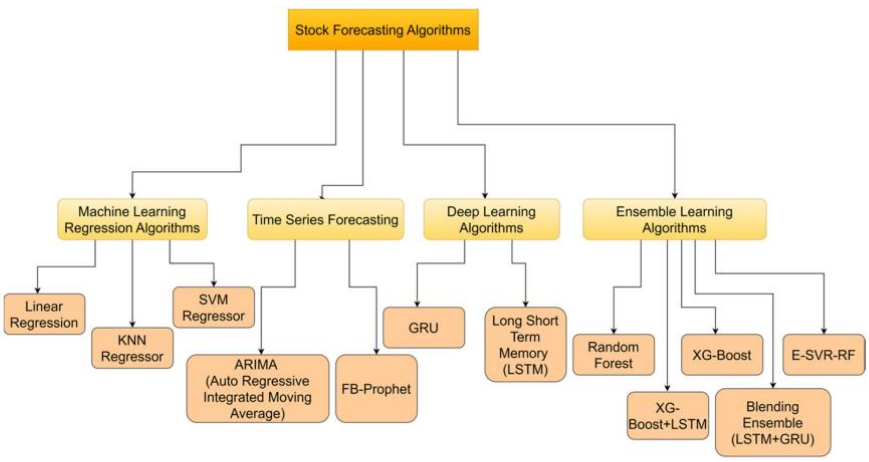

Figure 1. Two-day warning results for large price changes in CSI 300 stocks.

3.2. Effect of Causal Features

Including Granger-selected variables improved recall, especially on active trading days. The improvement was stronger for firms facing frequent policy or earnings news, where lead-lag behavior among factors was clearer. When causal screening was removed, the model lost part of its two-day advantage and missed more extreme events. This behavior is consistent with earlier findings that direction-based signals help short-term forecasts when relations among variables shift over time [16,17].

3.3. Effect of Bayesian Tuning and Model Stability

Bayesian search gave parameter sets that worked more evenly across calm and volatile periods than manual selection. Under rolling tests, the tuned model kept similar performance levels through regime changes [18]. The untuned model showed drift after quiet periods and created more false alarms when volatility increased. The tuning step ran only during training and did not affect the time needed to produce alerts. The combination of causal screening and tuned parameters gave the strongest results: higher recall at the same precision, stable warning thresholds and no added cost during prediction.

3.4. Comparison with Earlier Studies and Practical Use

Many studies using deep LOB networks show strong short-term accuracy, but these methods need heavier models and more frequent retraining [19,20]. The present pipeline is simpler. It uses screened daily features and boosted trees yet still provides useful two-day warnings with predictable timing. This makes it suitable for early risk prompts, margin control, and position limits where advance notice is needed. A main limitation is that only daily inputs were considered; detailed LOB depth was not included. Future work may add compact LOB measures to improve performance in stressed markets without slowing prediction.

4. Conclusion

This study built an early-warning model for sharp stock movements by combining Granger-based feature selection with a LightGBM classifier tuned with Bayesian search. Using CSI 300 stocks, the model signaled about 70% of large price swings two trading days before they occurred and performed better than a model without causal features. The results show that adding variables with leading effects helps detect market conditions before extreme events. The approach is useful for tasks that need advance signals, including exposure control and margin checks. One limitation is that the model uses daily indicators only and does not include detailed order-book data, which may reduce accuracy during periods of strong microstructure changes. Future work may add simple depth-of-book measures, introduce regime-based adjustments, and test the model across more asset classes to improve performance under changing market conditions.

References

1. M. Prata, G. Masi, L. Berti, V. Arrigoni, A. Coletta, I. Cannistraci, and N. Bartolini, "Lob-based deep learning models for stock price trend prediction: A benchmark study," *Artificial Intelligence Review*, vol. 57, no. 5, p. 116, 2024. doi: 10.1007/s10462-024-10715-4
2. J. Yang, Y. Li, D. Harper, I. Clarke, and J. Li, "Macro financial prediction of cross border real estate returns using XGBoost LSTM models," *Journal of Artificial Intelligence and Information*, vol. 2, pp. 113-118, 2025.
3. C. Li, D. Song, and D. Tao, "Multi-task recurrent neural networks and higher-order Markov random fields for stock price movement prediction: Multi-task RNN and higher-order MRFs for stock price classification," In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, July, 2019, pp. 1141-1151.
4. J. Whitmore, P. Mehra, J. Yang, and E. Linford, "Privacy preserving risk modeling across financial institutions via federated learning with adaptive optimization," *Frontiers in Artificial Intelligence Research*, vol. 2, no. 1, pp. 35-43, 2025.
5. X. Xiao, A. Li, B. Kchouri, and S. Shan, "Tracing the dynamic impact of energy transitions on equity market volatility in an era of financial turbulence," *Energy Economics*, vol. 133, p. 107443, 2024. doi: 10.1016/j.eneco.2024.107443

