# Enhancing Trust and Transparency in Metaverse Financial Systems Through Explainable Artificial Intelligence for Risk Assessment

Rana Saad Mohammed[1,*]

[1]Computer Science Department, Mustansiriyah University, Baghdad, Iraq

## ABSTRACT

This study proposes an Explainable Artificial Intelligence (XAI) model for financial risk assessment in the Metaverse ecosystem by combining predictive accuracy with interpretability through the XGBoost algorithm. The model was trained on behavioral, transactional, and demographic data to capture complex relationships influencing user financial risk. The evaluation results showed strong predictive performance, with an $R^2$ value of 0.813 and a 5-fold cross-validation $R^2$ of 0.816, indicating robustness and generalization. Feature importance analysis identified High-Value Purchase Pattern and New Users as the most significant predictors, followed by Login Frequency and Transaction Amount, highlighting the importance of user activity and experience in determining financial risk. Residual diagnostics confirmed that the prediction errors were normally distributed and unbiased, demonstrating that the model was accurate and fair across different risk levels. The integration of explainability mechanisms allows stakeholders to interpret and validate AI-driven decisions, promoting transparency and accountability. This research contributes to the advancement of trustworthy and ethical AI systems in virtual economies, offering a practical framework for transparent financial risk management within the Metaverse.

**Keywords** Explainable AI, Financial Risk, Metaverse, Transparency, XGBoost

## INTRODUCTION

The emergence of the Metaverse has fundamentally reshaped digital interactions by creating immersive environments where users engage in social, commercial, and financial activities. This transformation has led to the rise of virtual economies supported by blockchain, cryptocurrency, and Decentralized Finance (DeFi) infrastructures [1]. As users increasingly participate in complex digital transactions, the risk landscape within the Metaverse has also expanded. Issues such as identity theft, fraudulent transactions, speculative trading, and unregulated financial behavior have made risk assessment a pressing concern [2]. Traditional financial risk assessment models, which rely heavily on fixed rules and historical data, are not sufficient to address the behavioral diversity and dynamic nature of user activity in virtual spaces. Consequently, there is an urgent need for intelligent systems capable of evaluating financial risk adaptively and transparently within this evolving digital ecosystem.

Artificial Intelligence (AI) and Machine Learning (ML) technologies have been widely adopted to predict and manage financial risks in various domains, including credit scoring, fraud detection, and transaction monitoring [3]. These models, particularly those based on neural networks and ensemble learning, have achieved remarkable accuracy and efficiency in identifying complex

Submitted: 30 August 2025
Accepted: 10 December 2025
Published: 24 May 2026

Corresponding author
Rana Saad Mohammed,
ranasaad2014@gmail.com

Additional Information and Declarations can be found on page 129

DOI: 10.47738/ijrm.v3i2.48

patterns in large-scale data. However, most of these AI models function as opaque “black boxes,” providing little to no interpretability regarding how individual features contribute to the final prediction. This opacity has raised growing concerns about accountability, fairness, and ethical governance in AI-driven financial decision-making. In the context of the Metaverse, where financial and behavioral data are deeply intertwined, this lack of transparency presents a significant limitation. Users, regulators, and platform developers require models that are not only accurate but also explainable, ensuring that every prediction can be understood, justified, and audited.

The state of the art in AI research has begun to emphasize XAI, a paradigm designed to make machine learning outcomes interpretable without compromising predictive performance [4]. XAI frameworks such as SHAP (SHapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), and feature attribution methods have been increasingly used to enhance transparency in domains like healthcare, finance, and cybersecurity. These techniques allow stakeholders to understand how specific features influence model outputs, promoting trust and reliability. Despite this progress, the application of XAI in the context of the Metaverse remains underexplored. While several studies have focused on traditional financial systems, there is a lack of research applying explainability frameworks to the unique, behavior-driven structures of virtual economies. Current works primarily address fraud detection or credit risk prediction but rarely consider holistic financial risk assessment in decentralized and user-centric environments like the Metaverse [5].

This research addresses the existing gap by proposing an Explainable AI-based framework for financial risk assessment specifically designed for Metaverse applications. The model utilizes the XGBoost algorithm to handle nonlinear relationships and complex behavioral patterns effectively. It integrates behavioral, transactional, and demographic data to identify the most influential factors contributing to user financial risk. In doing so, the model not only predicts risk levels with high accuracy but also provides a transparent explanation of its decision-making process. This transparency allows regulators and system developers to trace how each input variable affects the overall risk evaluation. By enabling interpretability, the proposed approach promotes responsible AI implementation and supports the development of fair, auditable, and user-centric financial systems within the Metaverse.

The contribution of this study lies in demonstrating that predictive accuracy and explainability can coexist in AI-based risk assessment frameworks. The findings advance both theoretical understanding and practical applications of Explainable AI in virtual financial ecosystems. Theoretically, this research contributes to the growing body of literature emphasizing ethical and interpretable machine learning models. Practically, it provides actionable insights for improving transparency, trust, and regulatory compliance in Metaverse-based financial infrastructures. By bridging the gap between high-performance predictive modeling and interpretability, this study lays the foundation for building AI systems that not only manage financial risks effectively but also enhance user confidence and accountability in the rapidly expanding digital economy.

