Jurnal Informatika Ekonomi Bisnis

http://www.infeb.org

2024 Vol. 6 Iss. 2 Hal: 308-312 e-ISSN: 2714-8491

Anomaly Detection in Blockchain Transactions: A Machine Learning Approach within the Open Metaverse

Gregorius Airlangga[1]✉

[1]Atma Jaya Catholic University of Indonesia

gregorius.airlangga@atmajaya.ac.id

Abstract

This study investigates the application of machine learning models for anomaly detection and fraud analysis in blockchain transactions within the Open Metaverse, amid the growing complexity of digital transactions in virtual spaces. Utilizing a dataset of 78,600 transactions that reflect a broad spectrum of user behaviors and transaction types, we evaluated the efficacy of several predictive models, including RandomForest, LinearRegression, SVR, DecisionTree, KNeighbors, GradientBoosting, AdaBoost, Bagging, XGB, and LightGBM, based on their Mean Cross-Validation Mean Squared Error (Mean CV MSE). Our analysis revealed that ensemble methods, particularly RandomForest and Bagging, demonstrated superior performance with Mean CV MSEs of -0.00445 and -0.00415, respectively, thereby highlighting their robustness in the complex transaction dataset. In contrast, LinearRegression and SVR were among the least effective, with Mean CV MSEs of -224.67 and -468.57, indicating a potential misalignment with the dataset's characteristics. This research underlines the importance of selecting appropriate machine learning strategies in the context of blockchain transactions within the Open Metaverse, showcasing the need for advanced, adaptable approaches. The findings contribute significantly to the financial technology field, particularly in enhancing security and integrity within virtual economic systems, and advocate for a nuanced approach to anomaly detection and fraud analysis in blockchain environments.

Keywords: Blockchain Transactions, Anomaly Detection, Machine Learning, Open Metaverse, Fraud Analysis.

1. Introduction

The advent of the Open Metaverse heralds a new era in the digital domain, signaling the convergence of virtual reality, blockchain technology, and digital economies into a singular, expansive universe [1], [2], [3]. This digital frontier, characterized by its decentralized architecture and interoperable capabilities, promises unparalleled opportunities for innovation, social interaction, and economic transactions [4], [5], [6]. However, the very features that make the metaverse a bastion of potential also introduce significant challenges in ensuring secure, transparent, and equitable interactions among its inhabitants [7], [8], [9].

As blockchain technology underpins most transactions within the metaverse, ensuring the integrity and security of these transactions is paramount [10], [11], [12]. The literature is replete with studies focusing on blockchain's technical robustness, scalability, and security measures [13], [14], [15]. Yet, as the metaverse evolves, it becomes increasingly apparent that traditional blockchain security mechanisms are insufficient to address the nuanced and sophisticated threats emerging within this virtual ecosystem [16], [17], [18].

Recent research has increasingly turned its attention towards the application of machine learning and artificial intelligence in detecting anomalies and preventing fraud within digital transactions [19]. These studies highlight the potential of data-driven approaches in identifying patterns indicative of fraudulent activity, thereby mitigating risks, and enhancing user trust [20]. However, the literature also underscores a critical gap: the lack of comprehensive datasets that reflect the complexity and diversity of metaverse transactions, which is essential for training and validating effective anomaly detection models [21].

The urgency of developing robust anomaly detection and fraud analysis tools cannot be overstated [22]. With the metaverse poised to become a significant part of our digital future, the stakes for ensuring its security are high [23]. The proliferation of cyber threats, coupled with the increasing sophistication of fraudulent schemes, necessitates a proactive and adaptive approach to safeguarding transactions [24]. This is where the current research endeavors to make its mark.

Building upon the foundational work in blockchain security and fraud detection, this study introduces a novel dataset encompassing a wide array of transaction types, user behaviors, and risk profiles within the Open Metaverse [25]. Unlike previous studies, which often rely on limited or simulated data, this research employs a dataset of 78,600 records, meticulously crafted to mirror the complexity and dynamism of real-world metaverse transactions.

