ISSN 2616-5902 Vol. 5, Issue 19: 77-84, DOI: 10.25236/AJBM.2023.051912

A Financial Risk Control Method Based on XGBoost Algorithm

Zhaoxin Li[1,a,#], Yinkai Niu[2,b,#], Xinghan Chen[3,c,#], Cheng Huang[4,d,#]

[1]*Communication University of China, Beijing, China*
[2]*Kunming University of Science and Technology, Kunming, Yunnan, China*
[3]*TEDA NO. 2 Middle School, Tianjin, China*
[4]*Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, China*
[a]*2546607634@qq.com,* [b]*2035354859@qq.com,* [c]*3594992056@qq.com,* [d]*hc3210521930@163.com*
[#]*Co-first author*

Abstract: *For the past few years, with the continuous progress of technology, the rapid integration of traditional financial industry and the Internet has given rise to online financial businesses that deal with high-concurrency, large-scale, and multidimensional data. However, due to the high profitability and high risk of the financial industry, as well as the upgrading of fraudulent means, this transformation has also put higher requirements on financial risk control. Recently, advanced technologies represented by big data and artificial intelligence have provided new directions for improving risk control capabilities for commercial banks, with machine learning playing an increasingly important role. This paper aims to predict customer fraud behavior by conducting data analysis, data preprocessing, feature engineering , dataset partitioning, and using XGBoost and LightGBM algorithms in machine learning, in order to provide assistance in ensuring the stable operation of financial institutions and the security of customer assets.*

Keywords: *machine learning, digital finance, anti-fraud, ensemble learning, XGBoost*

1. Introduction

Due to the development of Internet financial services, traditional risk management methods face challenges. However, financial data has characteristics such as diversity, large data volume, and timeliness, and financial data may have problems such as missing values, outliers, and errors, which may lead to inaccurate analysis based on statistical models and human experience. At the same time, market behavior and trends can change at any time, and they are also affected by multiple factors such as politics, economy, and society, which makes manual financial risk control work extremely inefficient and difficult. Fraud risk is one of the main risks in consumer finance business, which refers to the risk that credit customers have no intention of repayment. Currently, fraud presents industrial chain characteristics, giving rise to related technological development, identity credit packaging, and fake identity provision industries. Therefore, for major financial institutions and Internet finance companies, risk control capability directly affects the profitability of their businesses and determines the survival of enterprises. Consumer finance risk control usually consists of three stages: pre-loan, in-loan, and post-loan, and the three stages interact with each other. However, this type of financial fraud has relatively few publicly available data, which makes it difficult to obtain data.

The emergence of machine learning technology has brought new opportunities for risk management. The core of big data risk control technology is to obtain information that helps risk control decisions from high-dimensional data. The significance of machine learning in financial risk control projects lies in providing a more accurate, efficient, and automated risk assessment and management method. By analyzing a large amount of historical data and real-time market data, machine learning algorithms can obtain a relatively complete understanding of customers and learn and identify potential risk patterns and trends, thereby helping financial institutions to timely discover and respond to risks and improve the decision-making capabilities of financial institutions. By building predictive models and optimizing algorithms, machine learning can help financial institutions better assess customer credit risk, market volatility risk, operational risk, etc., thereby supporting wiser decision-making and asset allocation.[1]

In previous research, Practitioners both domestically and internationally primarily engage in statistical analysis and data mining of fraudulent transactions to uncover patterns of fraud. For instance,

ISSN 2616-5902 Vol. 5, Issue 19: 77-84, DOI: 10.25236/AJBM.2023.051912

Syda et al. proposed in 2002 that there is a certain correlation between neural networks and specific fraud patterns based on the analysis of fraudulent transaction data. In 2011, Duman combined the advantages of genetic algorithms and distributed search to create a hybrid approach used for tracking customer financial activities in large banks and predicting credit card fraud. Wu Baohua (2010) proposed that anti-fraud models could consist of four main modules: data preprocessing, neural network, output, and tracking modules, and also suggested an improved backpropagation algorithm with weighted decay.[2]

Machine learning is a branch of artificial intelligence, a modeling technique for data, and trains the established model through massive data sets to achieve the expected effect, ultimately producing a reliable model. Its goal is to enable computer systems to learn and improve from data, thereby achieving task automation or prediction capabilities. Machine learning also has multiple methods.