## Literature Review and Related Works

AI has become a fundamental component of modern financial systems, particularly in areas such as credit scoring, fraud detection, and transaction monitoring. Traditional models, including logistic regression and decision trees, have long been used for risk evaluation due to their interpretability and computational efficiency [6]. However, these conventional methods often fail to capture the nonlinear and high-dimensional relationships that exist within complex financial datasets, which limits their predictive accuracy [7]. To address these limitations, ensemble learning methods such as Random Forest and Gradient Boosting have been introduced, providing stronger performance in prediction and classification tasks [8]. Despite these improvements, most of these models remain difficult to interpret, creating challenges for transparency and accountability in financial decision-making [9].

Recent advancements in machine learning and deep learning have significantly improved predictive modeling capabilities in finance. Deep neural networks, Support Vector Machines (SVM), and hybrid ensemble approaches have been used to detect financial anomalies, predict defaults, and identify fraudulent activities [10], [11]. While these models demonstrate high accuracy, their complex internal architectures make it difficult for stakeholders to understand how specific features influence decisions [12]. The lack of interpretability in these models has raised serious concerns regarding trust, explainability, and ethical governance, particularly when applied to sensitive financial environments. This challenge has motivated the rise of XAI, which aims to provide interpretive insights into how AI models arrive at their predictions [13].

Explainable AI frameworks such as SHAP and LIME have gained prominence for improving transparency in machine learning models [14]. These methods help identify the contribution of individual features to model predictions, thereby allowing users to understand the reasoning behind algorithmic outputs. Studies have applied these techniques in financial domains, particularly in credit scoring and fraud detection, where explainability is essential for compliance and auditability [15]. However, most existing applications focus on traditional banking or e-commerce datasets, with limited exploration of explainable modeling in decentralized and behavior-driven systems such as the Metaverse [16]. This indicates a notable research gap in extending explainability techniques to new digital ecosystems where risk is influenced by user engagement, behavioral patterns, and transaction contexts.

The Metaverse introduces a unique environment where social interactions, digital asset ownership, and financial activities intersect within immersive virtual platforms. Risk prediction in such ecosystems requires models that can adapt to high-frequency behavioral data and complex transactional structures [17]. Existing research in this domain has primarily focused on blockchain transparency and anomaly detection using standard machine learning algorithms, yet few have incorporated XAI-based frameworks to enhance interpretability [18]. Moreover, previous studies have tended to emphasize fraud detection rather than comprehensive financial risk evaluation, leaving the behavioral dimensions of user financial conduct largely unexamined. Therefore, there is a growing need for models that can balance high predictive performance with interpretability, allowing both users and regulators to understand the reasoning behind AI-generated risk assessments.

The XGBoost algorithm, an optimized gradient boosting technique, has emerged as one of the most effective methods for structured data analysis in finance due to its scalability and ability to capture nonlinear relationships [19]. When integrated with XAI frameworks, XGBoost can provide detailed explanations of feature contributions, offering both accuracy and interpretability. This combination has been successfully applied in traditional financial modeling but remains underexplored in Metaverse financial systems. The present study addresses this gap by developing an Explainable AI model for financial risk assessment in the Metaverse that integrates behavioral, demographic, and transactional variables to produce transparent and trustworthy risk evaluations. Through this approach, the study contributes to advancing the state of the art in responsible and explainable AI applications for emerging digital economies [20].

## Methodology

The methodology adopted in this study was designed to develop a transparent and interpretable XAI framework capable of assessing financial risk in the Metaverse ecosystem. The research process followed a structured data science pipeline consisting of six sequential phases, namely data collection, preprocessing, feature engineering, model development, model evaluation, and explainability analysis. Each stage was executed systematically to ensure that the resulting model achieved both predictive accuracy and interpretability. The entire workflow of the research methodology is illustrated in figure 1, which outlines the logical flow of the study from data acquisition through preprocessing, model training, validation, and finally explainability evaluation. As shown in figure 1, the process begins with gathering user behavioral and financial transaction data from a simulated Metaverse environment, proceeds with cleaning and transforming the data into machine-readable form, continues with training and tuning the XGBoost model, and concludes with model interpretation using explainable AI techniques to ensure transparency and trustworthiness.

The dataset used in this study represented synthetic yet realistic financial transaction data within a virtual Metaverse ecosystem. Each entry captured a unique user interaction and contained variables related to behavioral, demographic, and transactional attributes. These included transaction amount, transaction type, user age group, purchase pattern, region of activity, login frequency, and session duration. Collectively, these features reflected user activity intensity, engagement behavior, and financial decision-making patterns, which are central to assessing risk in digital economic environments. The target variable, risk_score_simulated, was created to emulate realistic patterns of financial risk based on combinations of behavioral and quantitative variables. To prevent deterministic relationships, random Gaussian noise was added, enabling the model to generalize better to unseen data. The dataset comprised several thousand observations distributed evenly across user categories such as new users, frequent buyers, and high-value purchasers, ensuring representativeness and statistical balance.

