Meta-universe Financial Transaction Anomaly Detection and Risk Prediction based on Machine Learning

Muxuan Li

School of Economics and Management, Xidian University, 266 Xifeng Road, Chang 'an District, Xi 'an, Shaanxi, China
22061300060@stu.xidian.edu.cn

Abstract. As blockchain, virtual reality, and artificial intelligence rapidly advance, the Metaverse is shifting from sci-fi to actuality. This evolution not only promises to transform human existence but also stands to profoundly influence financial transactions. Representing the next-gen Internet, the Metaverse strives to establish a fully immersive, temporally dynamic, self-sufficient virtual environment for human interaction across leisure, professional, and social domains. This paper delves into the analysis of blockchain financial transaction datasets within an open Metaverse environment, aiming to detect anomalous data and fraudulent activities. Employing a spectrum of machine learning models and deep learning methodologies, including support vector regression, linear regression, random forests, neural networks, and XGBoost, this study seeks to analyze and predict abnormal transactions and fraudulence. Furthermore, it aims to assess the risk associated with transactions within the Metaverse and establish a comprehensive Metaverse transaction risk scoring model. The findings underscore the efficacy of employing Random Forest and XGBoost models in crafting risk scoring models within the Metaverse context.

Keywords: Machine learning, ensemble learning, data analysis, anomaly detection

1 Introduction

Fueled by recent advancements in emerging technologies such as augmented reality, artificial intelligence, and blockchain, the metaverse is transitioning from science fiction to an imminent reality [1]. It represents a virtual realm where individuals can engage with each other via digital avatars. Leveraging cutting-edge technologies like blockchain, virtual reality, and artificial intelligence, the goal is to establish a seamless linkage between the physical realm and the virtual domain. Metauniverse financial transactions encompass a range of financial activities conducted within the metauniverse, such as digital currency transactions, digital asset transactions, decentralized finance, etc. Virtual and augmented reality technologies are converging on the topic, and the digital marketing fields are showing interest in the field, writes M Damar in his 2021 article Metaverse shape of your life for future: A bibliometric snapshot. Journal of Metaverse. In the next 15-20 years, the metaverse may enter many

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_14

areas of our lives, taking advantage of the opportunities presented by technological developments to shape our lives [2]. Therefore, financial transactions in the metaverse will be increasingly applied in future human society. In metaverse transactions, the complexity and diversity of multiple factors such as user behavior, transaction types, and transaction addresses make security a significant issue. Issues with abnormal transactions and fraud are inevitably going to arise within this context. Metaverse transactions are characterized by high frequency, small amounts, and anonymity, making traditional fraud detection methods difficult to apply. Traditional fraud detection methods typically rely on rules and thresholds; however, these methods are not ideal for the complex and ever-changing environment of metaverse transactions. Therefore, machine learning-based anomaly detection and fraud analysis technologies can play an important role. Machine learning-based anomaly detection and fraud analysis technologies can learn from historical transaction data to establish models that predict and classify new transactions, thereby identifying abnormal or fraudulent behavior [3].

We introduce a blockchain financial transaction dataset from kaggle's open metacomph, through which the research on blockchain financial transactions will be carried out in this study[4]. This dataset illustrates the multiple factors that affect financial transactions in the meta-universe, which will be pivotal in the examination of diverse meta-cosmic financial transactions in this paper.

The paper is organized as follows: Section 2 provides an overview of relevant literature in the domain. Section 3 explores the selected methodologies, elucidating their justification and the associated mathematical principles. Section 4 showcases experimental findings, dissecting the dataset to glean insights for anomaly detection and fraud analysis in transcendent financial transactions. Finally, Section 5 concludes the study and proposes a preventive model. The paper concludes with a reference section.

