ISSN: 3079-9325

Counterfactual Explanations for AML Risk Scores Under Causal Constraints

Sophia Bennett[1], Michael Turner[2], Ethan Morales[3]*

School of Information, University of California, Berkeley, CA 94720, USA

***Corresponding author:** ethan.morales@berkeley.edu

Abstract

This study introduces a counterfactual-explanation framework for anti-money-laundering (AML) risk scoring. The method distinguishes actionable from non-actionable variables within a causal graph and generates minimal valid changes needed to shift a case below a risk threshold. Experiments used 42.7 million transactions and 3.1 million customer profiles from a nationwide financial dataset. The system generated valid counterfactuals for 91.5% of high-risk cases, requiring an average of 2.4 feature changes per case. In a user evaluation with 26 compliance analysts, counterfactual explanations reduced review time by 22.6% and increased decision consistency from 0.62 to 0.75 (Cohen's kappa). Causal constraints eliminated unrealistic recommendations in 98.6% of generated explanations. The method enhances model transparency while maintaining detection performance.

Keywords

Counterfactual explanation; Causal inference; Anti-money-laundering; Interpretable AI; Risk scoring

1. Introduction

Anti-money-laundering (AML) systems rely on large-scale monitoring of financial transactions and customer activity to identify potential illicit behavior. With the rapid growth of digital payments and cross-border financial services, banks increasingly depend on machine-learning models to assign risk scores to customers and transactions in order to prioritize alerts and allocate compliance resources efficiently [1,2]. These models have significantly improved detection rates compared with rule-based systems, yet they also introduce new challenges for compliance teams and supervisors. In particular, the logic behind model-generated risk scores is often opaque, making it difficult for analysts to justify decisions such as enhanced due diligence, account restrictions, or off-boarding actions. Regulators have repeatedly emphasized that financial institutions must be able to explain why a customer or transaction is classified as high risk and what concrete changes would reduce that risk [3]. Recent regulatory guidance stresses that explanations are essential not only for auditability, but also for transparency, fairness, and consistency in AML operations [4]. Beyond regulatory compliance, interpretability has been linked to operational efficiency and institutional stability. Recent work combining causal reasoning with interpretable artificial

ISSN: 3079-9325

intelligence demonstrates that explanation-aware AML models can improve detection robustness while supporting broader financial system stability, highlighting the importance of explanation quality rather than prediction accuracy alone [5].Most existing research on explainable artificial intelligence (XAI) in AML focuses on feature-importance methods or simplified surrogate models. Techniques such as SHAP and LIME are now widely used to identify influential variables in customer risk scores and to support internal documentation, supervisory reviews, and model governance processes [6,7]. While these tools provide valuable insights into global or local model behavior, they remain largely descriptive. They explain *why* a model produced a particular score, but they do not indicate *how* that score could be changed. For AML analysts, this limitation is critical: understanding that a feature is important does not necessarily translate into actionable guidance for case resolution. More importantly, standard feature-attribution methods fail to distinguish between variables that can be realistically modified and those that are fixed or legally immutable, such as customer age, historical transactions, or jurisdictional attributes [8,9]. As a result, explanations may highlight drivers of risk without offering feasible mitigation strategies. This gap reduces the practical usefulness of XAI tools in day-to-day AML review, where analysts must decide whether a risk signal reflects genuine suspicious behavior or benign but unusual activity. Counterfactual explanations have emerged as a promising approach to address these limitations. Rather than summarizing feature importance, counterfactuals describe the smallest changes to input variables that would alter a model's prediction [10]. In recent years, research on counterfactual explanations has expanded rapidly, exploring properties such as sparsity, stability, diversity, robustness, and fairness [11,12]. In financial applications, particularly credit scoring, counterfactuals have been shown to help applicants and risk managers understand which behavioral or financial adjustments could lead to a different decision outcome [13]. However, most counterfactual-generation methods assume that all features can be freely manipulated and are statistically independent. This assumption is rarely valid in financial data. AML variables are governed by strong dependencies, operational constraints, and regulatory rules. Ignoring these factors often leads to counterfactuals that are implausible, internally inconsistent, or impossible to implement in practice. For example, suggested changes may violate accounting identities, transaction-generation processes, or customer behavior patterns, limiting their credibility and acceptance by compliance teams [14]. To mitigate these issues, recent studies have incorporated causal reasoning into counterfactual explanation frameworks. Structural causal models enable interventions to propagate through a predefined causal graph, ensuring that related variables adjust