1) Linear regression: a supervised learning method for predicting continuous values. It establishes a model based on linear relationships and fits the data by minimizing the error between the predicted value and the actual value.

2) Logistic regression: a supervised learning method for classification problems. It establishes a model based on the logistic function and can predict the probability that a sample belongs to a certain category.

3) Random forest: Random forest is an ensemble learning method composed of multiple decision trees. Each decision tree independently trains and predicts data, and the final result is obtained by voting or averaging.

2. Methods

The whole process can be divided into four parts(Fig.1). First, the dataset is imported and exploratory data analysis (EDA) is performed. Based on the discovered data features, feature engineering techniques are applied to preprocess the dataset. Then, modeling was performed using XGBoost and LightGBM algorithms. Finally, the predictive models were evaluated using k-fold cross-validation and AUC as evaluation metrics.

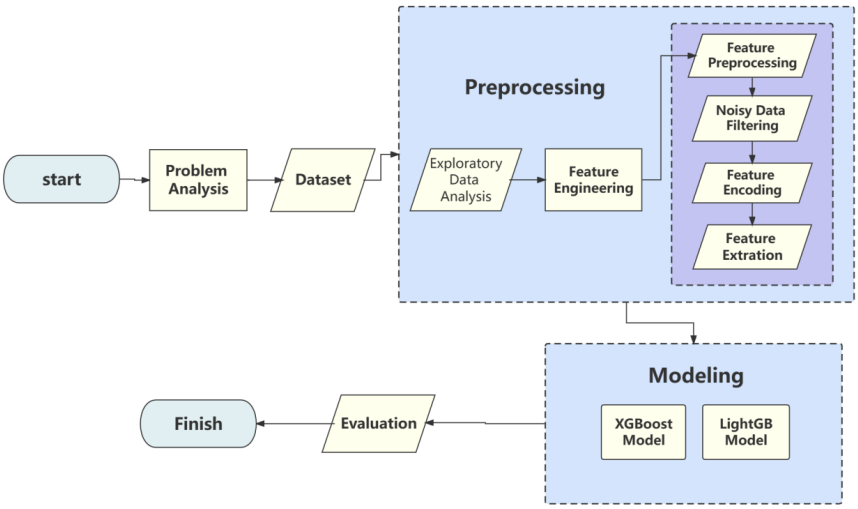

Figure 1: The Flow Chart of the Research

2.1. Exploratory Data Analysis (EDA)

The dataset is based on a large-scale, multidimensional, and complete dataset generated by an online digital finance business, with 800,000 instances and 47 features, as shown in Fig. 2, representing historical loan disbursements and some related personal information of past users. Among them, due to the confidentiality of financial data, there are 15 columns of anonymous variables.Due to the complexity and intricacy of financial data, the relationships between different data points are not always straightforward and can be complex.[3] Additionally, some data points may have weak or no correlation with each other, which can lead to extra resource consumption during calculations and reduce the efficiency and accuracy of machine learning models.At the same time, due to the small proportion of

ISSN 2616-5902 Vol. 5, Issue 19: 77-84, DOI: 10.25236/AJBM.2023.051912

fraudulent samples, the model's prediction performance fluctuates greatly and does not have stable output. The amount of financial data is extremely large, and intuitive analysis of the data may not reveal the relationships and features among them. By importing and exploring the data, we can make them more intuitive by indexing these data.

1	id	12	verificationStatus	23	pubRecBankruptcies
2	loanAmnt	13	issueDate	24	revolBal
3	term	14	purpose	25	revolUtil
4	interestRate	15	postCode	26	totalAcc
5	installment	16	regionCode	27	initialListStatus
6	grade	17	dti	28	applicationType
7	subGrade	18	delinquency_2years	29	earliesCreditLine
8	employmentTitle	19	ficoRangeLow	30	title
9	employmentLength	20	ficoRangeHigh	31	policyCode
10	homeOwnership	21	openAcc	32	n(anonymous data)
11	annualIncome	22	pubRec	33	isDefault

Figure 2: The Features in the dataset.

Random sampling: After importing the complete sample set, each experiment randomly extracts data from it to form a new data set for each experiment. Its purpose is twofold. On the one hand, it increases the randomness of the model, which helps improve the generalization ability of the model and avoid overfitting. On the other hand, for large databases, it can alleviate the impact of abnormal value distribution and the impact of the large cardinality of abnormal values on the final model learning.