2 Related work

As for the research on risk trading prediction, some scholars have carried out in the past. Rokach, L., & and Maimon, O. published in Expert Systems with Applications in 2014, investigated the application of these new technologies to fraud detection in financial transactions by using various machine learning techniques. The study found that rule-based, tree-based methods (such as decision trees and random forests), cluster-based methods (such as k-means algorithms), and integration-based methods (such as AdaBoost and Bagging) perform well in fraud detection [5].

In addition, Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. In an article published in JAMA in 2018, "Deep Learning-Based Financial Fraud Detection Model for Mobile Payments," The research introduces a deep learning-driven model for detecting financial fraud in mobile payments. The findings indicate that leveraging deep learning methods like Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM) yield promising results, fraud in financial transactions can be more accurately detected [6].

The outcomes of these investigations are beneficial. The former incorporates machine learning methodologies into fraud detection research, while the latter employs an array of deep learning techniques, including Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM), to address this issue, aiming to further improve the accuracy of the experiment. However, there are still some limitations in these experiments. The first is data sparsity. The data sets used in many experiments may have the problem of data imbalance or sparse data, resulting in deviations in the prediction of the model. Secondly, feature selection is not fully considered in some experiments on feature selection. There may be too many or irrelevant features, resulting in overfitting or performance degradation of the model.

3 Method

In this study, initially, an exploratory examination is conducted to delve into certain issues and patterns elucidated by the dataset, encompassing data visualization. For instance, analyzing the dataset's dispersion, such as the transaction amount's spread, the change of transaction type over time, the relationship between risk score and transaction type, amount, the relationship between login frequency and session duration, data correlation analysis and outlier detection. Then, the data will be preprocessed, and then the research will focus on solving one of the core problems of this paper in the meta-cosmic financial transaction: the risk analysis and prediction of the meta-cosmic financial transaction. We developed and trained several machine learning models, including linear regression (LR), support vector regression (SVR), Random Forest (RF), XGBoost regression (XGB) to obtain corresponding results for further analysis. Finally, this study will integrate all the data to find the most suitable model among these models and draw the final conclusion. The workflow described in Figure 1 illustrates the approach used in this article.

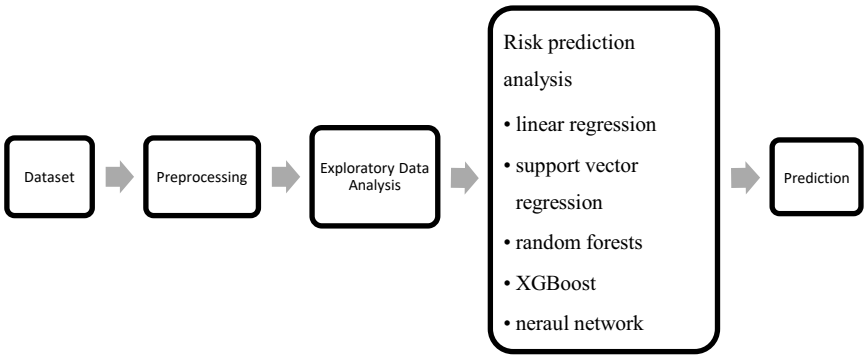

Figure 1. Research workflow

3.1 Preprocessing

Before establishing and training the trading risk prediction model, this paper needs to pre-process the data. First, the missing values in the data set are detected and processed. In the data preprocessing stage, after detection, there are no missing values in the data set. After that, given the correlation with the label used for prediction, the correlation with the transaction risk score is ranked, and the few features whose correlation coefficients are not high enough are discarded and no longer considered, and the remaining features are put into the model for training.

There are several non-numerical characteristics in the dataset, including timestamp, address, transaction type, geographic location, purchase pattern, and age group. These features need to be converted into numerical features to be used in machine learning models. Here, One-Hot Encoding is used to convert these non-numerical features. At the same time, the timestamp is broken down into separate numerical features such as year, month, day, and hour, as different parts of the date and time may have an impact on the risk score.