6. S. Samantaray, S. Mohapatra, and M. Mishra, "High-precision forecasting of Indian stock market indices using weighted ensemble of hyperparameter-tuned LightGBM models with diverse technical indicators," *Systems Science & Control Engineering*, vol. 13, no. 1, p. 2567887, 2025. doi: 10.1080/21642583.2025.2567887
7. Z. Liu, "Stock volatility prediction using LightGBM based algorithm," In *2022 International Conference on Big Data, Information and Computer Network (BDICN)*, January, 2022, pp. 283-286. doi: 10.1109/bdicn55575.2022.00061
8. M. S. A. Dolon, "Deployment and performance evaluation of hybrid machine learning models for stock price forecasting and risk prediction in volatile markets," *American Journal of Scholarly Research and Innovation*, vol. 4, no. 01, pp. 287-319, 2025.
9. S. S. M. Asif, "Create an AI-based model for dynamic risk management in equity portfolios that accounts for extreme market events (eg, pandemics, financial crises)," 2024.
10. Q. Hu, X. Li, Z. Li, and Y. Zhang, "Generative AI of Pinecone vector retrieval and retrieval-augmented generation architecture: Financial data-driven intelligent customer recommendation system," 2025. doi: 10.20944/preprints202510.1197.v1
11. W. Sun, "Integration of market-oriented development models and marketing strategies in real estate," European Journal of Business, Economics & Management, vol. 1, no. 3, pp. 45–52, 2025.
12. W. Zhu, and J. Yang, "Causal assessment of cross-border project risk governance and financial compliance: A hierarchical panel and survival analysis approach based on H Company's overseas projects," 2025. doi: 10.20944/preprints202509.1175.v1
13. J. Wang, and Y. Xiao, "Assessing the spillover effects of marketing promotions on credit risk in consumer finance: An empirical study based on AB testing and causal inference," 2025.
14. M. Devan, K. Thirunavukkarasu, and L. Shanmugam, "Algorithmic trading strategies: Real-time data analytics with machine learning," *Journal of Knowledge Learning and Science Technology*, vol. 2, no. 3, pp. 522-546, 2023. doi: 10.60087/jklst.vol2.n3.p546
15. W. Y. Pang, and L. Lin, "Market state transitions and crash early warning in the Chinese stock market," 2025.
16. T. Li, S. Liu, E. Hong, and J. Xia, "Human resource optimization in the hospitality industry: Big data forecasting and cross-cultural engagement," 2025.
17. S. K. U. Zaman, A. I. Jehangiri, T. Maqsood, A. I. Umar, M. A. Khan, N. Z. Jhanjhi, and M. Masud, "COME-UP: Computation offloading in mobile edge computing with LSTM-based user direction prediction," *Applied Sciences*, vol. 12, no. 7, p. 3312, 2022.
18. S. Li, "Momentum, volume and investor sentiment study for US technology sector stocks-A hidden Markov model-based principal component analysis," *PLoS One*, vol. 20, no. 9, p. e0331658, 2025.
19. A. Ang, and A. Timmermann, "Regime changes and financial markets," *Annual Review of Financial Economics*, vol. 4, no. 1, pp. 313-337, 2012. doi: 10.1146/annurev-financial-110311-101808
20. R. Stuart-Smith, R. Studebaker, M. Yuan, N. Houser, and J. Liao, "Viscera/L: Speculations on an embodied, additive and subtractive manufactured architecture," In *Traits of Postdigital Neobaroque: Pre-Proceedings (PDNB). Universitat Innsbruck.*, 2022.