2.2. Feature Engineering

Feature engineering is the most time-consuming and important step in data mining model development. Each attribute in the dataset provides different information gain for predicting the outcome. Therefore, it is often necessary to transform the original data through feature engineering to highlight the hidden features in the original data, thereby helping to decompose and aggregate the original data to better express the essence of the problem. Generally, feature engineering includes feature selection, feature transformation, and feature generation.

After performing exploratory data analysis (EDA), it was observed that the data contains missing values, standard time formats, and outliers. These types of data cannot be directly processed by the model and require preprocessing through feature engineering.

Filling missing values: Typically, for numerical data, missing values in columns with a small number of missing data are filled with the median or mean of that column, while nominal variables are filled with the mode. The mean is commonly used for filling missing values in data with a near-normal distribution, while the median is used for filling values in skewed distributions or data with outliers. Since the columns n11 and n12 have a high number of missing values (Fig.3), this set of data does not generally affect label determination and can be deleted. As the remaining data follows a normal distribution, median imputation is used for handling missing values.

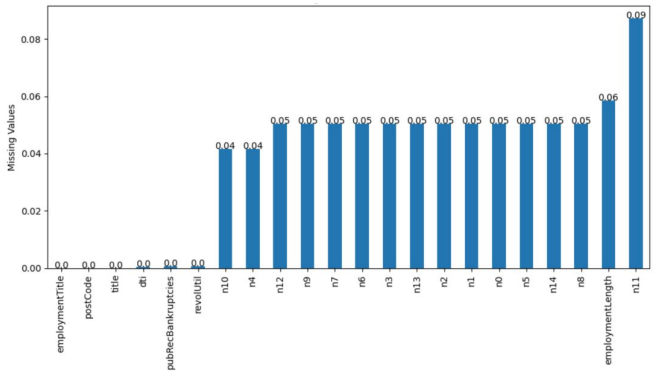

Figure 3: Distribution of the Missing Value

Converting time format data: For standard time formats in string type within a data table, which can't be recognized by the computer. Therefore, a function is used to convert the time format to an integer type. The issueDate column in the "data" is converted to datetime format[4]

Outlier Filtering: An outlier refers to an individual value in a sample that significantly deviates from

ISSN 2616-5902 Vol. 5, Issue 19: 77-84, DOI: 10.25236/AJBM.2023.051912

its average value. Values that are located more than 3 standard deviations (3σ) away from the mean are generally considered outliers(Fig.4). These outliers can affect the accuracy of the Prediction Model,causing it to be fluctuate and inaccurate.[5] So,the method which is usually applied is to removed to facilitate better information mining in subsequent analysis. The 3σ rule, which relies on the characteristics of a normal distribution, is commonly applied to identify and remove data points that fall outside this range.

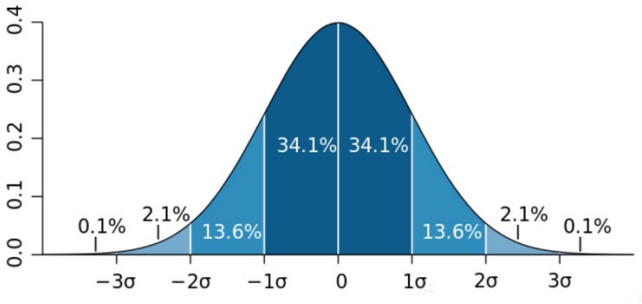

Figure 4: Probability distribution diagram of the standard normal distribution function.

Feature engineering enhancement: Feature engineering is the most time-consuming and important step in developing data mining models. Each attribute in the dataset provides different information gain for predicting the outcome. Therefore, it is necessary to transform the original data through feature engineering to highlight hidden features and facilitate the decomposition and aggregation of the data, ultimately better expressing the essence of the problem. Feature engineering typically includes feature selection, feature transformation, and feature generation.

1) Feature generation involves creating new data features by combining and cross-comparing the statistical features of the original data, such as mean and variance, to discover hidden patterns and correlations in the original data. By examining the object features in the data, it is found that there are features with inclusion relationships in the original data, such as grade and subGrade. To explore the relationship between the part of subGrade that is not included and the target feature, the data is separated and divided into a new feature attribute, which is included in the original dataset. The same data processing is applied. The heatmap shows that the separately divided data also exhibits correlation with the target feature during correlation analysis.