This article will employ the interquartile range (IQR) method for outlier detection. This method assumes a normal distribution of the data and divides it into four segments:

First quartile (Q1): Represents the value below which 25% of the data points lie.

Median (Q2): Represents the value below which 50% of the data points lie.The third quartile (Q3) represents the value below which 75% of the data points lie. Interquartile range (IQR) is the difference between Q3 and Q1.

Outliers are defined as data points below Q1-1.5 * IQR or above Q3 + 1.5 * IQR. These data points are considered anomalies because they fall outside the normal data distribution. Finally, the results are visualized with box diagram.

3.2 Exploratory analysis of data

Next, exploratory analysis of the data will be carried out to gain a more profound comprehension of the data's characteristics. In this paper, correlation analysis of each feature will be carried out through the feature matrix, the distribution of each data feature will be visualized, outliers in the data set will be detected by means of box diagram, and the importance of the features will be explored.

3.3 Model Selection and Construction

In this study, the ensemble learning model will be used to perform the task of predicting the transaction risk. Ensemble learning models can mitigate the risk of overfitting associated with individual models and enhance the overall generalization capability by amalgamating predictions from multiple base models. Meanwhile, with the use of different integration methods such as Bagging, Boosting and Stacking, ensemble learning can more effectively capture the complexity and diversity of data, thus improving the accuracy and robustness of the model. Compared with other machine learning algorithms, ensemble learning models, capable of amalgamating predictions from multiple base models, exhibit distinct advantages to improve the accuracy and

reliability of forecasts, thereby more effectively reducing trading risk and enhancing the credibility of investment decisions.

In machine learning, prevalent integration techniques encompass Bagging, Boosting, and Stacking. Bagging creates multiple subsamples from the training dataset using bootstrap sampling. These subsamples train multiple base models, and the integrated model prediction is obtained by averaging or voting. Random Forest is an example, consisting of multiple decision trees trained on different subsamples [7]. Boosting gradually improves base model performance through sequential iteration. It adjusts sample weights based on previous prediction errors, giving more attention to misclassified samples, thus reducing overall prediction error [8]. Stacking aggregates forecasts from various base models into a meta-model to derive final predictions. For instance, a linear regression model can serve as the meta-model. This study employs random forests and XGBoost for predicting trading risk [9]. Furthermore, to assess and compare the efficacy of different models, this research also incorporates Linear Regression and SVR.

- Random Forest

Random Forest is an ensemble learning framework consisting of multiple decision trees and makes predictions by means of voting or averaging. It has good generalization ability and anti-overfitting ability. By synthesizing the results of each decision tree, the accuracy of predictions can be enhanced.. In the construction process of each tree, random forest makes each tree based on different subsets by sampling the features and sampling the samples, thus increasing the diversity of the model. This diversity allows random forests to effectively mitigating overfitting risk while maintaining robust performance with extensive data volumes [10]. Bagging plays a pivotal role in a random forest, constructing numerous decision trees through randomized sampling of training datasets and then averaging or voting their predictions to get a final prediction.

$$\hat{f}_B(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \quad (1)$$

Where, B represents the count of decision trees, $\hat{f}B(x)$ denotes the prediction outcome of the b decision tree, and $f_b(x)$ signifies the ultimate prediction outcome of the random forest.

The meta-cosmic financial transaction dataset cited in this paper contains multiple features, and the random forest can automatically select features and deal with nonlinear relationships.

- XGBoost

XGBoost is developed by amalgamating numerous weak learners, typically decision trees. The core principle is to iteratively train the model by optimizing the gradient descent of the loss function, while incorporating regularization terms to prevent overfitting. Based on the framework of the gradient propulsion machine, the Taylor expansion is used to approximate the second derivative of the loss function under the current model. This makes the optimization process faster and the prediction performance higher [11]. Because the dataset contains a large number of features, the XGBoost model can effectively simplify the model's intricacy and improve the accuracy of risk prediction. At the same time, XGBoost model has strong nonlinear expression ability and can handle nonlinear relations in data sets [12]. A typical loss function can be expressed as,

$$\text{Loss} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1} \Omega(f_k) \quad (2)$$

L represents the data loss function, quantifying the disparity between the predicted value $\hat{y}_i$ and the true label y_i.

