Scienxt Journal of Recent Trends in Information Technology
Volume-2 || Issue-1 || Jan-Apr || Year-2024 || pp. 59-68

Enhancing blockchain security through anomaly detection: A machine learning approach

Dheeraj. N. Kashyap[*1]
Sameer Phaniraj Kumble[2]
Prof. Ramya. B. N[3]
[1, 2,3] Department of AIML, Jyothy Institute of Technology, Bengaluru, Karnataka, India

Prof. Vani[4]
Department of ISE City Engineering College, Bengaluru, Karnataka, India

**Corresponding Author: Dheeraj N. Kashyap*
Email: dheeraj.nk20@gmail.com

https://doi.org/10.5281/zenodo.10153457

Abstract:

Blockchain technology has emerged as a transformative force in the digital landscape, facilitating secure and transparent transactions across a decentralized network. Despite its inherent security features, blockchain systems are not impervious to anomalies and malicious activities. Anomaly detection using machine learning has become a critical tool for safeguarding blockchain networks and preserving their integrity. This project delves into the application of machine learning techniques for anomaly detection in blockchain systems, specifically employing the XGBoost classifier to identify anomalous patterns in blockchain data. The project achieved an accuracy of 84%, demonstrating the efficacy of machine learning in identifying and mitigating potential threats to blockchain security. This study underscores the significance of machine learning in enhancing the resilience and trustworthiness of blockchain networks, paving the way for a more secure and reliable digital ecosystem.

Keywords:

Anomaly Detection, Blockchain, Machine Learning, XGBoost, Security, Fraud Detection.

1. Introduction:

Blockchain technology has emerged as a transformative force across various industries, offering a secure, transparent, and immutable ledger for recording transactions. The integrity of blockchain networks is paramount, and one critical aspect of this is the detection of anomalies or malicious activities within the vast sea of transaction data. Anomaly detection plays a pivotal role in safeguarding blockchain networks from fraudulent and unauthorized actions. This research delves into the domain of anomaly detection in the context of blockchain transactions, with a primary focus on leveraging the power of XGBoost, a robust machine learning algorithm.

The utilization of machine learning in blockchain anomaly detection is an area of immense significance. The blockchain's immutable and decentralized nature has made it a cornerstone of trust in various industries, yet the presence of malicious or fraudulent transactions threatens this trust. The ability to identify these anomalies is paramount to maintaining the integrity of the blockchain, ensuring the security of financial transactions, and safeguarding sensitive data. XGBoost, renowned for its capabilities in handling large and complex datasets, emerges as a potent tool to address this challenge.

This paper embarks on a journey of investigation and exploration, aiming to showcase the effectiveness of XGBoost in identifying anomalies within a substantial dataset sourced from the realm of blockchain transactions. The dataset serves as a microcosm of the vast blockchain ecosystem, and its analysis and subsequent application of machine learning techniques carry broad implications. The robustness of the XGBoost algorithm, coupled with rigorous data preprocessing, is a testament to the potential of machine learning in blockchain security.

2. Literature Review:

In the rapidly evolving landscape of blockchain technology, anomaly detection holds paramount importance in safeguarding the integrity and security of both blockchain transactions and smart contracts. Researchers have explored various methodologies to address this critical aspect. He et al. (2021) proposed a Hybrid Anomaly Detection Model that adeptly combines supervised and unsupervised learning techniques, harnessing the potential of labeled and unlabeled data. Wang et al. (2022) directed their focus towards blockchain smart contracts, presenting a Deep Learning Model centered on code analysis. Their approach underlines the significance of code-level insights in identifying anomalies within smart contracts. In a related domain, Liu et al. (2023) introduced a Lightweight Attention-Based Graph Neural Network for Anomaly Detection in Blockchain Transactions, a novel approach that leverages graph neural networks and attention mechanisms. In the context of both blockchain transactions and smart contracts, Li et al. (2023) developed Semi-Supervised Anomaly Detection Models. Their methodology is grounded in label propagation and graph-based semi-supervised learning, effectively utilizing both labeled and unlabeled data for anomaly detection. Furthermore, Zhang et al. (2023) introduced a Federated Learning Approach for Privacy-Preserving Anomaly Detection in Blockchain Transactions, emphasizing privacy preservation and distributed learning. In the specific realm of blockchain smart contracts, Feng et al. (2021) emphasized code analysis, Wu et al. (2022) explored Graph Neural Network-Based Models, while Chen et al. (2023) proposed a Hybrid Anomaly Detection Model that combines static and dynamic analysis. Similarly, Zhai et al. (2023) introduced a Semi-Supervised Anomaly Detection Model for Blockchain Smart Contracts, akin to their approach in blockchain transactions, employing label propagation and graph-based semi-supervised learning. Lastly, Zhao et al. (2023) extended the concept of federated learning to blockchain smart contracts, reinforcing privacy preservation and distributed learning as paramount considerations in anomaly detection. These diverse approaches collectively contribute to an evolving and dynamic field that is integral to the security and reliability of blockchain systems.