2) Feature transformation involves processing the data to convert numerical values or discrete values that may affect model training into a suitable format, adjusting their impact on parameter selection. For example, continuous variables can be standardized or normalized, while discrete variables can be one-hot encoded or label encoded. By examining the distribution of the LoanAmnt attribute through data distribution plots, it is observed that a normal distribution can help the model converge faster (Fig.5). Therefore, log transformation is applied to check if the data conforms to a normal distribution.

Overall, outlier handling and feature engineering are important steps in data preprocessing. By effectively handling outliers and engineering informative features, the models can achieve better performance and accuracy in predicting and detecting risks.

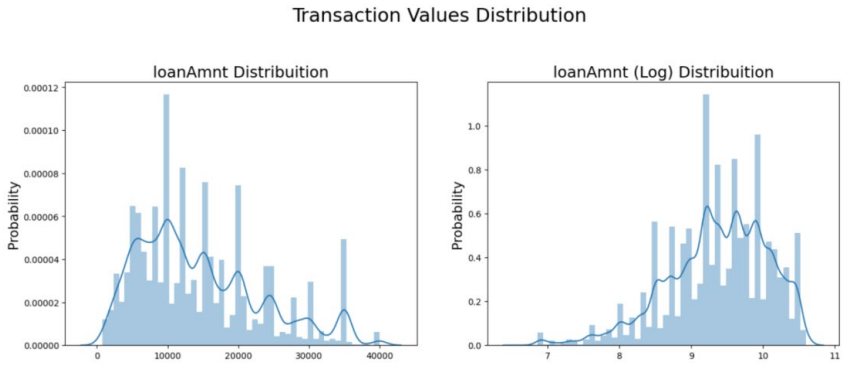

Figure 5: Transaction Values Distribution of the feature

ISSN 2616-5902 Vol. 5, Issue 19: 77-84, DOI: 10.25236/AJBM.2023.051912

3. Feature selection

By analyzing the correlation, variance, and other indicators between features and the target variable, features that have a high impact on the target variable are selected and retained, while features that have a small impact on model performance are removed. The correlation between each numerical feature (numerical_fea) and the target variable (isDefault) is calculated, and a heatmap is plotted. From the heatmap, it can be observed that applicationType, policyCode, n11, and n12 have very low correlation with the target variable (Fig.6), and they have little impact when used for prediction. The policyCode is a single value and has no impact on the label's judgment. The applicationType, due to its uneven data distribution (Fig.7), is close to a single value distribution, and according to statistical rules, its impact on the label can also be considered minimal.