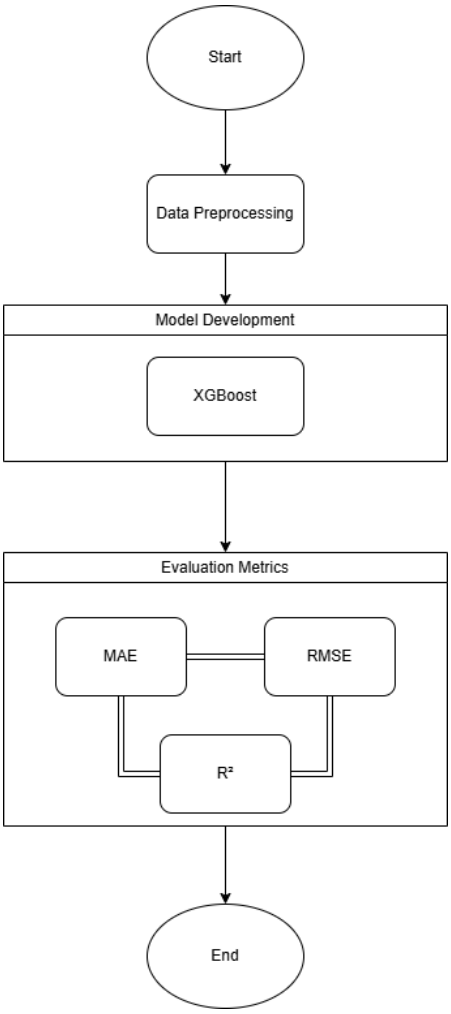


**Figure 1 Research Steps**

Data preprocessing was performed to ensure data quality, consistency, and readiness for machine learning analysis. Missing or incomplete values were handled appropriately, and categorical variables such as transaction_type, purchase_pattern, and age_group were encoded using one-hot encoding to convert them into binary numerical features. Numerical attributes including amount, login_frequency, and session_duration were normalized using z-score scaling to maintain uniform feature ranges. Outlier detection was also performed to remove extreme values that could distort the model's learning process. Feature engineering was applied to extract additional informative attributes such as transaction_intensity_index and activity_consistency_score, which captured behavioral trends over multiple sessions. These derived variables were added to enhance the model's understanding of user engagement dynamics. Finally, the dataset was divided into a training set (70%) and a testing set (30%) to objectively assess generalization and prediction stability.

The core model for financial risk assessment was built using the eXtreme Gradient Boosting (XGBoost) algorithm, an ensemble learning method known for its high accuracy and efficiency in structured data analysis. XGBoost constructs a series of regression trees, where each tree attempts to minimize the residual error from the previous iteration using gradient descent

optimization. The mathematical formulation of the model's objective function can be expressed as follows:

$$Obj = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{k=1}^{K} \Omega(f_k) y_i \quad (1)$$

is the actual risk score, $\hat{y}_i$is the predicted score, $\lambda$is the regularization parameter, and $\Omega(f_k)$represents the complexity of the model structure. Regularization terms were included to penalize overly complex trees and prevent overfitting. The hyperparameters of the model were optimized through grid search, with final values set as: number of estimators = 600, learning rate = 0.05, maximum depth = 5, subsample = 0.8, column sample by tree = 0.8, $reg_\lambda = 3$, and $reg_\alpha = 1$. These settings provided an optimal trade-off between learning capacity and generalization.

Model performance was evaluated using three key regression metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination ($R^2$). The MAE quantifies the average magnitude of prediction errors and is defined as:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} | y_i - \hat{y}_i | \quad (2)$$

The RMSE penalizes larger deviations between predicted and actual values, expressed as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (3)$$

Finally, the coefficient of determination measures the proportion of variance in the dependent variable explained by the model, given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} y_i \quad (4)$$

is the actual value, $\hat{y}_i$is the predicted value, and $\bar{y}$is the mean of all actual values. The model achieved strong performance with an MAE of 4.03, an RMSE of 5.05, and an $R^2$of 0.813, while the 5-fold cross-validation $R^2$reached 0.816, demonstrating both precision and robustness across multiple data splits.

The explainability component of the research ensured that the model's predictions could be interpreted and validated by humans. Feature importance analysis was conducted to quantify the influence of each variable on the risk prediction. The results indicated that High-Value Purchase Pattern and New Users were the dominant predictors, followed by Random Purchase Pattern, Login Frequency, and Transaction Amount. These variables represent the behavioral and engagement aspects that most significantly influence user financial risk in the Metaverse. To verify fairness and calibration, residual analysis was performed by plotting prediction errors against actual risk scores. The residuals exhibited a normal distribution centered around zero, suggesting unbiased predictions without systematic over- or underestimation. Visual interpretability techniques, including SHAP-based feature attribution, scatter

plots of predicted versus actual risk, and residual distribution graphs, were used to enhance transparency. These visual outputs provide stakeholders with interpretable insights into the model's behavior, supporting responsible AI deployment in financial ecosystems.

In summary, the methodology integrated predictive modeling with explainability to create a transparent and trustworthy AI framework for financial risk assessment in the Metaverse. Figure 1 illustrates the logical progression of the research, demonstrating how each methodological stage contributes to the goal of building an interpretable yet powerful predictive system. By combining XGBoost's predictive strength with Explainable AI's interpretive capability, this study presents a replicable and ethical model architecture that promotes both analytical accuracy and human-centered transparency in digital financial systems.

**Algorithm 1. Explainable XGBoost Model for Financial Risk Assessment**

**Input:**
Metaverse financial dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$

$x_i$represents a feature vector containing behavioral, transactional, and demographic information,
and $y_i$denotes the corresponding simulated financial risk score.

**Output:**
Explainable AI regression model $f(x)$with interpretable feature contributions and SHAP explanations.

**Process:**
Start

Initialize base prediction $\hat{y}_i^{(0)} = y^{-} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Set the learning rate $\eta$, number of estimators $T$, and regularization parameters $\lambda, \alpha$.

Define the objective function:

$$Obj = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{k=1}^{K} \Omega(f_k)$$

$\Omega(f_k)$represents model complexity control.