The goal of this research extends beyond the mere analysis of transaction patterns. It aims to advance the state of the art in anomaly detection and fraud analysis

Accepted: 04-04-2024 | Revision: 15-04-2024 | Publication: 30-06-2024 | doi: 10.37034/infeb.v6i2.864

by leveraging this comprehensive dataset to test and compare the efficacy of various machine learning models. In doing so, this study not only contributes to the theoretical understanding of metaverse transaction dynamics but also offers practical insights for developers, policymakers, and users within the digital economy.

The contribution of this research is threefold. Firstly, it enriches the academic discourse on blockchain transactions within the metaverse, offering empirical evidence and nuanced analysis that was previously lacking. Secondly, by evaluating the performance of different machine learning models, it provides a benchmark for future research and development in anomaly detection and fraud prevention technologies. Lastly, the identification of limitations and challenges within the current frameworks paves the way for future innovations and improvements.

Following this introduction, the article will proceed with an extensive methodology section, it will detail the processes involved in data collection, model selection, and analysis, ensuring transparency and replicability of the research findings. The results will be meticulously discussed, offering insights into their implications for both theory and practice. Finally, the conclusion will encapsulate the study's contributions and outline avenues for future research, underscoring the ongoing journey towards securing the Open Metaverse.

2. Research Method

This study adopts a quantitative research design, leveraging a combination of experimental and analytical methods to examine blockchain transactions within the Open Metaverse. The primary focus is on identifying patterns, anomalies, and risks associated with these transactions through machine learning models. This design enables a systematic investigation of the dataset, facilitating the exploration of relationships between different variables and the predictive performance of various models regarding anomaly detection and fraud analysis.

2.1. Population and Samples

The population of interest in this study encompasses blockchain transactions within the Open Metaverse, reflecting a diverse array of transaction types, user behaviors, and risk profiles. The sample for this research is derived from a comprehensive dataset consisting of 78,600 records. Each record represents a unique transaction, including attributes such as transaction type, location region, amount, and risk score. This dataset was meticulously compiled to ensure a representative cross-section of the broader population, allowing for generalizable insights into transaction dynamics within the metaverse environment.

2.2. Instruments

The primary instrument for this study is the dataset itself, described in the provided dataset overview. The dataset includes various attributes relevant to metaverse transactions, such as Timestamp, Hour of Day, Sending Address, Receiving Address, Amount, Transaction Type, Location Region, IP Prefix, Login Frequency, Session Duration, Purchase Pattern, Age Group, Risk Score, and Anomaly level. This rich dataset serves as the foundation for all subsequent analyses, providing the raw data necessary for model training, validation, and testing.

2.3. Data Collection and Analysis

The dataset was obtained through a sophisticated model that simulates blockchain transactions within the Open Metaverse [26]. This model incorporates distributions, behavioral patterns, and risk assessments to generate a dataset that closely mirrors the complexity and diversity of real-world metaverse activities. The data collection process was designed to ensure a balanced representation of various transaction types, user behaviors, and risk profiles, thereby enhancing the robustness and applicability of the research findings.

Furthermore, data analysis was conducted, the analysis was executed in several stages, beginning with preliminary data processing, including cleaning (removal of duplicates and handling missing values) and feature selection. The drop_columns (e.g., 'timestamp', 'sending_address', 'receiving_address', 'anomaly') were excluded to focus on attributes directly relevant to the study's objectives. Additionally, categorical variables (e.g., 'transaction_type', 'location_region', 'purchase_pattern', 'age_group') were encoded using Label Encoding to facilitate their use in machine learning models.

Subsequently, the dataset was split into training and test sets, maintaining a proportion of 80% for training and 20% for testing. A range of machine learning models, including RandomForestRegressor, LinearRegression, SVR, DecisionTreeRegressor, KNeighborsRegressor, GradientBoostingRegressor, AdaBoostRegressor, BaggingRegressor, XGBRegressor, and LightGBM, were applied to the training data. These models were selected for their varied approaches to regression and prediction, enabling a comprehensive evaluation of their effectiveness in predicting risk scores associated with metaverse transactions.

Cross-validation (using KFold with 5 splits) and grid search (for hyperparameter tuning) techniques were employed to assess model performance and optimize model parameters, respectively. The models' predictive performance was evaluated based on the mean squared error (MSE) metric, allowing for a quantitative comparison of their accuracy in forecasting transaction risk scores.

