Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/92/20241678

Research on risk detection and management based on machine learning

Jiayi Sun

East China University of Science and Technology, Hangzhou,Shanghai, 200237, China

S1533065114@163.com

Abstract. With the rapid advancement in financial risk management, machine learning techniques are playing an increasingly crucial role in credit assessment and risk detection. They provide financial institutions with more scientific and precise platforms and methodologies for risk management. This paper explores various traditional machine learning methods in risk assessment, introduces the XGBoost ensemble learning method, and integrates traditional machine learning with Long Short-Term Memory (LSTM) neural networks to enhance the generalization capability of risk control and credit assessment models. This approach offers new perspectives and improvement directions for risk assessment standards and practices in the financial industry, affirming the potential application of machine learning technologies in future risk management.The results show that the random forest and decision tree have excellent accuracy and recall rates for distinguishing fraudulent transactions. After the introduction of LSTM neural networks, the accuracy and recall rates of fraudulent transaction recognition models reach around 99%, indicating that these models have good adaptability for fraud risk recognition.

Keywords: Risk prediction, Machine learning, Ensemble learning, Long Short-Term Memory (LSTM) neural networks.

1. Introduction

With the expansion of the scale of the financial market, the inflow and outflow of large amounts of funds, and the increasingly complex market conditions, traditional artificial market credit evaluation methods will face more and more challenges. Most of these traditional risk fraud identification methods rely on simple linear analysis or manual judgment, and it is difficult to capture nonlinear relationships in the face of diverse and hidden fraudulent transactions. There may be inefficiency and a low recall rate of accuracy, leading to the prevalence of fraudulent transactions, resulting in user trust problems and fund security problems for financial institutions, which may affect the profit income of financial institutions and other capital flow industries, and reduce the efficiency and speed of capital flow.

As financial markets become more complex, traditional credit assessment methods face increasing challenges. These conventional methods mostly rely on linear analysis and human judgment, which is difficult to capture nonlinear relationships within complex data, thus impacting assessment accuracy and efficiency. Therefore, in recent years, machine learning technologies have garnered widespread attention due to their advantages in big data analysis and pattern recognition. This study aims to explore the application effectiveness of machine learning technologies in risk fraud, comparing traditional assessment methods with machine learning-based approaches. It looks at how well the latter handles complicated financial data and combines optimization methods, ensemble learning, and deep learning neural networks to make credit risk management much more accurate and effective. Using deep learning and more advanced machine learning methods, achieving a higher recall rate and accuracy rate can help the system to identify more fraudulent transactions, so that a smaller number of more extreme fraudulent transactions are ignored to protect the security and stability of capital flow, thus benefiting the entire industry and society.

2. Overview of machine learning methods and data set introduction

2.1. Data Preprocessing

The dataset comprises transactional records spanning a simulated period of 30 days, equivalent to 744 hourly time steps. With a total of 6,362,620 entries across 11 columns, each row represents a unique financial transaction captured within this simulated timeframe.

Because there are a large number of traders' IDs and transfer amounts, that is, those who have used the same ID to transfer or receive transfer may use the same ID in the future period for high -risk fraud transactions. The data with duplicate IDs are made for key labels. It can be found that 9313 and 3640258 IDs are repeated in the senders and the receipts.

Therefore, the new label called Risk Factor is used to record this invisible effect. Whenever the same ID appears in fraud transactions, it will increase the risk coefficient of the ID. The correlation coefficient of 0.427 between the label and the final target tag indicates a relatively close relationship, potentially serving as a key factor for the final model prediction [1].

As a special label, STEP indicates that the time sequence may have the relationship between the final fraud transactions. Therefore, it is important to prepare data for STEP in order to identify the unit time that is most susceptible to a high volume of fraud transactions. We discover that the percentage of fraudulent transactions is minimal, and the security environment improves at the onset. However among all of them, a large number of fraud transactions are happening (only some of the high -risk fraud transactions are intercepted here, and there are 8213 fraud transactions in the total data set).

2.2. Feature engineering and visual analysis

The histogram and nuclear density estimation curve of the data source accounts reveal a high frequency of small-scale transactions, while the frequency of large-scale transactions is relatively low. A significant number of transactions have taken place when the difference between the new and old balances of the source account is minimal. Under these circumstances, the frequency of transferring new balances is higher than the old balance. We speculate that a capital transfer might happen, leading to a generally low new balance in the source account. The balance will initially be low and will gradually increase to accommodate high-quota transactions.

Simultaneously, our analysis revealed that the non-fraud environment generally adheres to the previously mentioned dense regularity of small transactions, whereas in the fraud environment, the transaction amount increases, the upper limit reaches 1*1E7, and the frequency is also significantly high [2].

In summary, calculate the correlation coefficient between all tags and make the relevant matrix. The analysis reveals a clear correlation between the RISK_FACTOR factors and the final fraud transactions, while the numerical category tags like the transaction amount and the target label are less evident.