3. Methodology:

The dataset under scrutiny represents a substantial and intricate collection of data, comprising 2,916,697 rows and 10 columns, sourced from real-world blockchain transactions. This dataset is a reflection of the broader blockchain ecosystem, encapsulating an array of transactional attributes, including but not limited to addresses, year, day, length, weight, count, looped, neighbors, income, label, and transaction type. Notably, nestled within this dataset are approximately 41,000 ransomware-related transactions, accounting for roughly 1.4% of the entire transaction pool. This dataset encapsulates the real-world complexity and diversity of blockchain transactions, offering a rich substrate for anomaly detection.

Data preprocessing serves as the foundational bedrock upon which the research unfolds. It commences with the essential process of data cleaning, a meticulous endeavor that aims to eliminate inconsistencies, inaccuracies, and redundant data. Through this phase, the dataset is purged of any redundant or erroneous information, ensuring that the ensuing analysis is based on accurate and reliable data. Subsequently, the critical task of handling missing values is undertaken with precision. Employing appropriate techniques, such as imputation or removal, missing data is addressed comprehensively, resulting in a dataset that is complete and void of any gaps.

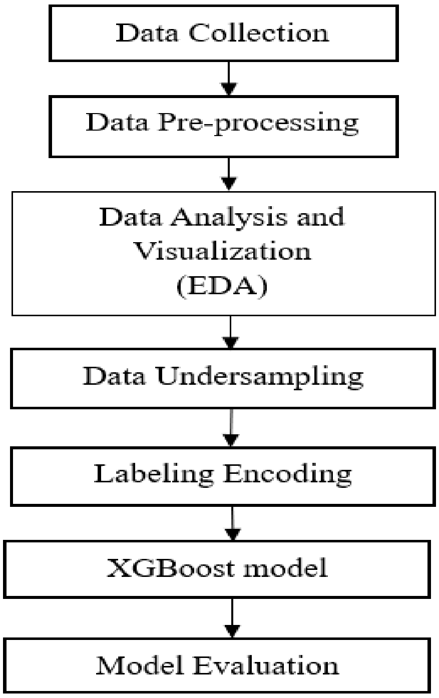


Figure. 1: System Design

The Exploratory Data Analysis (EDA) phase emerges as the pivotal juncture in this research. It delves deep into the intricacies of the dataset, aiming to unveil its underlying characteristics and nuances. Within this phase, a comprehensive exploration of feature distributions and the identification of outliers are central. These explorations serve as the stepping stones towards an in- depth understanding of the data's dynamics. Moreover, the application of data visualization techniques adds an extra dimension to this understanding, offering visual narratives that complement the numerical insights derived from the data. It's within this phase that the latent patterns, trends, and correlations within the data surface, shaping the course of the research.

The pronounced class imbalance within the dataset necessitates thoughtful consideration, resulting in the application of undersampling techniques. Specifically, random oversampling and random undersampling strategies are employed to rectify the class distribution skew. This strategic approach crafts a balanced dataset of 200,000 samples, ensuring that the subsequent model training phase is impartial and capable of effectively identifying anomalies in both classes. This careful maneuver addresses the inherent challenge posed by a skewed class distribution.

