Causal Machine Learning for Identifying Market Stress Drivers: A Leakage-Safe Walk-Forward Investigation

Lei Tian Zhu
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
zhulei@uc.edu

Rowan Moran
Department of Computer Science, University of Houston, Houston, TX, USA.
hellorowan@uh.edu

Abstract

The early identification of market stress drivers is essential for maintaining financial stability, yet conventional machine learning approaches suffer from persistent causal confusion and evaluation leakage that undermine their practical credibility. This paper develops a leakage-safe walk-forward framework that integrates causal machine learning with rigorous temporal validation to isolate structural drivers of market drawdowns. We argue that the standard cross-validation protocols commonly adopted in financial machine learning violate the temporal ordering of information, leading to inflated performance metrics and spurious causal attributions. By contrast, our proposed architecture enforces strict temporal independence across training, validation, and test partitions while embedding causal inference methods—including double machine learning and heterogeneous treatment effect estimation—within a walk-forward loop. The framework is designed to separate persistent market stress signals from transient volatility, thereby enabling more robust early-warning systems. We examine the system-level trade-offs between predictive accuracy, interpretability, and computational sustainability, and discuss the governance implications of deploying causal models in high-stakes financial regulation. Through a series of conceptual case illustrations drawn from recent literature on leakage-safe evaluation and causal discovery, we demonstrate how the proposed methodology mitigates overfitting to spurious correlations and enhances the economic credibility of stress detection. The paper concludes with forward-looking recommendations for the design of causal, leakage-resilient infrastructures in financial surveillance.

Keywords

causal machine learning; market stress detection; walk-forward validation; data leakage; financial stability; drawdown risk; interpretability; governance.

1. Introduction

The increasing frequency and severity of market stress events—ranging from flash crashes to systemic liquidity crises—have intensified the demand for reliable early-warning systems that can identify underlying drivers before they escalate into full-blown dislocations. Machine learning models, particularly those based on tree ensembles and deep neural architectures, have demonstrated remarkable capacity to capture nonlinear patterns in financial time series [1], [4]. Yet a growing body of evidence suggests that many reported successes in financial

forecasting are artifacts of methodological flaws, most notably information leakage arising from improper temporal splitting of data [5], [8]. When training and test periods intermingle through overlapping features or look-ahead bias, the resulting models exhibit deceptively high performance that disappears in out-of-sample deployment.

Beyond the problem of leakage, the machine learning community has increasingly recognized that predictive accuracy alone is insufficient for decision-making in complex socio-technical systems. Financial regulators and risk managers require not just forecasts of stress events but also causal explanations of why those events occur. Without causal understanding, interventions based on model outputs risk being misdirected or counterproductive. This has motivated the emergence of causal machine learning, a paradigm that seeks to estimate the causal effects of interventions or treatments on outcomes of interest, thereby enabling policy-relevant insights [2], [3]. In the context of market stress detection, causal machine learning can help distinguish between genuine structural drivers—such as sudden shifts in leverage or liquidity—and mere statistical correlations that arise from common confounding factors.

The central contribution of this paper is the design of a leakage-safe walk-forward framework that integrates causal machine learning for the identification of market stress drivers. We draw on recent advances in leakage-aware evaluation protocols [5], [6] and causally interpretable modeling to propose an architecture that respects the temporal ordering of financial data while recovering structural relationships. Our emphasis is on system-level considerations: the trade-offs between model complexity and generalizability, the computational overhead of walk-forward validation, the sustainability of repeated retraining cycles, and the governance challenges of deploying causal models in regulatory settings. By positioning the framework within the broader literature on robust financial infrastructure, we aim to bridge the gap between academic causal inference and practical market surveillance.

2. Background and Related Work

The challenge of market stress detection has historically been approached through volatility modeling, extreme value theory, and regime-switching models [4]. Volatility, however, is a noisy proxy for stress; periods of high volatility may reflect healthy market adjustments rather than systemic fragility. Recent work has introduced residual stress signals that isolate drawdown risk from ordinary volatility by conditioning on factor exposures and macroeconomic regimes [6]. These signals serve as candidate outcome variables for causal machine learning, as they embed economically meaningful definitions of stress that are less prone to false alarms.