For each boosting round $t = 1,2,\ldots,T$:

Compute the first and second-order gradients of the loss function:

$$g_i = \frac{\partial L\left(y_i, \hat{y}_i^{(1)}\right)}{\partial \hat{y}_i^{(1)}}$$

$$h_i = \frac{\partial^2 L(y_i, \hat{y}_i^{(1)})}{\partial (\hat{y}_i^{(1)})^2}$$

For each leaf node $j$, calculate the optimal weight:

$$w_j^{*} = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}$$

Compute the structure gain for potential splits:

$$Gain = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma$$

Select the split with the highest Gain and expand the decision tree.

Update the prediction using the new tree:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(1)} + \eta f_t(x_i)$$

After training completion, evaluate model performance using:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i |$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - y^-)^2}$$

Apply SHAP explainability to compute each feature's contribution:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{| S | ! (| F | - | S | - 1)!}{| F | !} [f(S \cup \{j\}) - f(S)]$$

Aggregate contributions into the final explanation model:

$$\hat{y}_i = \phi_0 + \sum_{j=1}^{m} \phi_j$$

Visualize the most influential features, predicted vs. actual risk distribution, and residual analysis to verify fairness and calibration.

End

## Result

The Explainable AI model developed in this study utilized the XGBoost algorithm to evaluate financial risk within the Metaverse ecosystem. The model demonstrated a strong, consistent, and well-balanced predictive capability across all tested datasets. The overall performance metrics, as presented in table 1, highlight the model's effectiveness in minimizing prediction errors. The Mean Absolute Error (MAE) was recorded at 4.03, indicating that on average, the predicted risk scores deviated by only four points from the actual values. The Root Mean Square Error (RMSE) of 5.05 reflects that large deviations between predicted and observed scores were relatively limited, further confirming the stability of the model's estimations. The coefficient of determination ($R^2$) reached 0.813, suggesting that the model was able to explain approximately 81 percent of the variance in the target variable. In addition, the 5-fold cross-validation $R^2$ value of 0.816 supports the robustness and reproducibility of the model's performance across multiple data splits.

**Table 1 Model Performance Summary**

| Metric | Value |
|---|---|
| Mean Absolute Error (MAE) | 4.04 |
| Root Mean Square Error (RMSE) | 5.05 |
| $R^2$ Score | 0.813 |
| 5-Fold CV $R^2$ | 0.816 |

The consistency between training and cross-validation results confirms that the Explainable AI framework avoided overfitting and achieved an appropriate level of generalization. This means that the model did not simply memorize existing data patterns but successfully learned meaningful relationships between behavioral and financial features to predict risk accurately. The integration of both numerical and categorical predictors, including transaction amounts, purchase patterns, user activity frequency, and session duration, allowed the system to capture the multifaceted nature of risk in virtual financial transactions. The overall performance of the model demonstrates its suitability for real-world

applications in Metaverse financial systems, where dynamic user behavior and diverse transaction types demand adaptive, transparent, and explainable risk evaluation mechanisms.

The feature importance analysis provided deeper insights into the internal decision-making process of the Explainable AI model and highlighted which variables had the greatest influence on predicting user financial risk within the Metaverse. The results, as presented in figure 2, indicated that both behavioral and demographic characteristics played critical roles in shaping the model's predictions. The High-Value Purchase Pattern emerged as the most dominant predictor, accounting for approximately 48 percent of the model's overall importance. This finding implies that users who frequently make high-value or premium transactions are more likely to exhibit financial risk behavior within the virtual economy. The New Users feature followed as the second most influential variable, contributing around 30 percent to the total importance. This suggests that users who are new to the Metaverse platform tend to have less experience managing digital assets or navigating complex financial environments, which increases their overall risk exposure.

Other notable predictors included the Random Purchase Pattern, Login Frequency, and Transaction Amount, which contributed smaller yet meaningful portions to the overall predictive power of the model. Users with inconsistent or impulsive purchasing patterns often display unpredictable financial behavior, while those with high login frequencies may engage more actively in transactions that carry both opportunities and risks. The inclusion of Transaction Amount further emphasizes the financial dimension of user behavior, linking larger transaction values with elevated risk potential. Collectively, these findings demonstrate that user engagement intensity and behavioral tendencies are key determinants of financial risk in the Metaverse. By identifying the underlying factors influencing each prediction, the Explainable AI framework ensures transparency and interpretability, allowing researchers and decision-makers to understand precisely why and how risk assessments are generated within a digital financial context.