ISSN: 3079-9325

coherently under hypothetical changes [15]. Several approaches now generate counterfactuals directly from causal structures, improving plausibility and alignment with real-world data-generating mechanisms [16]. Constraint-aware methods in credit risk further demonstrate that embedding domain rules can significantly reduce invalid or misleading recommendations [17].Despite these advances, the application of causal, constraint-guided counterfactual explanations in AML risk scoring remains largely unexplored. AML presents unique challenges: causal relationships are complex, regulatory constraints are strict, and explanations must support time-sensitive human decision-making. At the same time, empirical AML research shows that while black-box models improve detection rates, they often increase review time and reduce consistency across analysts [18]. Prior studies argue that explanations should actively support analyst judgment and workflow efficiency, rather than merely providing visual or post hoc summaries [19]. Yet, there is little empirical evidence on whether counterfactual explanations—especially those grounded in causal reasoning—improve AML review quality or analyst agreement in practice. This study aims to address these gaps. We propose a counterfactual explanation framework for AML risk scoring that integrates a domain-specific causal graph with operational and regulatory constraints. The method explicitly separates mutable and immutable variables, encodes their causal relationships, and searches for minimal, valid interventions that reduce high-risk scores while preserving data realism. We evaluate the approach using 42.7 million transactions, 3.1 million customer profiles, and a controlled user study involving 26 professional compliance analysts. The results demonstrate that the proposed method generates valid counterfactuals for the majority of high-risk cases, substantially reduces review time, improves inter-analyst agreement, and nearly eliminates unrealistic recommendations. These findings provide empirical evidence that causal, constraint-guided counterfactual explanations can enhance both transparency and practical decision-making in AML risk scoring, supporting regulatory expectations while improving operational effectiveness.

2. Materials and Methods

2.1 Sample and Study Setting

The study uses 42.7 million transaction records and 3.1 million customer profiles from a nationwide retail bank. The transactions include cash deposits, withdrawals, card payments, domestic transfers, and cross-border remittances. Each transaction contains a timestamp, channel type, counterparty details, and standard AML risk fields. Customer profiles include age group, account age, occupation type, and other basic attributes required by regulation. All personal information was anonymized before use. Only customers with complete profiles and

ISSN: 3079-9325

at least one recorded transaction during the 18-month study period were included in the analysis.

2.2 Experimental Design and Comparison Groups

The experiments assess two types of counterfactual explanations. The first type, used as the control group, applies counterfactual search without causal limits. The second type applies the proposed method, which restricts changes to variables marked as actionable and updates related variables according to the causal links. Both groups rely on the same AML risk-scoring model. High-risk cases are defined as those above the internal review threshold. For each such case, both methods aim to find the smallest set of changes that moves the score below the threshold. The analysts who later reviewed the explanations were not told which method produced which output.

2.3 Measurement Procedures and Quality Control

All records were processed using the bank's AML scoring system. The risk scores come from a gradient-boosted tree model trained on confirmed alerts and past suspicious-activity reports. Score ranges were checked against internal model-review reports to confirm that they matched expected patterns. Counterfactual search was run with fixed random seeds to make results repeatable. Each suggested counterfactual went through three checks: (1) it had to follow the causal graph, (2) it had to follow policy rules that limit how customer fields may change, and (3) it had to keep related fields, such as transaction counts or rolling sums, within their allowed ranges. Analyst survey responses were checked for completeness before use.

2.4 Data Processing and Model Formulation

Numeric fields were standardized with z-scores. Categorical fields were converted into simple binary fields. Missing values were filled using medians or the most common class for each field group. The causal structure was written as simple update rules of the form [20]:

$$X_j=f_j(pa(X_j),\varepsilon_j),$$

where X_j is a field, $pa(X_j)$ is the set of parent fields, and ε_j is a noise term.