Machine learning models applied to financial time series have proliferated in the past decade, including random forests, gradient boosting, and transformers [14], [15]. While these models achieve high in-sample fit, their out-of-sample performance often degrades sharply when standard k-fold cross-validation is replaced with temporal walk-forward validation [8]. The discrepancy arises because financial data exhibit serial dependence, non-stationarity, and evolving regimes that violate the i.i.d. assumption underlying cross-validation. A leakage-safe evaluation design enforces that the training window precedes the validation window in calendar time, and that no future information is used to construct features or labels at any point [5]. This design is particularly critical for early-warning systems intended to be deployed in real time.

Causal inference in time series has a rich tradition in econometrics, including vector autoregressions, Granger causality, and instrumental variable methods. However, these

classical approaches are often limited by linearity assumptions and the difficulty of handling high-dimensional controls. Causal machine learning relaxes these constraints by leveraging flexible function approximators to estimate conditional average treatment effects (CATE) and average treatment effects (ATE) in the presence of confounding [3], [16]. Methods such as double machine learning (DML) use Neyman-orthogonal scores and cross-fitting to achieve root-n consistency under weaker assumptions, making them attractive for financial applications where the true functional form is unknown [1]. The integration of DML into a walk-forward loop introduces additional layers of complexity: the orthogonalization step must be performed within each temporal window to avoid leakage, and the treatment variable must be defined in a way that respects temporal precedence.

The literature on financial stress detection has also explored interpretability as a governance requirement. Regulatory frameworks such as SR 11-7 in the United States mandate that model outputs be explainable and that assumptions be validated. Black-box models, even when accurate, face resistance in adoption because stakeholders cannot interrogate their reasoning [10], [11]. Causal machine learning offers a path toward interpretability by design: rather than post-hoc explanations, the models themselves estimate causal quantities that have direct economic interpretations. For instance, the effect of a sudden increase in margin requirements on market stress can be estimated as a treatment effect, providing actionable information to clearinghouses and central counterparties.

3. Causal Machine Learning in Financial Time Series

The application of causal machine learning to financial time series requires careful consideration of the treatment assignment mechanism. In a typical market stress context, treatments may correspond to policy interventions, regulatory changes, or exogenous shocks. However, many candidate drivers of market stress are not randomly assigned; they are correlated with the state of the economy, market sentiment, and latent confounders. For example, a central bank interest rate decision is influenced by the same macroeconomic conditions that affect market stress, creating a classical confounding problem. Standard predictive models would conflate correlation with causation, potentially leading to erroneous attributions.

Double machine learning addresses this by using machine learning to estimate nuisance functions—the conditional expectation of the outcome given confounders and the conditional probability of treatment given confounders—and then constructing an orthogonalized score that is insensitive to errors in these nuisance estimates [1], [3]. This procedure can be embedded within a walk-forward loop as follows. At each step of the walk-forward process, a fixed-length training window is used to fit the nuisance models and the causal parameter of interest. The orthogonalized score is then computed on the subsequent validation window using the treatment and outcome from that period, but the nuisance models are not refitted on the validation data, thereby preventing leakage of future information into the training phase. The resulting sequence of causal estimates provides a time-varying picture of how the effect of a given driver evolves across market regimes.

The choice of treatment variable is critical. A common approach is to define treatment as an indicator for whether a specific financial variable exceeds a threshold, such as a jump in volatility, a credit spread widening, or a liquidity dry-up. However, such binary treatments risk losing information. Continuous treatments can be accommodated using causal machine learning methods for dose-response functions, but these require stronger assumptions and larger sample sizes. In the leakage-safe framework, the treatment must also be defined in a

way that is predictable from past information only; otherwise, the causal estimate would incorporate look-ahead bias [5], [8]. For instance, if treatment is defined as a sudden drop in the VIX index, the drop itself must be measured at time t based on data available up to time t, and the outcome stress signal must occur at time t+1 or later.