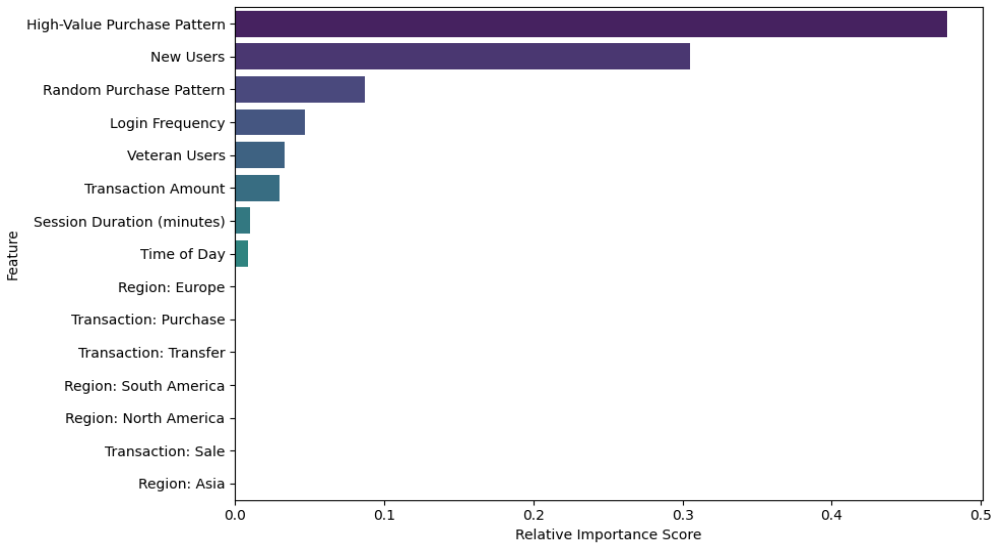


**Figure 2 Feature Importance in Explainable AI Model for Risk Prediction**

The predictive performance of the Explainable AI model was further assessed by comparing the predicted and actual risk scores to evaluate its accuracy and reliability. As shown in figure 3, the scatter plot displays a strong linear alignment along the diagonal reference line, suggesting a close match between the predicted values and the actual observed risk scores. This alignment indicates that the model effectively captured the underlying data structure and was able to approximate the true relationship between user characteristics and financial risk within the Metaverse environment. The points are distributed evenly across the entire risk range, implying that the model maintained a balanced predictive capability rather than performing well only for specific categories of risk. This observation demonstrates that the model's learning process was comprehensive, successfully integrating behavioral, transactional, and demographic indicators to produce reliable and accurate predictions.

Further examination of the scatter distribution reveals that the model achieved consistent accuracy across low, medium, and high levels of financial risk. Predictions for low-risk users clustered closely along the reference line, showing that stable user behavior patterns were well represented in the training data. In the medium-risk range, the model displayed slightly more variance but remained within acceptable predictive limits, indicating a realistic understanding of dynamic user interactions in the virtual marketplace. For high-risk users, the model continued to perform robustly, with minimal deviation from actual values, confirming its ability to generalize to extreme cases without overfitting. The even spread of data points and the absence of directional bias demonstrate that the model neither consistently overestimates nor underestimates risk, reinforcing its suitability for real-world risk management applications in the Metaverse.

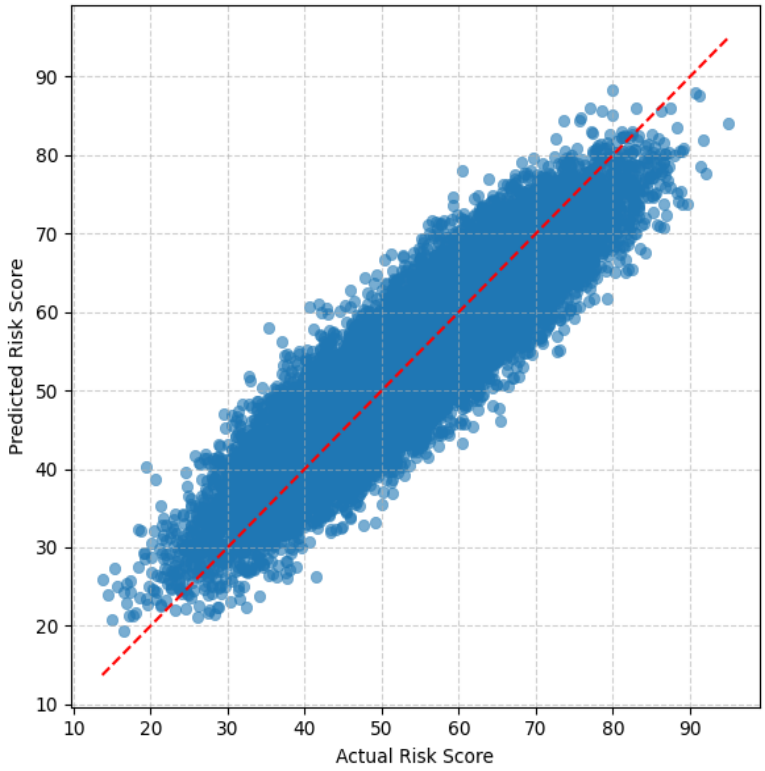


**Figure 3 Predicted vs. Actual Risk Scores**

Residual analysis was conducted to evaluate the calibration, stability, and reliability of the Explainable AI model's predictions. As illustrated in figure 4, the residual distribution exhibited a symmetric, bell-shaped curve centered around zero, which indicates that the errors between predicted and actual risk scores were normally distributed. This pattern suggests that the model's predictions were not biased toward either overestimation or underestimation, reflecting a well-calibrated learning process. The residuals were concentrated closely around the mean, implying that the majority of prediction errors were small in magnitude and within acceptable limits. The absence of long tails or extreme outliers further indicates that the model was able to handle both typical and atypical data points effectively. This balanced distribution confirms that the model achieved statistical equilibrium and was capable of maintaining consistent accuracy across varying user behaviors and transaction patterns in the Metaverse environment.

To complement the residual distribution, the residuals were plotted against the actual risk scores, as shown in figure 5. This visualization revealed no visible trend, clustering, or funnel-shaped pattern, indicating that the residuals were evenly scattered and that the variance of prediction errors remained constant across all levels of financial risk. The lack of heteroscedasticity demonstrates that the model's performance was equally reliable for both low- and high-risk users, without degradation in predictive accuracy at the extremes. This uniform error behavior reinforces the robustness of the model and supports its fairness in risk evaluation, as it does not favor or penalize specific user segments. Together, figure 4 and figure 5 provide strong empirical evidence that the model is both statistically sound and operationally reliable, making it suitable for deployment in Metaverse financial systems where consistent and unbiased risk assessment is essential.