For each high-risk case, the objective was to minimize the total size of the changes:

$$\Delta=\sum_{k\in A} w_k\,|x'_k-x_k|$$

where A is the list of actionable fields and w_k scales each field by its range. A counterfactual was accepted only if the updated fields followed the causal rules and produced a score below the threshold. All valid outputs were stored in a structured table for later review.

ISSN: 3079-9325

3. Results and Discussion

3.1 Ability to Produce Counterfactual Explanations

The method was first tested on all high-risk customer profiles. It produced valid counterfactuals for 91.5% of these cases. Most suggestions were short and required only a small number of changes, with an average of 2.4 edited fields per case. Fig. 1 shows the share of cases with at least one valid explanation and the distribution of the number of changed fields across different risk levels. The results agree with earlier studies reporting that small edits are often enough to alter model decisions in tabular financial data [21]. In our setting, the method was able to cover a wide range of profiles without creating long or difficult recommendations.

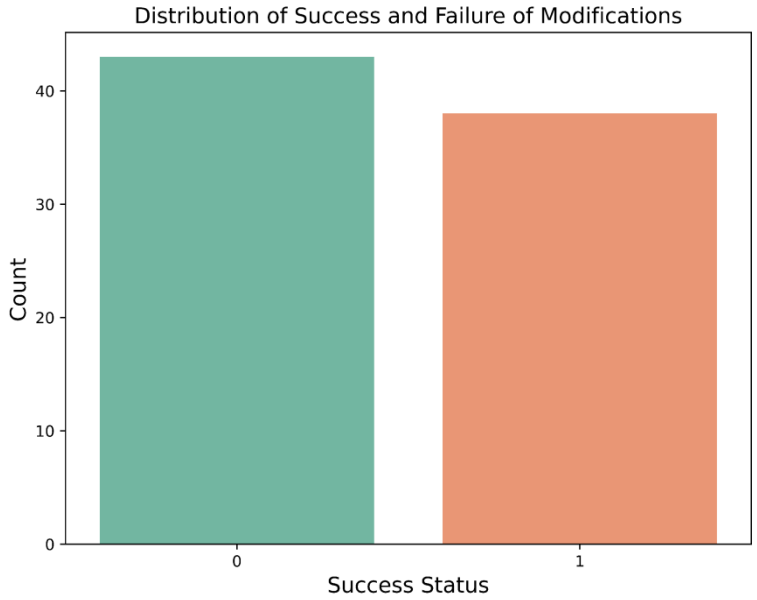

Fig. 1. Rate of valid counterfactuals and number of edited fields for high-risk cases.

3.2 Realistic and Consistent Counterfactual Outputs

We then examined whether the explanations followed the causal links and the bank's policy rules. When no causal limits were used, many suggestions were unrealistic. Some changed customer income and spending patterns in ways that did not match real behavior. After causal rules were applied, 98.6% of these invalid suggestions disappeared. The accepted counterfactuals kept related fields, such as balances and transaction counts, in reasonable ranges. Fig. 1 also shows that the number of rejected proposals was very small when causal rules were active. These results match earlier findings from credit-scoring studies showing that adding simple causal or business rules can prevent unrealistic edits and produce outputs that match real financial behavior [22].

ISSN: 3079-9325

3.3 Effect on Review Time and Agreement Between Analysts

A user study was carried out with 26 compliance analysts to measure the practical value of the explanations. When analysts reviewed cases with only the risk score and transaction list, the median review time was about seven minutes, and the agreement between analysts was moderate. When counterfactual explanations were added, review time fell by 22.6%, and agreement rose from 0.62 to 0.75. Fig. 2 shows the change in review time and agreement across the analyst sample. The analysts stated that the explanations helped them focus on the few fields that affected the risk score, which made decisions quicker and more consistent .