$\Omega(f_k)$ is a regularization term used to control model complexity and prevent overfitting.

- Linear regression

Linear regression model is simple and easy to interpret, can effectively capture the correlation between features and labels, thus holding significant importance in understanding the linear relationship of data sets and predicting the value of target variables. In the risk prediction in this paper, the linear regression model can provide a basic prediction result for the research, indicating the influence of characteristics on the target variable [13]. Parameter estimation formula of linear regression model,

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (3)$$

where: $\hat{\beta}$ is the estimated value of parameter β; X is a design matrix containing independent variables; Y is the dependent variable vector. This formula estimates the parameters of the model by minimizing the residuals

- SVR

SVR is a powerful supervised learning algorithm based on finding hyperplanes that can maximize the separation of data points in a high-dimensional space. SVM enhances classification robustness and generalization by maximizing the margin delimited by support vectors, the nearest data points to the hyperplane. By solving convex optimization problems, SVM can find the optimal hyperplane, maximizing the distance between the support vectors and the hyperplane [14]. Because support vector regression can deal with nonlinear relation, it has good performance in nonlinear relation prediction of data set. Through kernel technique, SVR can fit the nonlinear relationship in the data set well, thus improving the accuracy of prediction [15].

4 Result and Discussion

4.1 Overview of the data set

This paper uses kaggle's meta-cosmic financial transaction dataset, this dataset illustrates the multiple factors that affect financial transactions in the meta-universe, including the timestamp at which the transaction occurred, the transaction address, the transaction amount, the transaction type (e.g. Purchase, sale, transfer), prefix of transaction IP address, frequency of login, duration of session, purchase pattern, age group of user, risk score of transaction, mark whether transaction is abnormal, etc., these data will be significant in the analysis of various meta-cosmic financial transactions in this paper[4]. A few of the more important features are briefly described in table 1.

Table 1. Data characteristic information

timestamp	The time when the transaction occurred
amount	The amount of the transaction
transaction_type	Type of transaction (for example, purchase, sale, transfer)
risk_score	The risk score of the transaction
anomaly	Indicates whether a transaction is abnormal

This paper visually analyzes dataset attributes to identify predictors of trading risk. The correlation matrix displayed below illustrates relationships between attributes (Figure 2). For instance, a positive correlation exists between "risk score," "amount," and "session duration," suggesting that higher transaction amounts or longer session durations may correspond to higher risk scores.