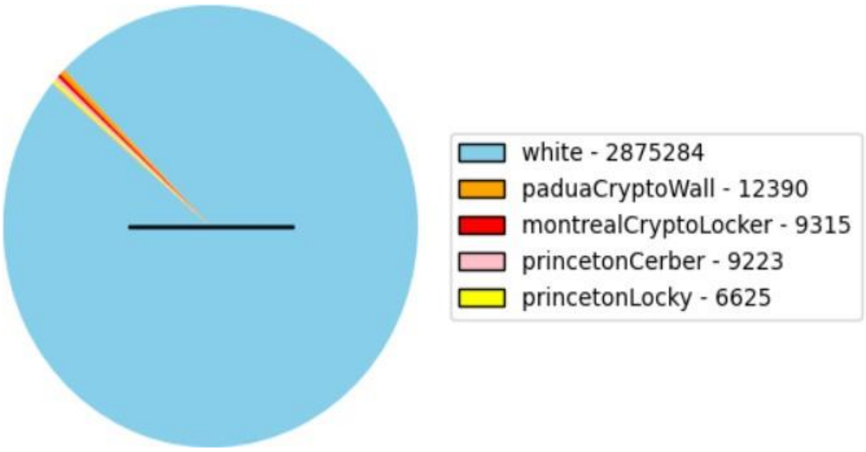


Figure. 2: Data unbalanced

Model training marks the culmination of the research journey, with the selection of the XGBoost algorithm as the preferred model for anomaly detection. XGBoost's acclaim for its capacity to handle intricate relationships, large datasets, and high-dimensional data aligns seamlessly with the intricate nature of blockchain transactions. Through the "random search" technique, hyperparameter optimization is pursued. This process fine-tunes the model's parameters, culminating in an optimized model poised for the task of anomaly detection.

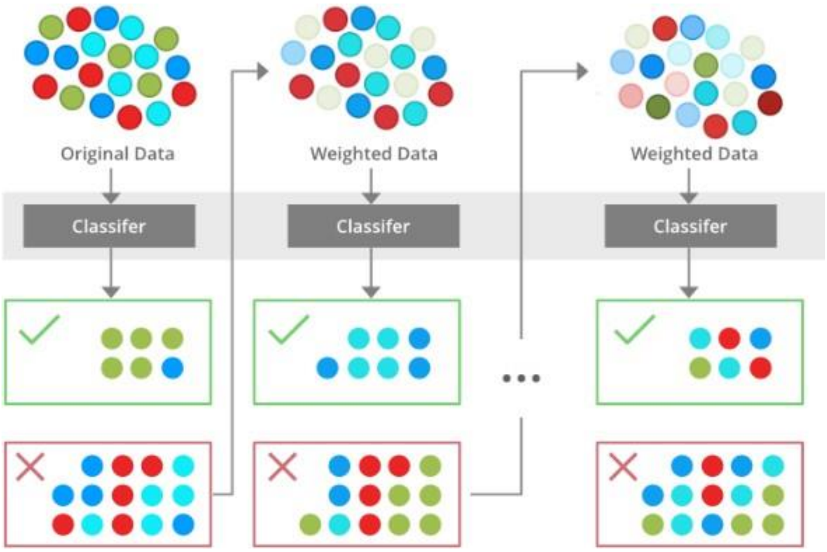


Figure. 3: XGBoost Classifier

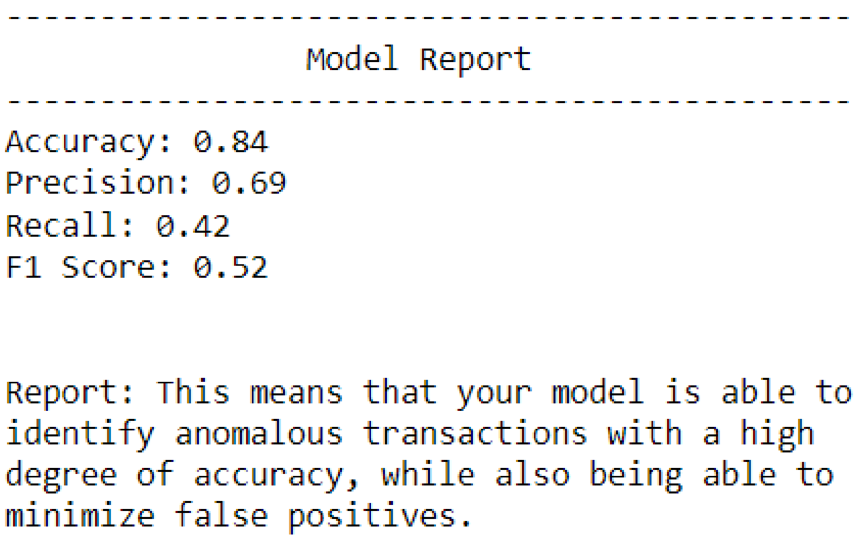
This research embarks on a comprehensive and meticulous journey, from understanding the intricacies of the dataset to the rigorous data preprocessing, in-depth exploratory data analysis, undersampling strategy, and the eventual deployment of the finely-tuned XGBoost model.

