Counterfactual Based Probabilistic Graphs for Explainable Money Laundering Detection

Xingzhe Sun, Dehui Du
Shanghai Key Laboratory of Trustworthy Computing, East China Normal University

Abstract

Anti-money laundering (AML) is a critical challenge for the global financial sector, and deep neural networks have become an essential tool for AML monitoring. However, existing black-box models often lack explainability and fail to provide in-depth analysis of the intent behind behaviors. The method proposed in this paper constructs a Bayesian network for the AML problem by injecting counterfactual examples into the dataset to explain the black-box model through inference. In addition, the method use backward inference to uncover the intent behind anomalous transaction behaviors. Experiments conducted on various AML models and datasets show that our approach provides model-agnostic explanations and can infer the intrinsic intent of money launderers, providing valuable insights for decision-makers.

Introduction

The issue of money laundering has consistently represented a substantial challenge within the financial sectors of countries across the globe. The United Nations defined money laundering in the 1988 Vienna Convention as the process of transferring or transmitting property with the knowledge that it is derived from illegal sources, in order to conceal or disguise its illicit origin, or to assist any person involved in criminal behavior in evading legal consequences.(Al-Suwaidi and Nobanee, 2020) A report by the World Bank indicates that approximately 2 to 5 percent of global GDP (equivalent to approximately $800 billion to $2 trillion) is laundered annually through illicit means. Money laundering is frequently associated with criminal activities such as drug trafficking, terrorism financing, and human trafficking.(Corselli, 2020) The United Nations Office on Drugs and Crime (UNODC) estimates that the global income from the transnational drug trade reaches several hundred billion dollars annually, with these funds often flowing into the formal financial system through money laundering behaviors.

The advent of AI technology has led to the emergence of AI predictors as a key tool for the monitoring of money laundering anomalies.(de Jesús Rocha-Salazar, Segovia-Vargas, and del Mar Camacho-Miñano, 2021) Nevertheless, the reliabilty of these predictors has become a significant financial concern, given the heightened security and privacy concerns surrounding their use. This is the explainability problem faced by many black-box models, or XAI. Because black-box models are often not transparent, even if they perform well, decision makers do not have access to the decision-making mechanisms within the model, which can raise serious safety or fairness issues if the model fails. Like figure 1. One potential solution to the XAI problem is to utilise AI models that are inherently explainable, such as linear regression models or decision trees.However, for tasks that use black-box models, it is necessary to explain the results after the decision is made. Using model-agnostic explanation methods allows for interpreting different models performing the same task. By using counterfactuals and probabilistic graphs, the existing model can be abstracted, allowing for an understanding of the impact of each input on the decision outcome.

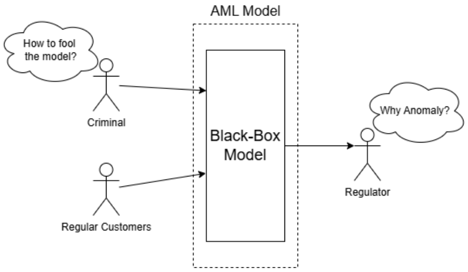

Figure 1: Black-box model is a ”wall”

Another problem in detecting money laundering is that criminals are very few in number compared to the general population, which makes money laundering transactions and accounts very rare and difficult to identify in all the data. Obviously, a large transaction can easily attract the attention of banks, but criminals today have many ways to deceive regulators, such as replacing a single large transaction with many scattered small transactions, which can be checked by pattern matching, like Figure 2 but the large volume and complexity of transaction data make it difficult to try to identify money laundering patterns. By combining domain knowledge and causal reasoning, our method can not only explain the model's decisions but also infer the criminal's money laundering intent from the behavioral data.