2.3. Model implementation

This experiment uses the perceptual machine, K-Nearest Neighbors, and Decision Tree—these three traditional machine learning methods [3][4]. The following experimental results and prediction accuracy rates were obtained: The author finds that the model task is relatively simple, and the traditional machine learning method can also achieve a higher accuracy rate. The overall model can be derived. The overall model's performance has gradually improved from the perceptual machine, K, located near the neighbor, to the decision tree. The decision-making tree model, for example, has been able to complete a very good fraud transaction classification task, and the accuracy rate and recall rate have reached 0.8940 and 0.9239, respectively. As a fraud task, the focus should be on the recall rate. Fraud transactions typically represent a small percentage, and a high recovery rate indicates the presence of actual fraud. The concealment and diversity of fraud transactions require the fraud detection system to be highly sensitive and accurate.

However, in the three traditional machine learning models, the recall rate of the perception machine is as high as 0.9741, but the accuracy rate is only 0.7465, which does not imply improved model performance. High recall but low accuracy means that the model may make a large number of negative predictions as a positive example, leading to frequent false alarms and reducing users' trust in the accuracy of the system. Therefore, in the field of fraud testing, a more balanced recall rate and accuracy rate are a symbol of high model performance.

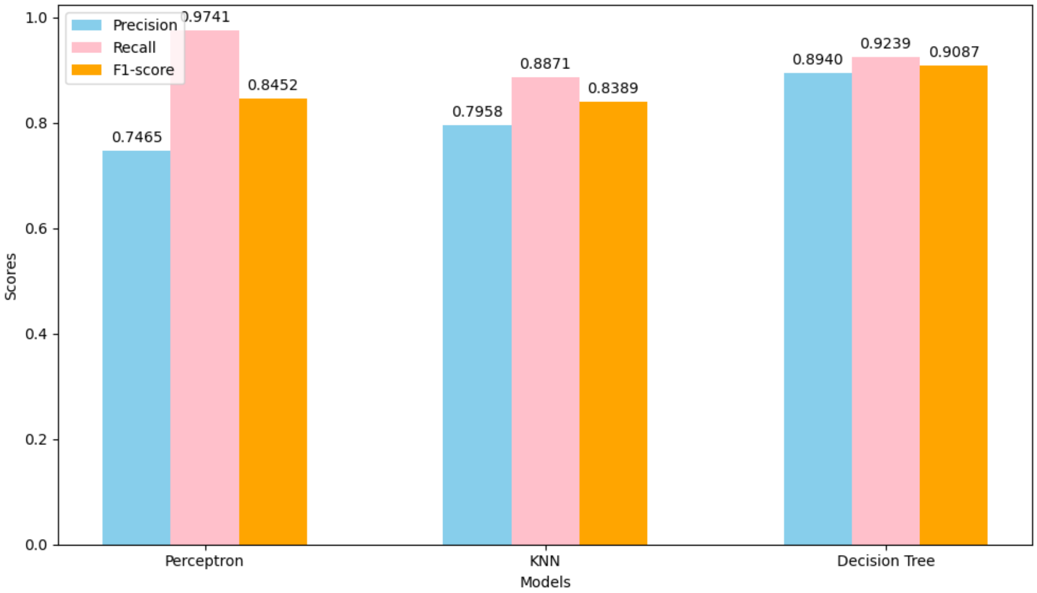

Figure 1. Comparison of Precision, Recall and F1-score

3. Model application and optimization

3.1. Integrated learning and deep neural network

Integrated learning is a paradigm of machine learning that aims to improve the overall prediction performance by combining the predictive results of multiple learning devices. Its core idea is to build multiple learning devices.

A deep neural network is a type of neural network model composed of multi -layer neurons. A deep neural network has a multi -layered non -linear transformation. Except for the input layer and output layer, each contains many neurons [5].

XGBOOST (Extreme Gradient Boosting) is an integrated learning algorithm that uses gradient boosting to train decision trees through iterative training trees [6].

Because fraud transactions may have a certain connection with time, frequent fraud transactions in a unit of time, the introduction of long -term memory neural networks to analyze fraud risk recognition. Long Short-Term Memory (LSTM) has proven that they have the effectiveness of understanding and capturing the dynamics of financial time sequences, predicting price trends, and evaluating long-term market stability [7][8].

On the basis of the decision tree in this experiment, the accuracy and recall rate of XGBOOST, random forest and LSTM neural network algorithms are used. Performance data such as recall rates have improved, especially LSTM neural networks, up to 0.996. Secondly, random forests have reached a recall rate of about 0.9875. The effect of XGBOOST's effect in this experiment is not obvious. Compared to traditional learning decision trees, the improvement is small, only around 0.97.

3.2. Model optimization technology

Due to the excellent comprehensive performance of the model, we need to consider the fitting problem in advance, and use the early stop method in LSTM to reduce the impact of overfitting on the model as much as possible.

The original LSTM neural network tentatively set its EPOCH to 10, but when the performance of the verification set (test set) deteriorated, the model halted at EPOCH = 8. At this time, the loss rate on the verification set (test set) fluctuates from 0.035 to 0.005 it can be found that in the LSTM neural network, the loss rate of the model is extremely low, which can well complete the task of monitoring fraud risk.