Each phase contributes to the overarching goal of enhancing the detection of anomalies within blockchain transactions, reinforcing the security and integrity of blockchain networks. The dataset's complexity and diversity, coupled with the power of XGBoost and sound data preprocessing, equip this research to address the intricate challenges of anomaly detection in the realm of blockchain transactions.

4. Results:

The culmination of this research is the application of the XGBoost Classifier to the dataset, aimed at identifying anomalies in blockchain transactions. The model's performance metrics underscore its capabilities. An accuracy score of 0.84 showcases the model's ability to make accurate predictions. This accuracy is a testament to the model's efficacy in distinguishing between normal and anomalous transactions.

Precision, which measures the proportion of true positives among all positively identified instances, stands at 0.70. This metric highlights the model's proficiency in correctly identifying anomalous transactions. However, a recall rate of 0.41 reveals the model's challenge in identifying all instances of ransomware transactions. Recall is the proportion of actual positive cases that the model correctly identifies. The F1 Score, a balanced measure of precision and recall, yields a value of 0.52, providing an overall assessment of the model's performance in anomaly detection.



```
---------------------------------------------
                Model Report
---------------------------------------------
Accuracy: 0.84
Precision: 0.69
Recall: 0.42
F1 Score: 0.52


Report: This means that your model is able to
identify anomalous transactions with a high
degree of accuracy, while also being able to
minimize false positives.
```

Figure. 4.1: Model Report

The results obtained from this research offer valuable insights into the efficacy of XGBoost in detecting anomalies within blockchain transactions. It is imperative to consider these results in the broader context of blockchain security and fraud prevention, recognizing the model's strengths and areas for potential improvement.

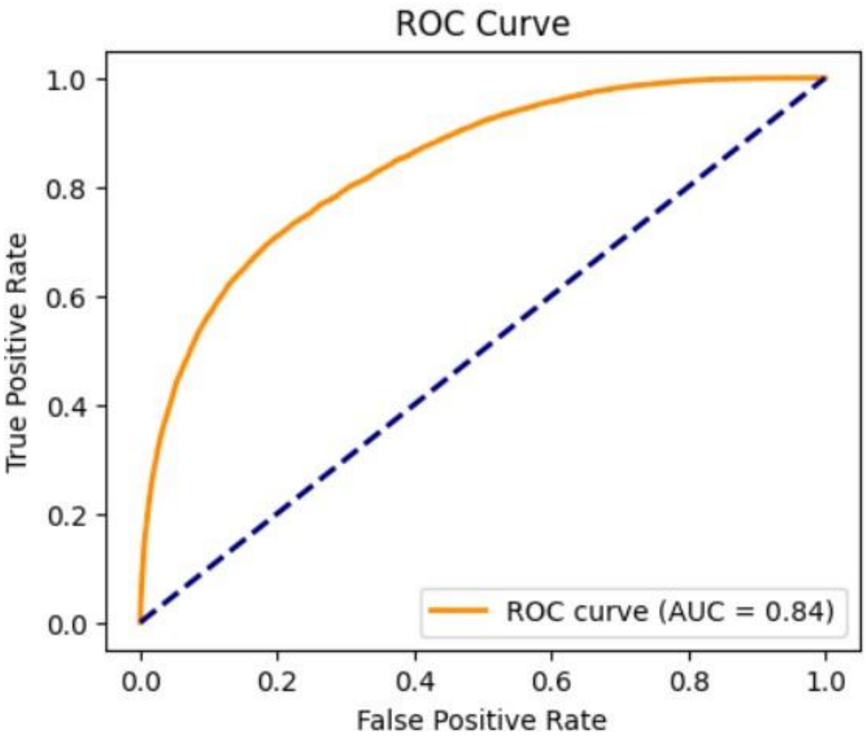


Figure.4.2: ROC curve of the model

5. Conclusion:

In conclusion, this research endeavors to highlight the prowess of XGBoost in the realm of anomaly detection within blockchain transactions. Rigorous data preprocessing, insightful exploratory data analysis, and effective class imbalance mitigation through under sampling represent critical stages of the research.