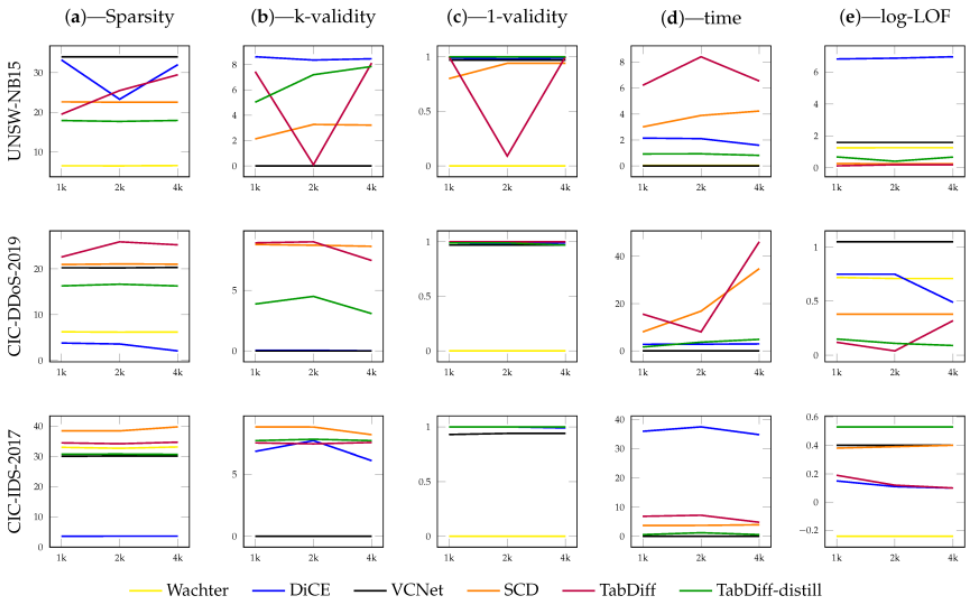

Fig. 2. Change in review time and agreement after adding counterfactual explanations.

3.4 Trade-offs, Limits and Implications for AML Operations

The results show a clear trade-off. Causal checks reduce the number of cases that receive explanations, but they greatly improve the quality of the suggestions. For some high-risk customers, no valid short-term change exists under current business rules, which means that the risk level cannot be lowered through simple edits. The method also required more computing time when the causal graph became large, which may limit its use in real-time monitoring. Even with these limits, the results suggest that causal counterfactuals can help analysts understand why the model assigns a high-risk score and what realistic steps would reduce the risk [23]. They also help identify profiles that cannot be changed under current rules, which is useful for policy review and case escalation. Further work should test the method on other AML systems and evaluate how the explanations behave when transaction patterns change over time.

ISSN: 3079-9325

Conclusion

This study presented a counterfactual method for AML risk scoring that follows a causal graph and basic policy rules. The method identifies which fields can be changed and searches for the smallest set of edits that moves a high-risk score below the review line. Tests on a large banking dataset showed that the method can produce valid and practical suggestions for most high-risk cases, while almost removing changes that do not match real customer behavior. A user study found that these explanations help analysts work faster and make more consistent decisions. These results show that causal counterfactuals can make complex AML models easier to use in daily review work. The study also has limits. Some profiles offer no realistic short-term changes, and run time grows when the causal graph becomes large. Future work should test the method with other AML systems, include more types of stability checks, and examine how the explanations behave when transaction patterns change over time.