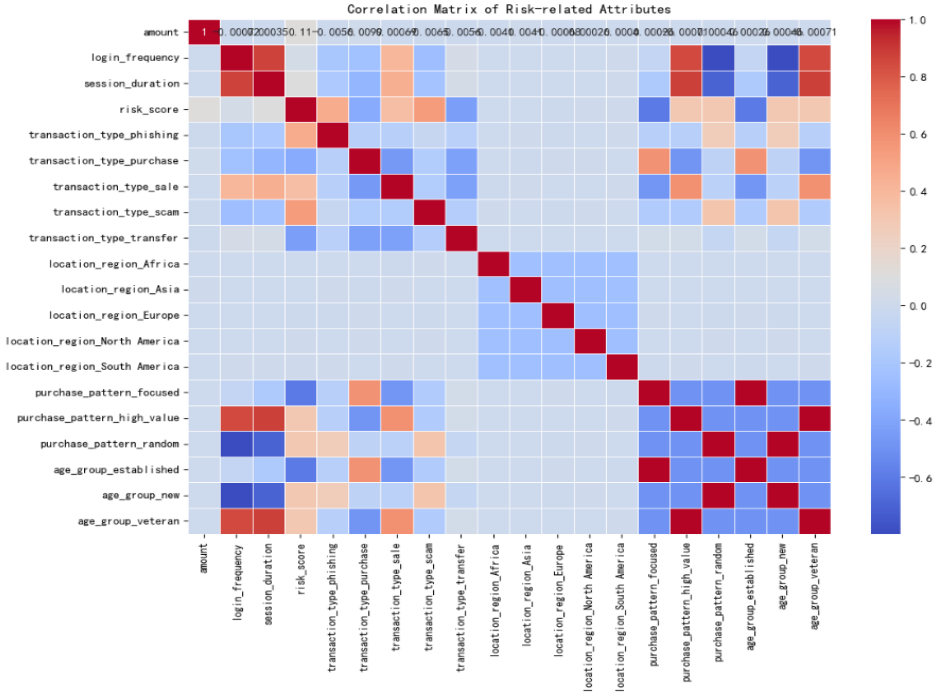

Figure 2. The correlation analysis of the dataset

This paper additionally visualizes the distribution of each attribute, facilitating the understanding of fundamental data characteristics, thereby aiding in the detection of data quality issues and the identification of abnormal patterns (Figure 3). It also presents several visualizations pertinent to anomaly detection and fraud analysis. The subsequent chart illustrates the relationship between transaction amount, session duration, and risk score. Furthermore, the paper visualizes the distribution of risk scores for different transaction types, revealing that “sale” transactions typically exhibit higher risk scores.

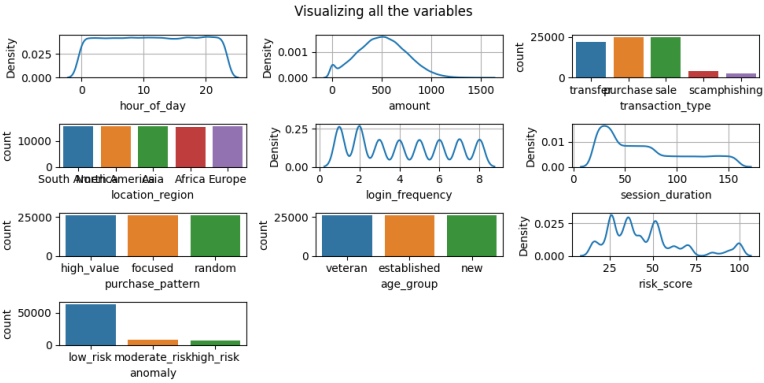

Figure 3. Variables of the features

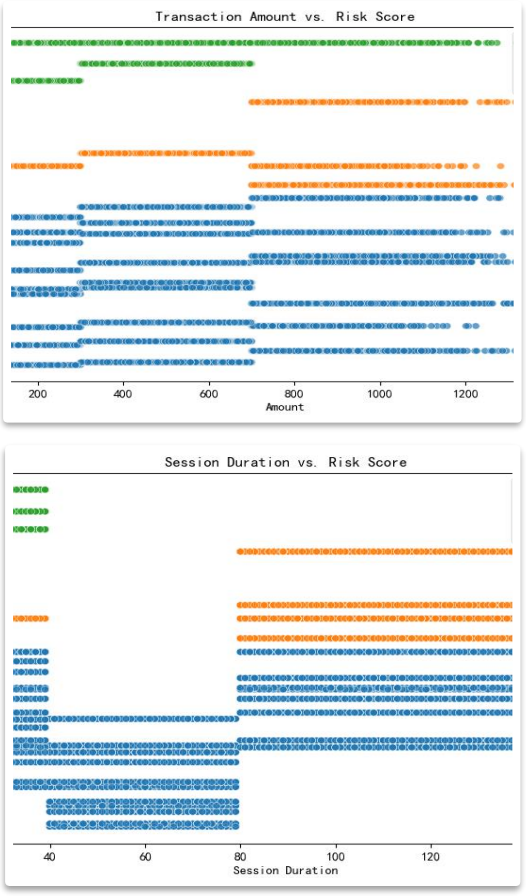

Figure 4. Relationship between risk score and several related characteristics

Next, we selected a few characteristics we wanted to explore and visualized their relationship to the risk score (Figure 4). First, it shows the risk score distribution of different transaction types (such as transfer, purchase, sale, etc.). Each type of transaction has a certain distribution of risk scores, and certain types of transactions appear to have higher risk scores.