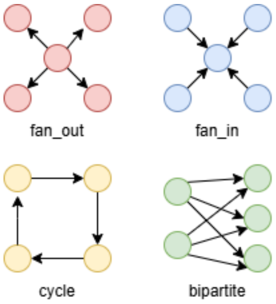

Figure 2: Money laundering patterns

Related Work

There are many machine learning anomaly detection methods for anti-money laundering, as summarized in (Chen et al., 2018), including decision trees, random forests, and support vector machines. Recent studies have increasingly focused on graph-based and neural network approaches(Pourhabibi et al., 2020), such as (Weber et al., 2019), which uses graph neural networks to address Bitcoin money laundering, and (Lo et al., 2023), which employs self-supervised graph neural networks for money laundering analysis. However, these high-performance models are often opaque black-box models that lack explainability.

In the field of explainable artificial intelligence (XAI), models such as decision trees provide inherent explainability. For black-box models, post hoc explanation methods have been proposed, such as (Ribeiro, Singh, and Guestrin, 2016), which suggests using LIME to explain different classification models, and (Lundberg and Lee, 2017), which defines a game-theoretic explanation framework, SHAP. Additionally, (Wachter, Mittelstadt, and Russell, 2017b) introduces the use of counterfactual explanations, which describe the minimal changes required to achieve a desired outcome, a concept that forms one of the theoretical foundations of this paper. The explainability of money laundering models has also garnered attention from researchers. For example, (Konstantinidis and Gegov, 2024) combines deep neural networks (DNNs) with SHAP to improve the transparency of anti-money laundering tasks, while (Li et al., 2024) employs a similar approach, using SHAP for post hoc explanations in anti-money laundering tasks.

Methodology

Overview

In the methodology of this paper, we first process the data. On the one hand, we adapt it to the probabilistic graph through discretization. On the other hand, we select anomalous or near-anomalous data from it, which will be used to generate counterfactual samples with different levels of intervention according to the statute. (Schulam and Saria, 2017)The reason for using counterfactual samples to interact with the model is that starting from the anomalies allows us to obtain the decision boundaries more efficiently and reduces the interference of a large amount of invalid data in PGM. At the same time, the PGM adds knowledge of money laundering strategies based on causal structure and domain knowledge and uses this to infer the intent hidden behind the transaction behavior.

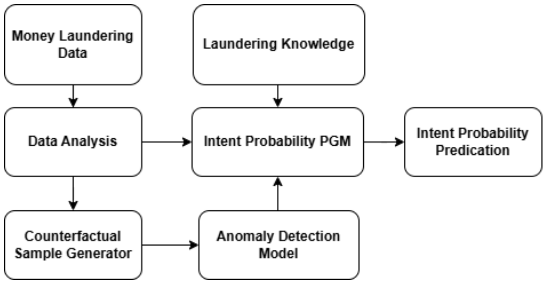

Figure 3: Counterfactual based probabilistic graphs for explainable money laundering detection

Counterfactual

The main question of Counterfactual is: ”How would the model’s prediction change if certain features of the input were different?” (Wachter, Mittelstadt, and Russell, 2017a)Counterfactual explanations are part of causal inference methods,We define the counterfactual as follow:
Given a trained black-box model $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space, let $x_0 \in \mathcal{X}$ be a specific input instance, and $y_0 = f(x_0)$ be the model’s prediction for that input. The goal of counterfactual generation is to find a new input x^* such that:

$$f(x^*) \neq f(x_0) \quad \text{and} \quad d(x^*, x_0) \text{ is minimized}$$

Where $d(x', x_0)$ is the distance between the original instance x_0 and the counterfactual instance x'.The primary objective of counterfactual generation is to minimize the difference $d(x_0, x^*)$ between the original instance x_0 and the generated counterfactual instance x^*, while ensuring that the model’s prediction changes, i.e., $f(x^*) \neq f(x_0)$. This can be formulated as the following optimization problem:

$$x^* = \arg\min_{x' \in X} \left(d(x', x_0) + \lambda \cdot \mathbb{I}(f(x') \neq f(x_0)) \right)$$

Where:$d(x', x_0)$ is the distance between the original instance x_0 and the counterfactual instance x'.

- λ is a regularization parameter that controls the trade-off between the distance and the prediction change.
- $\mathbb{I}(f(x') \neq f(x_0))$ is an indicator function that is 1 if $f(x') \neq f(x_0)$, ensuring that the counterfactual instance leads to a different model prediction.