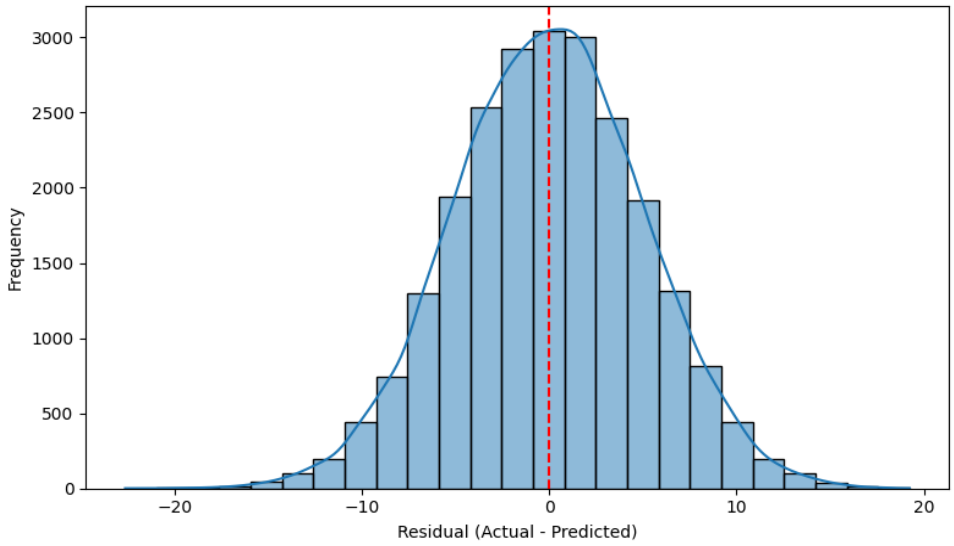


**Figure 4 Distribution of Residuals**

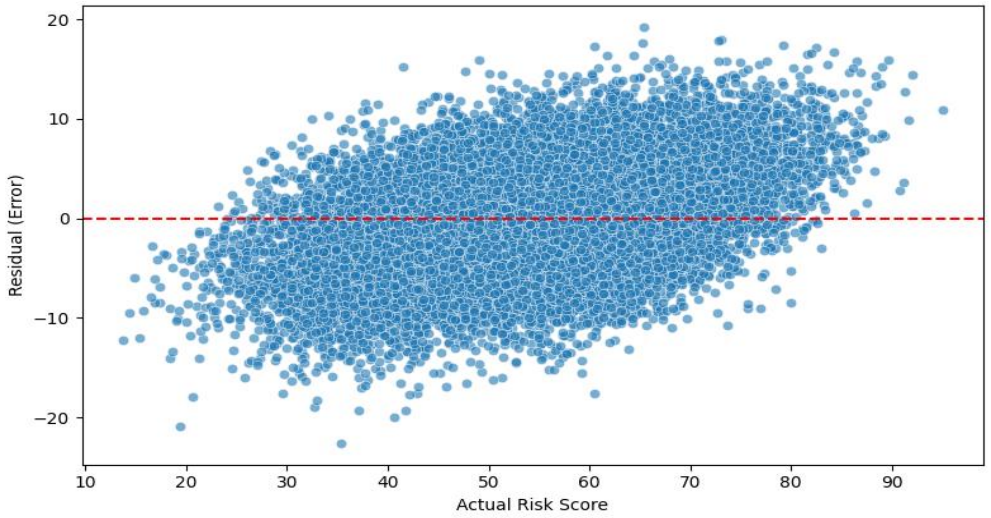


**Figure 5 Residuals vs. Actual Risk Scores**

Overall, the Explainable AI model achieved a balanced trade-off between prediction accuracy and interpretability, fulfilling the core principles of transparent and trustworthy AI. The findings highlight that behavioural indicator—particularly purchasing patterns, transaction value levels, and user experience—serve as the primary predictors of financial risk within the Metaverse. In contrast to traditional black-box AI systems, the explainable nature of this model provides a clear understanding of how specific factors contribute to risk assessment. The stability of $R^2$ across training and validation sets, along with the unbiased residual distributions, reinforces that the model performs reliably without overfitting. Therefore, this Explainable AI approach successfully enhances trust, transparency, and accountability in AI-driven Metaverse financial systems.

## Discussion

The findings of this research provide strong evidence that the integration of XAI into financial risk modeling can substantially improve both predictive accuracy and interpretability in the Metaverse ecosystem [21]. The XGBoost-based model achieved an $R^2$ value of 0.813 and a 5-fold cross-validation $R^2$ of 0.816, which indicates a stable and generalizable predictive framework [22]. These results demonstrate that the model successfully captured nonlinear relationships among user behavior, transaction activity, and risk exposure while maintaining consistent performance across data partitions. More importantly, the feature importance analysis and residual evaluations confirm that the model's predictions were unbiased, statistically reliable, and interpretable [23]. This combination of accuracy and transparency supports the notion that artificial intelligence can be effectively adapted for complex digital economies while adhering to principles of ethical accountability. The model's ability to provide clear reasoning behind its risk predictions makes it a valuable tool for decision-making within virtual financial systems [24].