The second is the relationship between risk score and transaction amount. As we can see in the figure, obvious linear relationship can be clearly observed between risk score and transaction amount, but large transactions seem to have higher risk score. Finally, the relationship between login frequency and session duration shows data points according to login frequency and session duration. Different age groups and purchase patterns are represented in different colors and styles, and different age groups and purchase patterns may cluster in certain intervals.

4.2 Anomaly detection

As mentioned in 3.1, the boxplot is used here to analyze the transaction amount and risk score. It shows a boxplot of the risk score and the transaction amount, respectively. The first is a boxplot of the transaction amount. The box plot shows the distribution of transaction amounts, with the box representing Q1 and Q3 of the data, and the line in the box representing the median (Q2). As you can see, there are points outside the box that are the detected outliers. Next is a boxplot of the risk score. Again, this box plot shows the distribution of risk scores (Figure 5). It can be seen that there are many outliers in the risk score, and most of the outliers have very high risk scores.

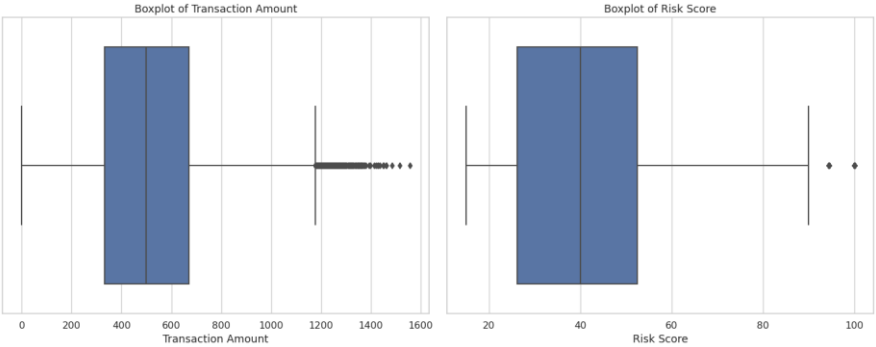

Figure 5. Visual outlier

After further analysis of the data set, the dataset shows transaction amounts averaging $502.57, the standard deviation is $245.90, the range is from $0.01 to $1557.15. Most transactions fall between $331.32 and $669.53. Risk scores average 44.96, and standard deviation is 21.78, ranging from 15 to 100, and login frequency averages 4.18, ranging from 1 to 8. Session duration averages 69.68, ranging from 20 to 159. Using the IQR method, 274 outliers were found for transaction amounts and 5,869 for risk scores. These outliers span various transaction types and often have very high risk scores, indicating elevated risk levels.

4.3 Feature importance exploration

From the figure 6, apart from risk_score, hour_of_day, transaction_type were of significant importance, followed by amount, session_duration, age_group, purchase_pattern, In addition, login_frequency also has a certain impact on the result, while sending_address, receiving_address, location_region features are not significant.