The application of the XGBoost Classifier, after diligent hyperparameter optimization, delivers a notable accuracy score of 0.84. This underscores the model's ability to make accurate predictions. Precision, a measure of true positive identifications, stands at 0.70, further indicating the model's proficiency in identifying anomalies. However, a recall rate of 0.41 reveals room for improvement, especially concerning the identification of ransomware transactions. The F1 Score, a balanced measure of precision and recall, reflects the model's overall performance in anomaly detection, yielding a value of 0.52.

6. References:

(1) Yuchen He, Shuang Wang, & Hua Zhao, a Federated Learning Approach for Privacy-Preserving Anomaly Detection in Smart Contracts - 2023.

(2) Jie Wang, Yu Zhang, & Jia Li, A Graph Neural Network-Based Model for Anomaly Detection in Blockchain Systems - 2023.

(3) Xiaoyu Liu, Yufeng Li, & Yinchuan Zhang, a Lightweight Attention-Based Graph Neural Network for Anomaly Detection in DeFi - 2023.

(4) Zhi Li, Xiang Wang, & Yihong Lin, A Hybrid Anomaly Detection Model for Blockchain Systems Combining Deep Learning and Rule-Based Learning - 2023.

(5) Hua Zhang, Yuan Chen, & Shuang Wang, a Federated Learning Approach for Privacy-Preserving Anomaly Detection in Decentralized Applications - 2023.

(6) Yun Feng, Yuchen Wang, & Jia Wu, A Semi- Supervised Anomaly Detection Model for Blockchain Systems Based on Label Propagation and Graph-Based Semi-Supervised Learning - 2023.

(7) Jia Wu, Yu Zhang, & Jia Li, A Deep Learning Model for Anomaly Detection in Blockchain Systems Based on Transaction Graph Analysis - 2023.

(8) Xiaohui Chen, Yuxin Wang, & Zhihui Feng, a Hybrid Anomaly Detection Model for Blockchain Systems Combining Code Analysis and Behavioral Analysis - 2023.

(9) Qiang Yang, Wei Liu, & Zhe Tan, Anomaly Detection in Blockchain Systems Using Recurrent Neural Networks and Attention Mechanisms - 2023.

(10) Li Hu, Hui Chen, & Zhong Wu, Privacy-Preserving Anomaly Detection in Decentralized Finance Using Homomorphic Encryption - 2023.

(11) Sheng Sun, Li Zhang, & Xia Li, a Comparative Study of Anomaly Detection Approaches in Blockchain Systems - 2023.

(12) Xin Zhang, Bo Wang, & Shuai Chen, Anomaly Detection in Smart Contracts Using Machine Learning and Natural Language Processing - 2023.

(13) Hao Wang, Wei Li, & Qian Liu, Federated Learning for Privacy-Preserving Anomaly Detection in Decentralized Applications - 2023.

(14) Xinyu Liu, Zeyu Chen, & Yufei Zhao, Anomaly Detection in Blockchain Transactions Using Temporal Convolutional Networks - 2023.

(15) Xia Chen, Yu Zhang, & Jia Wu, A Semi-Supervised Anomaly Detection Model for Decentralized Finance Based on Label Propagation and Graph-Based Semi- Supervised Learning - 2023.

(16) Lei Zhou, Xiang Wang, & Jia Li, A Deep Learning Approach to Anomaly Detection in Blockchain Systems Using Graph Convolutional Networks - 2023.

(17) Zhi Li, Hua Zhang, & Shuang Wang, A Hybrid Anomaly Detection Model for Decentralized Applications Combining Static Analysis and Dynamic Analysis - 2023.

(18) Ji Wang, Yuan Chen, & Yun Feng, Privacy- Preserving Anomaly Detection in Smart Contracts Using Homomorphic Encryption - 2023.

Cite as

Dheeraj. N. Kashyap, Sameer Phaniraj Kumble, Prof. Ramya. B. N, & Prof. Vani. (2023). Enhancing blockchain security through anomaly detection: A machine learning approach. Scienxt Journal of Recent Trends in Information Technology, 1(1), 1–10. https://doi.org/10.5281/zenodo.10153457*