2.4. Decision Tree

Decision Trees operate as a non-parametric method under supervised learning, suitable for both classification and regression tasks. They utilize a tree-like model of decisions that are made according to the values of different features. By dividing the dataset into smaller subsets on the basis of input feature values, a decision tree recursively applies these splits, creating a branching tree architecture. To decide on these splits, metrics such as Gini Impurity and Information Gain (Entropy) are frequently employed. These principles are detailed in Equations 1 and Equation 2.

$$Gini = 1 - \sum_{k=1}^{N}(p_k)^2 \qquad (1)$$

$$H(S) = -\sum_{k=1}^{N} p_k log_2(p_k) \qquad (2)$$

Where p_k is the proportion of samples that belong to class k in the set S.

2.5. Random Forest

The Random Forest is an ensemble learning technique used for classification and regression that builds numerous decision trees during the training phase. Its output is the combined result of the individual trees' predictions. Random Forest utilizes Bootstrap Aggregating, or Bagging, which generates multiple subsets from the original data through replacement, forming bootstrap samples. Each tree within the Random Forest is then trained using one of these bootstrap samples. For a dataset D with size N, a bootstrap sample D_i, also size N, is generated by sampling with replacement from D. This sampling is replicated to produce as many datasets as there are trees in the forest.

In the Random Forest, each decision tree is crafted by randomly selecting a subset of features at each decision point. Given M total features, a smaller number m (m << M) is chosen to ensure that at each decision point in the tree, only m features are randomly picked from the M available, and the node is split based on the best among these m features. The value of m remains fixed while growing the forest. For tasks involving regression, the Random Forest's output is the average of all individual trees' predictions. Mathematically, if h(x, Θ_i) represents the prediction by the i-th tree, then the overall prediction of the Random Forest, H(x), for an input x is formulated in Equation 3.

$$H(x) = \frac{1}{N}\sum_{i=1}^{N} h(x_i, \Theta_i) \qquad (3)$$

Where N is the number of trees, and Θ_i represents the parameters of the i_{th} tree. For classification, the output is the class selected by most trees (majority voting). Each tree gives a 'vote' for a class, and the class with the most votes is chosen as the final prediction.

3. Result and Discussion

The evaluation of various machine learning models on the blockchain transactions dataset as presented in the Table 1.

Table 1. Table Results of Machine Learning Performance

Model	Mean CV MSE
RandomForest	-0.004451
LinearRegression	-224.670175
SVR	-468.568389
DecisionTree	-0.005183
KNeighbors	-311.254124
GradientBoosting	-0.572428
AdaBoost	-39.269765
Bagging	-0.004154
XGB	-0.034141
LightGBM	-64.077053

Table 1 reveals a nuanced landscape of predictive capabilities. This reflected in their Mean Cross-Validation Mean Squared Error (Mean CV MSE) scores. These metrics provide critical insights into the effectiveness of each model in deciphering the complex patterns inherent in the transaction data.

The RandomForest and Bagging regressors exhibit remarkably low Mean CV MSEs of -0.00445 and -0.00415, respectively. These models, by leveraging ensemble techniques that aggregate predictions from multiple decision trees, achieve a balance between bias and variance, leading to a more reliable and robust prediction mechanism. Their success underscores the strength of ensemble methods in handling the diverse and complex nature of blockchain transaction data, where the amalgamation of multiple learning algorithms can effectively mitigate the limitations and noise present in individual models.

In stark contrast, LinearRegression and SVR (Support Vector Regression) encountered significant challenges, as indicated by their substantially higher Mean CV MSEs of -224.67 and -468.57. The linear model's poor performance suggests a fundamental misalignment with the nonlinear patterns and intricate relationships in the data, highlighting its limitations in capturing complex, multifaceted interactions. Meanwhile, the SVR's considerable error margin points to difficulties in kernel selection and parameter optimization, which are crucial in adapting the model to the specificities of high-dimensional transaction data. The DecisionTree model, with a Mean CV MSE of -0.00518, demonstrates a competent handling of the data's hierarchical structure. Despite its straightforward approach, where decisions are made based on the feature values, it appears to be slightly less effective than its ensemble counterparts, likely due to its vulnerability to overfitting and sensitivity to noisy data.