4. Discussion

Although machine learning has performed well in this fraud risk data set, there are still problems, such as data imbalance and feature selection. The problem of data imbalance is particularly prominent in detecting fraud risk. This is because there are very few real fraud data in the risk of fraud, but the final model prediction has high requirements for the recall rate. During the preprocessing process, the normal trading instances of the screening and reduced some of the normal trading instances for data balance and model optimization. The feature selection is also a problem in terms of fraud. Due to the large amount of numerical data, there is usually no obvious connection with the predicted target label. At this point, the author will explore the use of main component analysis and other techniques to resolve the issue. In summary, overcoming these challenges requires comprehensive consideration of data distribution, data preprocessing technology, feature engineering, etc., and continuous monitoring and updating models [9][10].

5. Conclusion

In short, the decision tree, perceptual machine, deep learning neural network and other algorithms are included in the fraud risk prediction system of the financial market, providing a comprehensive and dynamic method. The organized process of feature selection, preprocessing, model training and optimization ensures that the system can identify the complex mode in historical fraud data. By analyzing the accuracy recall rate of both input and output modules, the system is able to identify the majority of fraud transactions and enhance its overall credit rating. In short, the risk of financial fraud is closely related to our lives, not only affecting the security of our funds, but also affecting the security and stability of financial institutions and even social fund flows. Traditional machine learning and even manual screening methods can no longer meet our urgent needs for the identification of fraudulent transactions. I incorporate decision trees, perceptrons, deep learning neural networks and other algorithms into the financial market fraud risk prediction system to provide a comprehensive and dynamic approach. Organized feature selection, data preprocessing, model training and optimization processes ensure that the system can recognize complex patterns in historical fraud data. By analyzing the recall rate of input and output modules, the system can identify the most fraudulent transactions and improve its credit. This guarantees the flow of capital and ensures its safety in our country.

References

[1] Z. Ding, "Construction and Exploration of a Financial Risk Control Model Based on Machi ne Learning," 2023 International Conference on Evolutionary Algorithms and Soft Comp uting Techniques (EASCT), Bengaluru, India, 2023, pp. 1-6, doi: 10.1109/EASCT59475. 2023.10393547.

[2] S. Geetha, C. Suwetha, A. Sangavi and S. Shabana Nazrin, "Time Series Financial Market Forecasting Based on Machine Learning," 2024 2nd International Conference on Artificial

Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA), Namakkal, India, 2024, pp. 1-5, doi: 10.1109/AIMLA59606.2024.10531346.

[3] Y. Zhu, "Research on Financial Risk Control Algorithm Based on Machine Learning," 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 2021, pp. 16-19, doi: 10.1109/MLBDBI54094.2021.00011.

[4] W. Zhang, "Research on Corporate Financial Risk Prediction Model Based on Machine Lea rning," 2024 IEEE 2nd International Conference on Control, Electronics and Computer Technology (ICCECT), Jilin, China, 2024, pp. 535-540, doi: 10.1109/ICCECT60629.202 4.10545713.

[5] M. El-Bannany, A. H. Dehghan and A. M. Khedr, "Prediction of Financial Statement Fraud using Machine Learning Techniques in UAE," 2021 18th International Multi-Conference on Systems, Signals & Devices (SSD), Monastir, Tunisia, 2021, pp. 649-654, doi: 10.1109/SSD52085.2021.9429297.

[6] K. Tsarapatsani et al., "Machine Learning Models for Cardiovascular Disease Events Prediction," 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, Scotland, United Kingdom, 2022, pp. 1066-1069, doi: 10.1109/EMBC48229.2022.9871121.

[7] A. Mashrur, W. Luo, N. A. Zaidi and A. Robles-Kelly, "Machine Learning for Financial Risk Management: A Survey," in IEEE Access, vol. 8, pp. 203203-203223, 2020, doi: 10.1109/ACCESS.2020.3036322.

[8] Y. Y. Abdulla and A. I. Al-Alawi, "Advances in Machine Learning for Financial Risk Management: A Systematic Literature Review," 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), Manama, Bahrain, 2024, pp. 531-535, doi: 10.1109/ICETSIS61505.2024.10459536.

[9] K. Chaturvedi, P. K. Goel, Y. K. Bansal, T. Kaushik, A. Sharma and R. Srivel, "Fin Safe - Empowering Financial Security through the Synergy of Machine Learning and Blockchain," 2024 10th International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, India, 2024, pp. 988-992, doi: 10.1109/ICCSP60870.2024.10544039.

[10] Guo Feng, Zhuang Xudong, Wang Renzeng. Digital transformation of banks, Exogenous Fi ntech and Credit risk governance: An empirical test based on text mining and machine l earning [J]. Securities Market Review,2023(4):15-23. https://d.wanfangdata.com.cn/period ical/ChlQZXJpb2RpY2FsQ0hJTmV3UzIwMjQwNzA0Eg96cXNjZGIyMDIzMDQwMDIaC HgyN2E3eWZp