Consider a money laundering detection model $f(x)$, where x_0 represents a customer’s account behavior data. The model predicts whether the customer is involved in money laundering. If the model predicts that a customer x_0 is involved in money laundering (i.e., $f(x_0) = 1$), we might want to generate a counterfactual instance x^* such that $f(x^*) = 0$ (i.e., the model predicts the customer is not involved in money laundering), while minimizing the difference between x^* and x_0. The generated counterfactual instance x^* helps the PGM understand which account behaviors are critical in causing the model to change its prediction.

Probabilistic Graphical Models

Probabilistic Graphical Models (PGMs) are models that represent and reason about variables and their dependencies or causal relationships using a graph structure. PGMs include both directed and undirected graphs, where directed graphs can express causal relationships, while undirected graphs only represent dependencies. In cases where the structure is not well-defined, generating a PGM requires structural learning. In the task of anti-money laundering, we combine expert knowledge and data classification to construct a directed acyclic Bayesian network(Heckerman, Geiger, and Chickering, 1995), which serves as the structure for our probabilistic graphical model.

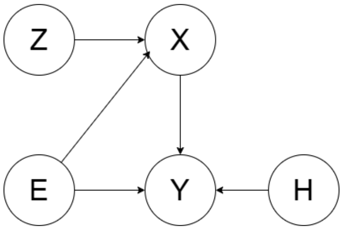

Figure 4: Causal model of money laundering

In this graph structure, X represents the user's transaction behavior, and Y represents the money laundering detection outcome. There exists a direct causal relationship between these two variables. H represents the account's history of violations and warnings, indicating the risk level of the account based on its past activities, which serves as a reference for the money laundering detection outcome. E denotes the transaction environment, such as the currency and the and countries involved in transactions. Different transaction environments can influence the detection results, as varying levels of regulatory control in different regions can affect user transaction patterns. Z represents the evasion intent, which refers to whether the trader intentionally seeks to engage in money laundering while avoiding regulatory oversight through their transaction behavior. The intent is typically not observable in the data, but through the counterfactual sample generation method mentioned earlier, we can infer it. Since "avoiding regulation" aligns with the goal of generating counterfactual examples, it is reasonable to treat the counterfactual samples as "money laundering activities that successfully evade regulation." This approach distinguishes counterfactual data from regular transactions, providing a meaningful way to capture the intent.

In the case of a known structure, the parameters of the probabilistic graphical model are learned, which involves obtaining the conditional probability distributions (CPD). Since we have discretized data such as total transaction amounts (e.g., using transaction volume binning), the initial distribution of the conditional probability table can be directly derived using the frequency counting method, i.e.,$P(X|Pa(X)) = \frac{\text{Count}(X \cap Pa(X))}{\text{Count}(Pa(X))}$. However, for continuous data that has not been discretized or when there is insufficient data, maximum likelihood estimation (MLE) or Bayesian estimation should be used for computation, though these are not discussed further here.After constructing the probabilistic graphical model, inference can be performed based on the CPD and domain knowledge. In forward inference, the probability distribution of the result nodes can be directly observed. For nodes requiring backward inference, Bayesian inference is applied, i.e.,$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$. This allows the inference of the confidence of parent nodes using the observable data. In the structure proposed in this paper, this path is used to compute the intent.

Experiments and Results

In the absence of an intuitive evaluation criterion for the XAI problem, the experiments presented in this paper combine the project with the selection of multiple datasets and classification models to demonstrate the validity of the method and the model-agnostict ability. Classification experiments are also conducted based on synthetic data to verify the ability to uncover ML intent.

Dataset and model

The data used in this study comes from two open-source AML task datasets and two confidential datasets provided by a partner bank. A significant portion of the data is derived from transactions. Therefore, the first step is to extract account-level features from these transactions. For each specific account, we collect its warning records as historical risk indicators, transaction currency and bank as environmental influences, and transaction frequency, average transaction amount, and maximum transaction value as transaction details. Additionally, to construct the probabilistic graph, the data is discretized (e.g., classifying transaction amounts based on their magnitude). For intent inference, labels are derived from expert-verified anomalous accounts and counterfactual samples.