A further nuance arises from the presence of multiple, interacting drivers. Market stress events are typically the result of a confluence of factors—rising leverage, falling liquidity, and negative sentiment—that amplify one another. Causal machine learning can estimate heterogeneous treatment effects by modeling how the impact of one driver varies with the values of other drivers [16], [13]. This is particularly valuable for early-warning systems because it allows the identification of threshold conditions under which a given driver becomes dangerous. For example, a modest increase in margin debt may have no effect on stress when liquidity is abundant, but the same increase can be destabilizing when liquidity is already scarce. Tree-based causal models, such as causal forests, naturally capture such interactions without requiring explicit specification.

4. System Architecture for Leakage-Safe Walk-Forward Identification

Building a robust system for market stress driver identification requires not only an appropriate causal estimator but also an infrastructure that enforces temporal integrity and supports continuous monitoring. We propose an architecture organized around a walk-forward backtesting engine, a causal estimation module, a signal aggregation layer, and a governance dashboard. Each component must be designed to prevent information leakage at every level of data flow.

The walk-forward engine partitions the historical data into a sequence of expanding or rolling windows. At each iteration, a training window of fixed length is used to fit the nuisance functions—these are the machine learning models that estimate the conditional mean of the stress outcome and the propensity score for the driver of interest. The engine then advances one step forward, applies the trained nuisance functions to the new validation window to compute orthogonalized scores, and updates the causal effect estimate. This process is repeated across the entire time series, generating a trajectory of causal estimates. The engine must ensure that no future observations are used to tune hyperparameters or select features. Hyperparameter optimization, if performed, must be conducted using an early validation set that is temporally separated from both training and subsequent test windows [5], [8].

The causal estimation module houses the core algorithmic machinery. For the double machine learning approach, the module must implement cross-fitting within each training window to avoid overfitting in the nuisance estimation step. This means that the training window itself is further split into two sub-windows: one for fitting the machine learning models and one for computing the cross-fitted scores. The splitting must be performed temporally: the sub-window used for fitting must precede the sub-window used for scoring, because the scoring step involves constructing residuals that will be used to estimate the causal parameter. This temporal nesting of splits ensures that the residuals are not contaminated by in-sample overfitting. The computational cost of nested walk-forward cross-fitting is substantial, but it is necessary for achieving the statistical guarantees that make the causal estimates credible.

The signal aggregation layer translates the time-varying causal effect estimates into actionable early-warning indicators. A single driver may show a nonzero causal effect during only certain regimes; the aggregation layer must thus incorporate regime detection, possibly through a separate unsupervised clustering of market states. For each regime, the layer

computes the average causal effect and its confidence interval, flagging drivers whose effects are statistically and economically significant. The aggregation also produces a composite stress index that integrates multiple drivers, weighting them by their estimated causal contributions. This composite can be monitored in real time and can trigger alerts when it crosses predefined thresholds.

The governance dashboard provides transparency to regulators, risk managers, and model auditors. It displays the historical trajectories of causal estimates, the stability of the nuisance models, and the results of leakage-check diagnostics. Auditors can inspect the temporal splits to verify that no future information has been used. The dashboard also tracks the computational resources consumed by each walk-forward iteration, enabling sustainability assessments. In production deployment, the architecture must support periodic retraining—daily or weekly—without requiring a full historical re-estimation. Incremental update strategies, where new observations are appended to the training window and nuisance models are updated online, can reduce computational overhead while maintaining leakage safety.

5. Empirical Considerations and Validation

The validation of a leakage-safe causal machine learning system poses unique challenges because ground truth causal effects are rarely observable in financial data. Unlike in randomized controlled trials, we cannot expose a financial system to a known intervention and observe the counterfactual outcome. Therefore, validation must rely on a combination of synthetic experiments, stress-test scenarios, and out-of-sample predictive checks. The predictive check is particularly instructive: if the causal model identifies a driver as having a large effect, then an early-warning system based solely on that driver should outperform a baseline system that ignores it, at least in periods where the driver is active.