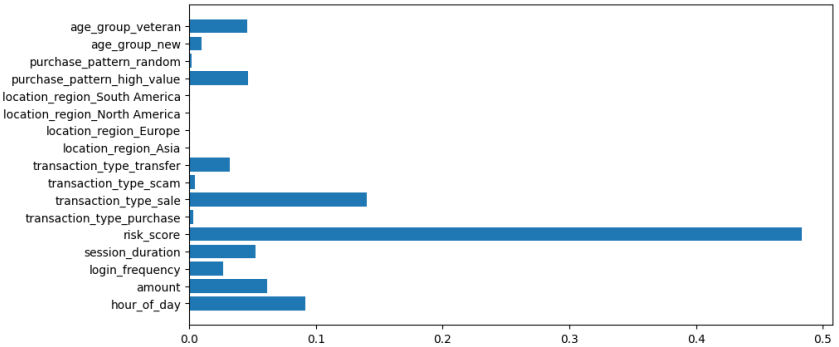

Figure 6. The importance of the features

4.4 Evaluation Indicators

The metrics used to evaluate the model are MSE, MAE, recall, and f1 scores. MSE, MAE, R-squared, and Max Error are metrics used to evaluate the performance of regression models. MSE and MAE measure the average discrepancies between predicted and true values, with MSE emphasizing larger errors due to squaring. R-squared assesses the proportion of variance explained by the model, with values closer to 1 indicating better fit. Max Error represents the maximum deviation between predicted and true values, highlighting the largest individual error.

4.5 Model Evaluation

MSE and MAE serve as fundamental measures for assessing predictive accuracy in various models. For Random Forest and XGBoost, both exhibit notably low MSE values—0.001945 and 0.000676, respectively—indicating minimal average deviation between predicted and true values. This underscores their robust predictive accuracy. Similarly, their MAE values, 0.001219 and 0.000541 respectively, further corroborate their ability to maintain low average absolute deviations (Table 2).

In contrast, linear regression and SVR display substantially higher MSE and MAE values. Linear regression yields an MSE of 216.21, and SVR yields 455.07, indicating significant average deviations between predicted and true values. Correspondingly, their MAE values stand at 9.52 and 14.43, respectively, reflecting notable average absolute deviations from the true values.

R-squared, another crucial metric, sheds light on the models' ability to explain variations in the target variable. Both Random Forest and XGBoost demonstrate R-

squared values close to unity—0.999996 and 0.999999, respectively—suggesting exceptional explanatory power and excellent model fit. Conversely, linear regression and SVR exhibit markedly lower R-squared values of 0.53 and 0.02, respectively, indicating their inadequate ability to explain target variable variations and subpar model fitting.

Maximum Error, reflecting the maximum deviation among predicted values, further emphasizes model performance. Both Random Forest and XGBoost display minimal maximum errors—3.31 and 2.85, respectively—underscoring their consistency in predictions. In contrast, linear regression and SVR show notably higher maximum errors at 55.38 and 69.52, respectively, indicating substantial deviations in certain prediction samples and suggesting potential limitations in their predictive capabilities.

Table 2 results of Performance Evaluation

Model	MSE	MAE	R-squared	Max Error
Random Forest	0.00194	0.00121	0.99999	3.30750
XGBoost	0.00067	0.00054	0.99999	2.84777
Linear Regression	216.21094	9.52446	0.53423	55.38389
SVR	455.07481	14.43454	0.01967	69.52351

Based on the experimental results (Table 2), the scatter plot which used to show the difference between the simulated prediction and the real situation of the four regression models is drawn to visualize the effects of the four methods.

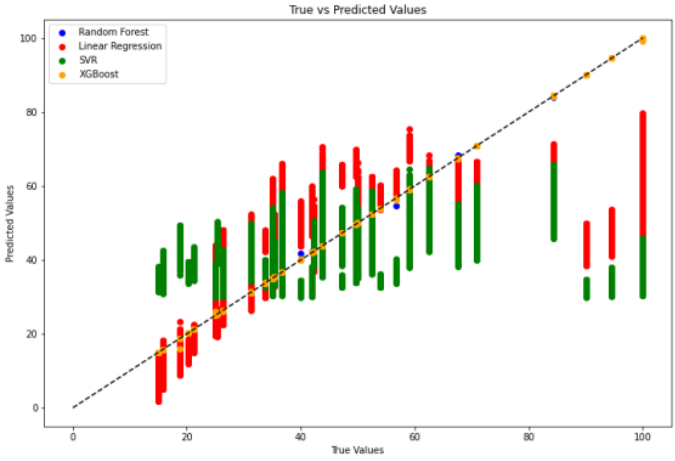

Figure 7. The visualization of model effects

In the figure 7 we can find a positive correlation between the predicted value and the true value of the random forest model and XGBoost model, and most of the data points are concentrated on the line y=x, indicating that their prediction accuracy is high. However, there are also some data points that are far from the line, which may be due to overfitting or sample bias.