Among the boosting models, GradientBoosting and AdaBoost reported Mean CV MSEs of -0.572 and -39.27, highlighting the impact of iterative error correction in enhancing predictive accuracy. These models sequentially refine their predictions, focusing on the most challenging aspects of the dataset. The XGBRegressor, an advanced implementation of gradient boosting, achieves a more favorable Mean CV MSE of -0.03414, attesting to its optimized algorithms and capability to handle complex data structures efficiently. LightGBM, with a Mean CV MSE of -64.08, further illustrates the efficacy of gradient boosting frameworks, particularly in large datasets and high-dimensional spaces. Its performance, while not surpassing the ensemble tree-based methods, demonstrates a competitive edge in speed and efficiency, catering to the need for scalable and effective analysis in large-scale blockchain environments.

These results offer a comprehensive view of the models' performances, with ensemble methods like RandomForest and Bagging taking the lead in predictive accuracy. The findings suggest that in the context of blockchain transaction analysis, where the data can exhibit a high degree of variability and complexity, ensemble and boosting methods are particularly adept at capturing the nuanced patterns necessary for effective anomaly detection and fraud analysis. This comparative analysis not only contributes to the understanding of machine learning applications in financial transaction monitoring but also guides future research in selecting appropriate modeling techniques for similar tasks in the Open Metaverse.

4. Conclusion

This study extensively explored anomaly detection and fraud analysis in the Open Metaverse's blockchain transactions, utilizing various machine learning models. It found that ensemble methods like RandomForest and Bagging were most accurate, based on their Mean CV MSE scores, highlighting the complexity of blockchain data and the need for sophisticated analysis to identify fraud effectively. Linear models and SVR were less effective, pointing to the importance of avoiding linear assumptions and optimizing parameters. Meanwhile, advanced models like XGBRegressor and LightGBM showed promise, suggesting areas for future research. This work contributes to financial technology knowledge, especially in the Open Metaverse, by showing how different machine learning approaches can help secure blockchain transactions. It encourages further investigation into advanced machine learning techniques and their integration with blockchain technologies to improve fraud prevention. This research supports ongoing efforts to enhance the security and reliability of blockchain transactions in virtual environments.

References

[1] Renieris, E. M. (2023). *Beyond data: Reclaiming human rights at the dawn of the metaverse*. MIT Press.

[2] Chen, H., Duan, H., Abdallah, M., Zhu, Y., Wen, Y., Saddik, A. E., & Cai, W. (2023). Web3 Metaverse: State-of-the-art and vision. *ACM Transactions on Multimedia Computing, Communications and Applications, 20*(4), 1-42. https://doi.org/10.1145/3630258

[3] Clark, D. (2023). *Learning and the Metaverse: What this Technology Means for L&D.* Kogan Page Publishers.

[4] Rozas, D., Tenorio-Fornés, A., Díaz-Molina, S., & Hassan, S. (2021). When ostrom meets blockchain: exploring the potentials of blockchain for commons governance. *Sage Open, 11*(1), 21582440211002526. https://doi.org/10.1177/21582440211002526

[5] Fritsch, F., Emmett, J., Friedman, E., Kranjc, R., Manski, S., Zargham, M., & Bauwens, M. (2021). Challenges and approaches to scaling the global commons. *Frontiers in blockchain, 4*, 578721. https://doi.org/10.3389/fbloc.2021.578721

[6] Manski, S., & Bauwens, M. (2020). Reimagining new socio-technical economics through the application of distributed ledger technologies. *Frontiers in Blockchain, 2,* 29. https://doi.org/10.3389/fbloc.2019.00029

[7] Quaquebeke, N. V., & Gerpott, F. H. (2023). The now, new, and next of digital leadership: How Artificial Intelligence (AI) will take over and change leadership as we know it. *Journal of Leadership & Organizational Studies*, *30*(3), 265-275. https://doi.org/10.1177/15480518231181731

[8] Kayyali, M. (2023). *Introduction to HE Higher Education Ranking: Methodology, Criteria, and Indicators*. Online Submission.

[9] Prayitno, W., Tumangkar, T., & Mulyani, S. (2023). *The metaverse symbol of civilization transfer in the middle of digital economic hegemony: synthesis of progressive law of the post-modern era.* Seven Editions.