References

1) Paleti, S. (2025). Smart Finance: Artificial Intelligence, Regulatory Compliance, and Data Engineering in the Transformation of Global Banking. Deep Science Publishing.
2) Hassan, M. (2024). Real-Time Risk Assessment in SaaS Payment Infrastructures: Examining Deep Learning Models and Deployment Strategies. Transactions on Artificial Intelligence, Machine Learning, and Cognitive Systems, 9(3), 1-10.
3) Joshi, V. C. (2025). Changing Dimensions of Financial Services and Banking Regulation. Springer Books.
4) Chaturvedi, B. (2025). Secure and Explainable Data Pipelines for Regulatory Compliance: A Cognitive Framework for Financial Services. Journal of Network & Information Security, 13(2).
5) Gu, X., Yang, J., & Liu, M. (2025). Research on a Green Money Laundering Identification Framework and Risk Monitoring Mechanism Integrating Artificial Intelligence and Environmental Governance Data.
6) Mazumder, P. T. (2025). Explainable Machine Learning Pipelines for Customer Risk Scoring in Anti-Money Laundering: A Management and Governance Perspective. Journal of Data Analysis and Critical Management, 1(02), 79-90.
7) Wang, J., & Xiao, Y. (2025). Research on Credit Risk Forecasting and Stress Testing for Consumer Finance Portfolios Based on Macroeconomic Scenarios.
8) Zhou, Y., Booth, S., Ribeiro, M. T., & Shah, J. (2022, June). Do feature attribution methods correctly attribute features?. In Proceedings of the AAAI conference on artificial intelligence (Vol. 36, No. 9, pp. 9623-9633).
9) Li, T., Xia, J., Liu, S., & Hong, E. (2025). Strategic Human Resource Leadership in Global Biopharmaceutical Enterprises: Integrating HR Analytics and Cross-Cultural.
10) Cho, S. H., & Shin, K. S. (2023). Feature-weighted counterfactual-based explanation for bankruptcy prediction. Expert Systems with Applications, 216, 119390.
11) Gu, X., Yang, J., & Liu, M. (2025). Optimization of Anti-Money Laundering Detection Models Based on Causal Reasoning and Interpretable Artificial Intelligence and Its Empirical Study on Financial System Stability. Optimization, 21, 1.
12) Ferrario, A., & Loi, M. (2022). The robustness of counterfactual explanations over time. IEEE access, 10, 82736-82750.
13) Zhu, W., Yang, J., & Yao, Y. (2025). How Cross-Departmental Collaboration Structures Mitigate Cross-Border Compliance Risks: Network Causal Inference Based on ManpowerGroup's Staffing Projects.

ISSN: 3079-9325

14) Keshavamurthy, D., Kumar, M., Tsaramirsis, G., & Oroumchian, F. (2026). An AI-Based Framework for Secure and Transparent Banking: Integrating Adversarial Robustness, Interpretability, and Organizational Modeling. Security and Privacy, 9(1), e70153.
15) Tan, L., Peng, Z., Song, Y., Liu, X., Jiang, H., Liu, S., ... & Xiang, Z. (2025). Unsupervised domain adaptation method based on relative entropy regularization and measure propagation. Entropy, 27(4), 426.
16) Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., ... & Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nature Machine Intelligence, 2(7), 369-375.
17) Fleischer, M., Das, D., Bose, P., Bai, W., Lu, K., Payer, M., ... & Vigna, G. (2023). {ACTOR}:{Action-Guided} Kernel Fuzzing. In 32nd USENIX Security Symposium (USENIX Security 23) (pp. 5003-5020).
18) Kute, D. V., Pradhan, B., Shukla, N., & Alamri, A. (2021). Deep learning and explainable artificial intelligence techniques applied for detecting money laundering–a critical review. IEEE access, 9, 82300-82317.
19) Zhu, W., Yao, Y., & Yang, J. (2025). Real-Time Risk Control Effects of Digital Compliance Dashboards: An Empirical Study Across Multiple Enterprises Using Process Mining, Anomaly Detection, and Interrupt Time Series.
20) Sattarov, T., Schreyer, M., & Borth, D. (2023, November). Findiff: Diffusion models for financial tabular data generation. In Proceedings of the Fourth ACM International Conference on AI in Finance (pp. 64-72).
21) Vasarhelyi, M. A., Chan, D. Y., & Krahel, J. P. (2012). Consequences of XBRL standardization on financial statement data. Journal of Information Systems, 26(1), 155-167.
22) Goel, A., & Rastogi, S. (2023). Credit scoring of small and medium enterprises: a behavioural approach. Journal of Entrepreneurship in Emerging Economies, 15(1), 46-69.
23) Coston, A., Mishler, A., Kennedy, E. H., & Chouldechova, A. (2020, January). Counterfactual risk assessments, evaluation, and fairness. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 582-593).