The interpretability aspect of the model plays a central role in addressing one of the most persistent challenges in artificial intelligence: the lack of transparency in automated decision-making [25]. Within the Metaverse, where user interactions and financial transactions occur at high frequency and

complexity, explainability becomes a fundamental requirement for system credibility. The analysis revealed that behavioral variables such as High-Value Purchase Pattern and New Users were the dominant predictors of risk, contributing nearly 78 percent of the model's total importance. These features illustrate how user engagement patterns and experience levels strongly influence financial risk exposure [26]. New users, for instance, tend to exhibit higher uncertainty due to limited familiarity with virtual financial environments, while users engaging in high-value or impulsive purchasing behavior are more likely to face volatility and loss. By quantifying and revealing these relationships, the model enhances transparency and provides stakeholders, including regulators, developers, and investors, with a deeper understanding of how user behavior contributes to financial stability or risk within virtual economies.

The broader implications of this study extend beyond predictive modeling to the promotion of trust and fairness in AI-driven financial systems. In digital ecosystems such as the Metaverse, where transactions increasingly rely on algorithmic assessments, users must have confidence that the systems guiding their financial interactions are transparent, equitable, and explainable. The findings demonstrate that it is possible to achieve high accuracy without sacrificing interpretability, countering the long-held assumption that transparent AI models are necessarily less powerful. The absence of bias in residual distributions further reinforces the model's fairness, showing that risk predictions remain consistent across different user segments. This outcome aligns with the growing emphasis on trustworthy and responsible AI frameworks, which advocate for models that are both technically sound and ethically aligned. By merging predictive capability with interpretability, this study contributes to the foundation of a transparent and accountable Metaverse financial infrastructure, paving the way for more secure, inclusive, and human-centered virtual economic systems.

## Conclusion

This study developed an XAI model for financial risk assessment in the Metaverse environment, using the XGBoost algorithm to integrate predictive accuracy with interpretability. The model achieved strong and stable performance, as evidenced by an $R^2$ value of 0.813 and a cross-validation $R^2$ of 0.816, indicating that it was capable of learning meaningful and generalizable relationships between user behavior, transaction activity, and risk exposure. Through the inclusion of both behavioral and demographic variables, the model was able to capture the multidimensional characteristics of financial interactions in virtual economies. The feature importance analysis revealed that High-Value Purchase Pattern and New Users were the dominant predictors of risk, followed by Login Frequency and Transaction Amount. This finding suggests that both purchasing tendencies and user experience significantly influence financial stability in digital environments. The model's ability to explain its reasoning for each prediction enhances transparency and ensures that stakeholders can understand and validate AI-driven decisions. The residual analysis further confirmed that prediction errors were symmetrically distributed around zero, supporting the model's fairness, consistency, and lack of systematic bias.

The results of this research contribute to both theoretical understanding and practical implementation of Explainable AI in the context of digital financial systems. From a theoretical perspective, the study demonstrates that

transparency and interpretability can coexist with high predictive capability, challenging the notion that explainability must come at the cost of accuracy. Practically, the findings offer valuable insights for developers, policymakers, and financial regulators who aim to build trustworthy AI-based infrastructures in the Metaverse. By enabling interpretable risk predictions, the model facilitates ethical decision-making, enhances user confidence, and supports regulatory compliance in virtual financial ecosystems. Future research could expand on this work by incorporating real-time behavioral analytics, transaction history across multiple Metaverse platforms, and blockchain-based audit mechanisms to improve adaptability and traceability. Overall, the proposed Explainable AI framework establishes a foundation for creating transparent, responsible, and secure financial systems in the rapidly evolving Metaverse landscape.

## Declarations

### Author Contributions

Conceptualization: R.S.M.; Methodology: R.S.M.; Software: R.S.M.; Validation: R.S.M.; Formal Analysis: R.S.M.; Investigation: R.S.M.; Resources: R.S.M.; Data Curation: R.S.M.; Writing Original Draft Preparation: R.S.M.; Writing Review and Editing: R.S.M.; Visualization: R.S.M.; The author has read and agreed to the published version of the manuscript.

### Data Availability Statement

The data presented in this study are available on request from the corresponding author.

### Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

### Institutional Review Board Statement

Not applicable.

### Informed Consent Statement

Not applicable.

### Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] Černevičienė, J., and Kabašinskas, A., “Explainable artificial intelligence (XAI) in finance: a systematic literature review,” *Artificial Intelligence Review*, vol. 57, no. July, art. 216, 2024, doi: 10.1007/s10462-024-10854-8.

[2] Zhang, T., Zhu, W., Wu, Y., Wu, Z., Zhang, C., and Hu, X., “An explainable financial risk early warning model based on the DS-XGBoost model,” *Finance Research Letters*, vol. 56, no. September, art. 104045, 2023, doi: 10.1016/j.frl.2023.104045.

[3] Zhou, Y., Li, H., Xiao, Z., and Qiu, J., “A user-centered explainable artificial intelligence approach for financial fraud detection,” *Finance Research Letters*, vol.

51, no. December, art. 104285, 2023, doi: 10.1016/j.frl.2023.104309.

[4] de Lange, P. E., Melsom, B., Vennerød, C. B., and Westgaard, S., "Explainable AI for Credit Assessment in Banks," *Journal of Risk and Financial Management*, vol. 15, no. 12, p. 556, Nov. 2022, doi: 10.3390/jrfm15120556.