A crucial aspect of validation is distinguishing between causal drivers and mere leading indicators. A leading indicator, such as an inverted yield curve, may predict recessions well without being a cause of them. Causal machine learning can help separate these by conditioning on a rich set of confounders. If the estimated causal effect of the yield curve slope becomes negligible after conditioning on monetary policy expectations and credit conditions, then the yield curve is likely a proxy rather than a driver. This type of analysis requires high-quality data on potential confounders, which is often unavailable at the required frequency. Missing confounders can lead to bias, underscoring the need for sensitivity analyses.

Recent research has developed leakage-safe benchmark designs for market stress early-warning systems, providing evaluation frameworks that compare candidate models on a level playing field [5]. These benchmarks enforce temporal ordering, require out-of-sample predictions for economic loss functions (such as portfolio drawdown), and penalize models that overfit to specific historical windows. The methods proposed in this paper can be evaluated against such benchmarks. Preliminary results from related work indicate that leakage-safe causal models outperform purely predictive models in terms of the stability of their performance across different evaluation periods [6], [12]. They also exhibit lower variance in their feature importance rankings, which enhances trust among human decision-makers.

Another validation concern is the finite sample behavior of causal estimators. In short financial time series with infrequent stress events, the effective sample size for estimating causal effects can be very small. The walk-forward procedure exacerbates this by further

subdividing the data. One remedy is to pool information across drivers using hierarchical Bayesian methods, though this introduces modeling assumptions that must be carefully checked. Alternative approaches include using synthetic control methods to construct counterfactual stress paths, or leveraging high-frequency data to increase the number of effective observations per window.

6. Governance, Fairness, and Sustainability Implications

The deployment of causal machine learning for market stress identification raises a host of governance challenges that extend beyond technical performance. Financial regulators and systemically important financial institutions must ensure that the models are transparent, accountable, and free from discriminatory biases. Causal machine learning, with its emphasis on structural understanding, offers advantages over black-box models in meeting these governance requirements. For instance, if a model identifies a particular asset class or demographic group of investors as a driver of stress, regulators need to understand whether the effect is causal or spurious. Causal estimates provide a clearer basis for policy action: if the effect is causal, then regulating that asset class may reduce systemic risk; if the effect is merely correlational, regulatory intervention could be ineffective or harmful.

Fairness considerations enter when the treatment variable is correlated with protected characteristics. For example, if a driver such as “retail investor margin debt” is identified as a cause of stress, the subsequent regulatory response might disproportionately affect retail investors who may have less access to alternative funding sources. A causal machine learning system should therefore include fairness audits that examine whether the estimated treatment effects vary across subpopulations defined by income, geography, or other sensitive attributes. If the effect is concentrated in a particular subpopulation, the regulatory response should be targeted rather than blanket. The walk-forward framework allows for temporal fairness monitoring: one can test whether the magnitude of the causal effect on a disadvantaged subpopulation is stable over time or whether it changes during crisis periods.

Sustainability is another key concern. The computational cost of running a nested walk-forward double machine learning pipeline with deep neural networks or large tree ensembles can be substantial. For a daily retraining cycle spanning decades of data, the cumulative energy consumption and carbon footprint may be non-trivial. System architects must balance the demand for statistical rigor—which favors more complex models and longer training windows—against the environmental cost and the latency constraints of real-time monitoring. One sustainability strategy is to use simpler base learners, such as linear models with engineered features, in the nuisance estimation step, sacrificing some flexibility for reduced computational intensity. Another is to implement pruning and early stopping within the walk-forward loop to avoid training models that are unnecessarily complex for the available data.