[10] Ud Din, I., Awan, K. A., Almogren, A., & Rodrigues, J. J. P. C. (2023). Integration of IoT and blockchain for decentralized management and ownership in the metaverse. *International Journal of Communication Systems, 36*(18), e5612. https://doi.org/10.1002/dac.5612

[11] Zhang, X., Min, G., Li, T., Ma, Z., Cao, X., & Wang, S. (2023). AI and Blockchain Empowered Metaverse for Web 3.0: Vision, Architecture, and Future Directions. *IEEE Communications Magazine.* https://doi.org/10.1109/MCOM.004.2200473

[12] Tan, A. (2021). Metaverse Realities: A Journey Through Governance, Legal Complexities, and the Promise of Virtual Worlds. *Legal Complexities, and the Promise of Virtual Worlds (December 9, 2021)*.

[13] Rebello, G. A. F., Camilo, G. F., de Souza, L. A. C., Potop-Butucaru, M., de Amorim, M. D., Campista, M. E. M., & Costa, L. H. M. (2024). A Survey on Blockchain Scalability: From Hardware to Layer-Two Protocols. *IEEE Communications Surveys & Tutorials* https://doi.org/10.1109/COMST.2024.3376252

[14] Al Hwaitat, A. K., et al. (2023). A new blockchain-based authentication framework for secure IoT networks. *Electronics, 12*(17), 3618. https://doi.org/10.3390/electronics12173618

[15] Javaid, U., Aman, M. N., & Sikdar, B. (2020). A scalable protocol for driving trust management in the internet of vehicles with blockchain. *IEEE Internet of Things Journal*, *7*(12), 11815-11829. https://doi.org/10.1109/JIOT.2020.3002711

[16] Nassif, A. B., Talib, M. A., Nasir, Q., & Dakalbab, F. M. (2021). Machine learning for anomaly detection: A systematic review. *IEEE Access, 9,* 78658-78700. https://doi.org/ 10.1109/ACCESS.2021.3083060

[17] Tian, X., He, J. S., & Han, M. (2021). Data-driven approaches in FinTech: a survey. *Information Discovery and Delivery, 49*(2), 123-135. https://doi.org/10.1108/IDD-06-2020-0062

[18] Otoum, Y., Gottimukkala, N., Kumar, N., & Nayak, A. (2024). Machine Learning in Metaverse Security: Current Solutions and Future Challenges. *ACM Computing Surveys*. https://doi.org/10.1145/3654663

[19] Vanini, P., Rossi, S., Zvizdic, E., & Domenig, T. (2023). Online payment fraud: from anomaly detection to risk management. *Financial Innovation, 9*(1), 66. https://doi.org/10.1186/s40854-023-00470-w

[20] Elshenraki, H. N. (2024). Forecasting Cyber Crime in the Metaverse Era: Future Criminal Methods-Readiness Requirements. *In Forecasting Cyber Crimes in the Age of the Metaverse* (pp. 1-23). https://doi.org/10.4018/979-8-3693-0220-0.ch001

[21] Hassan, M., Aziz, L. A.-R., & Andriansyah, Y. (2023). The role of artificial intelligence in modern banking: an exploration of AI-driven approaches for enhanced fraud prevention, risk management, and regulatory compliance. *Review of Contemporary Business Analyses, 6*(1), 110-132.

[22] Swati, S., & Kumar, M. (2023). Innovations in Blockchain Using Artificial Intelligence. In *Blockchain and its Applications in Industry 4.0* (pp. 179-210). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-8730-4_7

[23] Truong, V. T., Le, L., & Niyato, D. (2023). Blockchain meets metaverse and digital asset management: A comprehensive survey. *IEEE Access, 11, 26258-26288.*

[24] Pomerleau, P. L., & Lowery, D. L. (2020). *Countering Cyber Threats to Financial Institutions: A Private Public Partnership Approach to Critical Infrastructure Protection*. Springer.

[25] Radanliev, P. (2024). The rise and fall of cryptocurrencies: defining the economic and social values of blockchain technologies, assessing the opportunities, and defining the financial and cybersecurity risks of the Metaverse. *Financial Innovation*, *10*(1), 1. https://doi.org/10.1186/s40854-023-00537-8

[26] Janjua, F. I. (2023). Metaverse Financial Transactions Dataset. Retrieved from https://www.kaggle.com/datasets/faizaniftikharjanjua/metaverse-financial-transactions-dataset