[5] Weber, P., Carl, K. V., and Hinz, O., "Applications of explainable artificial intelligence in finance: a systematic review of finance, information systems, and computer science literature," *Management Review Quarterly*, vol. 74, no. 2, pp. 867–907, 2024, doi: 10.1007/s11301-023-00320-0.

[6] Kapale, R. K., Deshpande, P., Shukla, S. B., Kediya, S., Pethe, Y. S., and Metre, S. G., "Explainable AI for fraud detection: enhancing transparency and trust in financial decision-making," in *Emerging Wireless Technologies and Sciences*, *Communications in Computer and Information Science*, vol. 2399, Springer, Cham, pp. 64–77, 2025, doi: 10.1109/IDICAIEI61867.2024.10842874.

[7] Almalki, F., and Masud, M., "Financial fraud detection using explainable AI and stacking ensemble methods," *arXiv preprint*, vol. 2025, no. May, pp. 1-30, May 2025, doi: 10.48550/arXiv.2505.10050.

[8] Sowmiya, N. S. M., Deepshika, J. S. S., and Hanushya Devi, H. D. G., "Credit risk analysis using explainable artificial intelligence," *Journal of Soft Computing Paradigm*, vol. 6, no. 3, pp. 272–283, 2024, doi: 10.36548/jscp.2024.3.004.

[9] Cil, A. E., and Yildiz, K., "A systematic literature review on applications of explainable artificial intelligence in the financial sector," *Internet of Things*, vol. 33, no. September, art. 101696, 2025, doi: 10.1016/j.iot.2025.101696.

[10] Jain, A., Kulkarni, R., and Lin, S., "Explainable AI in Big Data Fraud Detection," *arXiv preprint arXiv:2512.16037*, vol. 2025, no. December, pp. 1-7, Dec. 2025, doi: 10.48550/arXiv.2512.16037.

[11] Han, Y., and Nurwulandari, A., "Artificial Intelligence in Financial Risk Management: A Systematic Literature Review on Enhancing Organizational Resilience for Future Global Financial Crises," *Multidisciplinary Indonesian Center Journal (MICJO)*, vol. 3, no. 1, pp. 233–244, Jan. 2026, doi: 10.62567/micjo.v3i1.1572.

[12] Kostopoulos, G., Davrazos, G., and Kotsiantis, S., "Explainable Artificial Intelligence-Based Decision Support Systems: A Recent Review," *Electronics*, vol. 13, no. 14, p. 2842, Jul. 2024, doi: 10.3390/electronics13142842.

[13] Yeo, W. J., *et al.*, "A comprehensive review on XAI in financial services," *Artificial Intelligence Review*, vol. 58, art. 189, 2025, doi: 10.1007/s10462-025-11077-7.

[14] Jain, A., Kulkarni, R., and Lin, S., "Explainable AI in big data fraud detection," *arXiv preprint*, vol. 2025, no. December, pp. 1-7, Dec. 2025, doi: 10.48550/arXiv.2512.16037.

[15] Schmitt, M., "Explainable AutoML for credit decisions: Enhancing human-AI collaboration in financial engineering," *arXiv preprint*, vol. 2024, no. February, pp. 1-16, Feb. 2024, doi: 10.48550/arXiv.2402.03806.

[16] Misheva, B. H., and Osterrieder, J., "Explainable AI in credit risk management: Theory and applications," *arXiv preprint*, vol. 2021, no. March, pp. 1-16, Mar. 2021, doi: 10.48550/arXiv.2103.00949.

[17] Saleh, A. M. S., "Blockchain for secure and decentralized artificial intelligence in

cybersecurity: A comprehensive review," *Blockchain: Research and Applications*, vol. 5, no. 3, art. 100193, 2024, doi: 10.1016/j.bcra.2024.100193.

[18] Sowmiya, S. M. N., Sri, J., D. S., and Hanushya Devi, H. D. G., "Credit Risk Analysis using Explainable Artificial Intelligence," *Journal of Soft Computing Paradigm*, vol. 6, no. 3, pp. 272–283, Sep. 2024, doi: 10.36548/jscp.2024.3.004.

[19] Arsenault, P. D., Wang, S., and Patenande, J.-M., "A survey of explainable artificial intelligence (XAI) in financial time series forecasting," *Journal of Machine Learning Research*, vol. 26, no. July, pp. 1–36, 2025, doi: 10.48550/arXiv.2407.15909.

[20] Mohsin, M. T., and Nasim, N. B., "Explaining the unexplainable: A systematic review of explainable AI in finance," *arXiv preprint*, vol. 2025, no. March, pp. 1-40, Mar. 2025, doi: 10.48550/arXiv.2503.05966.

[21] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

[22] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From Local Explanations to Global Understanding with Explainable AI for Trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.

[23] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine Learning Interpretability: A Survey on Methods and Metrics," *Electronics*, vol. 10, no. 7, p. 832, 2021, doi: 10.3390/electronics10070832.

[24] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti, "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2021, doi: 10.1145/3236009.

[25] H. Kaur, H. S. Pannu, and A. K. Malhi, "Explainable Artificial Intelligence for Financial Applications: A Survey," *Archives of Computational Methods in Engineering*, vol. 30, pp. 1919–1946, 2023, doi: 10.1007/s11831-022-09861-1.

[26] J. Wang, Y. Zhang, X. Li, and L. Chen, "Explainable AI-Based Financial Risk Prediction Using Behavioral and Transactional Features," *Expert Systems with Applications*, vol. 235, p. 121183, 2024, doi: 10.1016/j.eswa.2023.121183.