From a policy perspective, the integration of causal machine learning into financial surveillance requires institutional buy-in and regulatory sandboxing. Central banks and supervisory authorities are traditionally cautious about adopting methods that lack long track records. A phased approach, where the causal early-warning system runs in parallel with existing models and is evaluated over several years, would build confidence. The leakage-safe design is particularly important for gaining regulatory acceptance, because it addresses a common criticism of machine learning models—that they perform well in backtests but fail in live deployment. By demonstrating that the causal estimates hold up under rigorous temporal validation, system developers can make a stronger case for deployment.

7. Conclusion

This paper has presented a comprehensive framework for identifying market stress drivers using causal machine learning within a leakage-safe walk-forward architecture. The central thesis is that predictive accuracy alone is insufficient for early-warning systems that must underpin regulatory and risk-management interventions. Causal machine learning, when embedded in a temporally sound evaluation protocol, can recover structural relationships that separate genuine drivers from spurious correlations. The framework addresses the twin challenges of information leakage and causal confounding by enforcing strict temporal independence and using orthogonalized estimation techniques.

The system architecture we have outlined integrates a walk-forward engine, a causal estimation module, a signal aggregation layer, and a governance dashboard. Each component is designed to prevent leakage while maintaining computational feasibility. We have discussed the trade-offs between model complexity and generalizability, the need for sensitivity analyses in the presence of unobserved confounders, and the importance of fairness and sustainability considerations. The adoption of such a framework by financial regulators and institutions could significantly improve the reliability of market stress detection, reducing the incidence of false alarms and missed warnings.

Future research directions include the development of online learning algorithms that maintain leakage safety while updating causal estimates in real time, the extension of the framework to handle multivariate treatment regimes, and the integration of counterfactual explanations to aid human interpretation. A particularly promising avenue is the combination of causal discovery methods—which learn the underlying causal graph from data—with the causal estimation framework described here. As the financial system becomes increasingly data-rich and interconnected, the need for credible, causally grounded early-warning systems will only grow. The leakage-safe walk-forward methodology provides a principled foundation for meeting that need.

References

1. Athey, S., & Imbens, G. (2016). The state of applied econometrics: Causality and policy evaluation. Journal of Economic Perspectives, 30(3), 3–32.
2. Pearl, J. (2009). Causality: Models, reasoning and inference (2nd ed.). Cambridge University Press.
3. Hernán, M. A., & Robins, J. M. (2020). Causal inference: What if. Chapman & Hall/CRC.
4. Andersen, T. G., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. International Economic Review, 39(4), 885–905.
5. Liu, T. (2026). Leakage-Safe Benchmark Design for Market-Stress Early Warning: An Economically Credible Evaluation.
6. Liu, T. (2026). Beyond volatility: A leakage-safe residual-stress signal for drawdown risk monitoring. Available at SSRN 6503179.
7. Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.
8. Liu, T. (2026). Interpretable Machine Learning for Volatility Forecasting Under Realistic Walk-Forward Constraints.

9. Liu, T. (2026). PCA-APT Stress Index for Market Drawdowns.

10. Varian, H. R. (2014). Big data: New tricks for econometrics. Journal of Economic Literature, 52(2), 331–359.

11. Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable. Independently published.

12. Liu, T. (2026). A Comparative Study of Transformer-Based and Classical Models for Financial Time-Series Forecasting. Journal of Risk and Financial Management, 19(3), 203.

13. Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. Review of Financial Studies, 33(5), 2223–2273.

14. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).

15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30.

16. Liu, T. (2026). Volatility Forecasting and Early-Warning Market Stress Detection: A Leakage-Safe Evaluation with Tree Ensembles and Transformers.

17. Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, 33(1), 3–56.

18. Carhart, M. M. (1997). On persistence in mutual fund performance. Journal of Finance, 52(1), 57–82.

19. Liu, T. (2022, December). Financial Constraint'Impact on Firms' ESG Rating Based on Chinese Stock Market. In 2022 4th International Conference on Economic Management and Cultural Industry (ICEMCI 2022) (pp. 1085-1095). Atlantis Press.

20. Hu, L., & Shen, Y. (2026). A predictive analytics approach for forecasting global stock index returns using deep learning techniques. Decision Analytics Journal, 100685.