Table 1: Anti-money laundering dataset

Dataset	Transactions	Laundering	Rate
IBM-AML	5M	5.1k	0.1%
SAML-D	9M	9.8K	0.1%
BANK-sim	251k	15k	6.32%
BANK-real	9.6k	97	1%

Two of the selected datasets are publicly accessible, namely IBM-AML(Altman et al., 2023) and SAML-D(Oztas et al., 2023). The remaining two BANK datasets are proprietary and have been provided by the project's partner banks. The money laundering rate indicates that the AML problem is a classical sample imbalance problem. Even when a targeted selection is made during the generation of the dataset, the frequency of money laundering anomalies is significantly lower than that of common classification and anomaly detection problems. Consequently, further screening is necessary when explaining the model to obtain meaningful results and avoid ineffective learning time.

The models selected for the experiment were provided by the project's partner bank. These models have been pre-trained on a variety of datasets and have demonstrated robust performance in anti-money laundering tasks. The mean test results for the two models across the four datasets are presented in the table. It is important to acknowledge that, due to the class imbalance inherent in the money laundering task, greater emphasis should be placed on recall metrics when evaluating model performance. This is because the core task of anti-money laundering is to identify all anomalous samples, rather than to achieve high accuracy in classifying a large number of normal samples. The two models used in the experiment are different black-box models, and this experiment aims to illustrate that our method is model-agnostic.

Table 2: Anti-money laundering model

Dataset	Precision	Recall	F1 Score	ACC
MODEL-1	90%	94%	92.0%	98%
MODEL-2	90.5%	95%	92.7%	99%

Explainability

In the experiment, we injected counterfactual samples into the dataset, using 80% of the data for training, 10% for validation, and 10% for model evaluation. The obtained classification results are presented below. To demonstrate the Explainability of the model's decisions, the classification results here are compared to those of the prediction model, without considering the actual labels of the samples.

Table 3: The confidence of black-box model

	IBM-AML	SAML-D	BANK-sim	BANK-real
MODEL1	0.6143	0.6720	0.7053	0.7497
MODEL2	0.6265	0.6609	0.7275	0.7608

The experimental results in the table 3 demonstrate that our method can provide explanations for the prediction outcomes of different models across various datasets. This capability allows our method to offer interpretable insights into the model's predictions, aiding decision-makers. However, due to the limited number and dimensions of input features, achieving a 100% explanation is challenging, which is one of the common issues faced in the field of Explainable AI (XAI).

Intent Prediction

To test the method's ability to infer the money laundering evasion intent, we used a mixture of normal samples and counterfactual samples (i.e., samples that successfully evade detection). These samples were classified as normal by the anti-money laundering model. A standard threshold was defined, and by comparing the threshold with the probability inferred from the behavior, the output was converted into a binary classification. The figure5 shows the confusion matrix of the classifier's test results. It can be observed that even when the transaction behavior was modified to successfully deceive the detection model, the proposed method was still able to identify the underlying intent from the behavior.

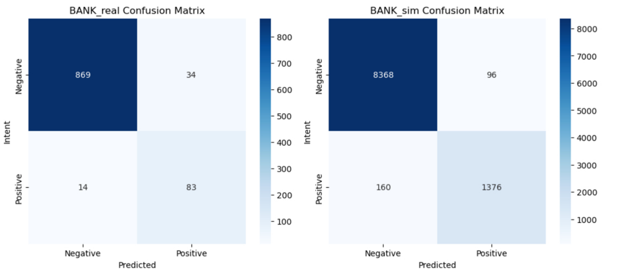

Figure 5: Intent prediction result

Limitations

This proposed model also has several limitations:Unlike the SHAP method, which can provide a unique explanation, the approach proposed in this paper provides explanations based on a probabilistic graphical model structure. Depending on the underlying structure and the method used to generate counterfactuals, different explanations can be derived. Furthermore, due to the nature of the probabilistic graphical model, it can only model a small subset of the extracted features. When faced with a large number of input features, computational difficulties arise as the problem is NP-hard(Chickering, Heckerman, and Meek, 2004).