The relationship about the two kinds of value of the linear regression model seems to be less obvious, and some data points even show a negative correlation. This may mean that linear regression models are poorly adapted to such data sets, or that there are problems such as multicollinearity. The SVR model is similar to the linear regression model, but the prediction is better.

5 Conclusion

This study conducted comprehensive analysis on metaverse financial transaction data utilizing various machine learning and deep learning models, encompassing anomaly detection, user behavior analysis, fraud analysis, and risk prediction. Results demonstrated excellent performance of Random Forest and XGBoost models across multiple evaluation metrics. Specifically, the Random Forest model exhibited an MSE of 0.001945, MAE of 0.001219, an R-squared value close to 1 (0.999996), with a maximum error of 3.31; while the XGBoost model displayed an MSE of 0.000676, MAE of 0.000541, an R-squared value approaching 1 (0.999999), with a maximum error of 2.85. These metrics indicate their high accuracy and exceptional fitting capability in prediction tasks.

In contrast, the performance of linear regression and support vector regression models was relatively poor. The linear regression model yielded an MSE of 216.21, MAE of 9.52, an R-squared value of 0.53, with a maximum error of 55.38; whereas the support vector regression model had an MSE of 455.07, MAE of 14.43, an R-squared value of 0.02, with a maximum error of 69.52. These metrics indicate significant deviations from true values and weaker ability to explain data variability.

Considering these findings, it is recommended to prioritize the utilization of Random Forest or XGBoost models in constructing risk assessment models for the metaverse financial market. Furthermore, future research endeavors could focus on algorithm optimization and application of smart contracts to enhance prediction accuracy and generalization capabilities, thereby fostering the intelligent, efficient, and secure development of the metaverse financial market.

References

1. Wang, Y., Su, Z., Zhang, N., et al. A survey on metaverse: Fundamentals, security, and privacy. arXiv preprint 2022, arXiv:2203.02662.
2. Damar, M. Metaverse shape of your life for future: A bibliometric snapshot. Journal of Metaverse, 2021, 1(1), 1-8.
3. Alom, M. Z., Taha, T. M., & Yakopcic, C. A state-of-the-art survey on deep learning theory and architectures. Electronics,2020, 9(3), 481.
4. Metaverse Financial Transactions Dataset, https://www.kaggle. com/datasets/faizaniftikharjanjua/metaverse-financial-transactions-dataset/data
5. Rokach, L., & Maimon, O. A survey on machine learning techniques for fraud detection in financial transactions. Expert Systems with Applications, 2014, 41(10), 4756-4772.

6. Caruana, R., & Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd International Conference on Machine Learning 2006,1-11.
7. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. Deep learning based financial fraud detection model for mobile payments. Journal Of The American Medical Association, 2018,320(14), 1453-1454.
8. Ting, K. M. A comparative study of cost-sensitive boosting algorithms. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(8), 1468-1481.
9. [Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. Super learner. Statistical Applications in Genetics and Molecular Biology, 2007, 6(1).
10. Breiman, L. Random forests. Machine Learning, 2001, 45(1), 5-32.
11. Tianqi Chen, Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System." Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining. 2016.
12. Tianqi Chen, Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. KDD, 2016.
13. Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning. Springer, 2001.
14. Cortes, C., & Vapnik, V. Support-vector networks. Machine learning, 1995, 20(3), 273-297.
15. Christopher M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006.