Conclusion and Future Work

Recent studies have demonstrated the increasing variety of money laundering methods, with strategies now capable of evading AI-driven regulation. The method proposed in this paper can explain existing anti-money laundering (AML) black-box models and analyze their prediction results. With the support of causal inference and domain knowledge, it further infers criminal intent. Our method has been tested on multiple datasets and AML models, providing effective explanations for these models. Additionally, by reasoning intent, it can distinguish between intentional money laundering evasion strategies and unintentional behavior by ordinary users. Future work includes: (1) constructing an anti-money laundering knowledge graph by integrating domain knowledge, enabling more detailed classification and analysis of laundering behaviors; (2) expanding transaction and account features to uncover high-dimensional risk causal relationships.

Acknowledgments

This work was supported in part by National Key R&D Program of China (No. 2022ZD0120302).

References

Al-Suwaidi, N.; and Nobanee, H. 2020. Anti-money laundering and anti-terrorism financing: a survey of the existing literature and a future research agenda. *Journal of Money Laundering Control*, ahead-of-print.

Altman, E. R.; Blanusa, J.; von Niederhäusern, L.; Egressy, B.; Anghel, A.; and Atasu, K. 2023. Realistic Synthetic Financial Transactions for Anti-Money Laundering Models. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Chen, Z.; Khoa, L. D. V.; Teoh, E. N.; Nazir, A.; Karuppiah, E. K.; and Lam, K. S. 2018. Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review. *Knowl. Inf. Syst.*, 57(2): 245–285.

Chickering, M.; Heckerman, D.; and Meek, C. 2004. Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5: 1287–1330.

Corselli, L. 2020. Italy: money transfer, money laundering and intermediary liability. *Journal of Financial Crime*.

de Jesús Rocha-Salazar, J.; Segovia-Vargas, M.-J.; and del Mar Camacho-Miñano, M. 2021. Money laundering and terrorism financing detection using neural networks and an abnormality indicator. *Expert Systems with Applications*, 169: 114470.

Heckerman, D.; Geiger, D.; and Chickering, D. M. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20: 197–243.

Konstantinidis, G.; and Gegov, A. 2024. Deep Neural Networks for Anti Money Laundering Using Explainable Artificial Intelligence. In *2024 IEEE 12th International Conference on Intelligent Systems (IS)*, 1–6.

Li, P.-Y.; Chang, T.-T.; Kuo, Y.-C.; Lin, C.-Y.; and Chang, H.-Y. 2024. Unveiling the Black Box: An XAI-based Anti-Money Laundering Model. In *2024 International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, 293–294.

Lo, W. W.; Kulatilleke, G. K.; Sarhan, M.; Layeghy, S.; and Portmann, M. 2023. Inspection-L: self-supervised GNN node embeddings for money laundering detection in bitcoin. *Appl. Intell.*, 53(16): 19406–19417.

Lundberg, S. M.; and Lee, S. 2017. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 4765–4774.

Oztas, B.; Cetinkaya, D.; Adedoyin, F.; Budka, M.; Dogan, H.; and Aksu, G. 2023. Enhancing Anti-Money Laundering: Development of a Synthetic Transaction Monitoring Dataset. In *2023 IEEE International Conference on e-Business Engineering (ICEBE)*, 47–54.

Pourhabibi, T.; Ong, K.-L.; Kam, B. H.; and Boo, Y. L. 2020. Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, 133: 113303.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 1135–1144. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342322.

Schulam, P.; and Saria, S. 2017. Reliable decision support using counterfactual models. *Advances in neural information processing systems*, 30.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017a. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31: 841.

Wachter, S.; Mittelstadt, B. D.; and Russell, C. 2017b. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *CoRR*, abs/1711.00399.

Weber, M.; Domeniconi, G.; Chen, J.; Weidele, D. K. I.; Bellei, C.; Robinson, T.; and Leiserson, C. E. 2019. Anti-Money Laundering in Bitcoin: Experimenting with Graph Convolutional Networks for Financial Forensics. *CoRR*, abs/1908.02591.