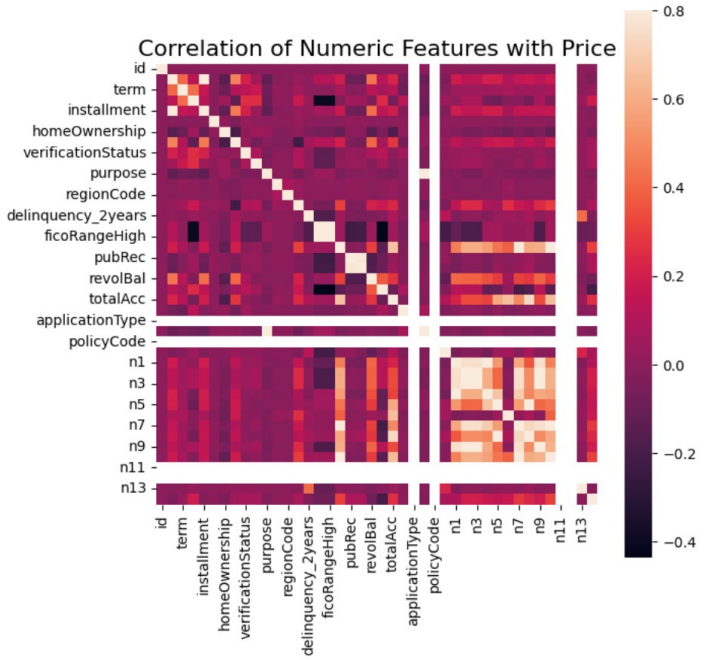

Figure 6: Correlation of Numeric Features with Price

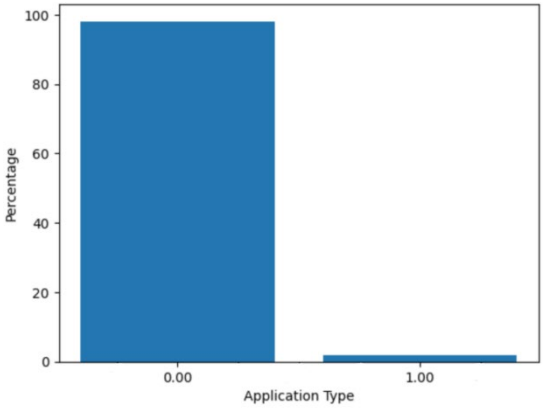

Figure 7: Distribution of the feature Application Type

3.1. Modeling

3.1.1. Algorithm

XGBoost Algorithm: XGBoost (extreme gradient boosting) is an improved version of gradient boosting based on GBDT. It is a decision tree machine learning algorithm. It has been widely recognized for its classification performance in fields such as medicine, physics, and finance, and is widely used in data mining and model prediction. The objective function is as follows[6].

ISSN 2616-5902 Vol. 5, Issue 19: 77-84, DOI: 10.25236/AJBM.2023.051912

$$\lambda^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_i) \tag{1}$$

By performing second-order Taylor expansion on the above equation and substituting GBDT, we obtain the objective function of XGBoost:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)}) + g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i) + \Omega(f_i) + cons\tan t \tag{2}$$

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \tag{3}$$

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \tag{4}$$

Compared with the GBDT algorithm, XGBoost uses the second-order expansion of the Taylor function, which increases the accuracy and also supports custom loss functions. By introducing regularization terms L1 and L2 in the objective function, it reduces the variance of the model and learns from the practice of random forests, supporting column sampling, which effectively suppresses overfitting and improves learning space. XGBoost can handle the imbalanced distribution of positive and negative samples in financial fraud data and the large data volume.[7]

LightGBM Algorithm: LightGBM (Light Gradient Boosting Machine) is proposed by Microsoft to solve the problems encountered in massive data. Compared with XGBoost, LGBM first pre-sorts all features according to their numerical values and changes the level-wise growth strategy of GBDT to a leaf-wise growth strategy. It calculates each layer's splitting node separately and only splits the leaf with the maximum gain, improving efficiency in multi-machine parallelism, achieving faster computation, and ensuring that a single machine can use more data and reducing the communication cost of parallel computing on multiple machines, improving efficiency in multi-machine parallelism.[8]

3.1.2. Model Evaluation

K-fold: To address overfitting and other issues during the training process, the K-fold method is used to divide the original data into K subsets, with one subset used as the test set and the remaining K-1 subsets used as independent training sets, increasing the randomness of the training data. The validation set is not involved in training and is independent of the prediction model, used for the final model evaluation.

AUC: Since the task is a typical binary classification problem of whether a customer has financial fraud behavior, the evaluation criterion used is AUC (Area Under the ROC Curve), which is less sensitive to data imbalance and can better characterize the performance of the model.[9]

$$TRP = \frac{TP}{TP + FN} \tag{5}$$

$$FPR = \frac{FP}{FP + TN} \tag{6}$$

TPR: True Positive Rate (TPR), also known as sensitivity or recall, is the ratio of correctly predicted positive instances to the total actual positive instances. It measures the proportion of positive instances that are correctly identified.FPR: False Positive Rate (FPR) is the ratio of incorrectly predicted positive instances to the total actual negative instances. It represents the proportion of negative instances that are wrongly classified as positive.

3.1.3. Algorithm comparison

XGBoost and LightGBM are both ensemble algorithm tools based on decision tree boosting (Tree Boosting). Ensemble learning mainly includes three strategies: diversity enhancement, learner training, and learner combination. Both algorithms have the characteristics of being insensitive to input requirements, low computational complexity, and good performance, and they are widely used in the industry. The performance of the two models on different-sized datasets is shown in the graph. When the data exceeds 10,000 records, the overall performance of the models improves, but the improvement rate is slow. However, when the data is reduced to 10,000 records, there is a significant decrease in the model's

ISSN 2616-5902 Vol. 5, Issue 19: 77-84, DOI: 10.25236/AJBM.2023.051912

accuracy.

Based on the analysis and comparison of the two models(Fig.8), it can be concluded that XGBoost performs better than LightGBM when the dataset is large or small. The performance of the two models is similar when the dataset is small, but as the dataset size increases, XGBoost's performance improvement becomes faster than LightGBM's, demonstrating XGBoost's good generalization ability.

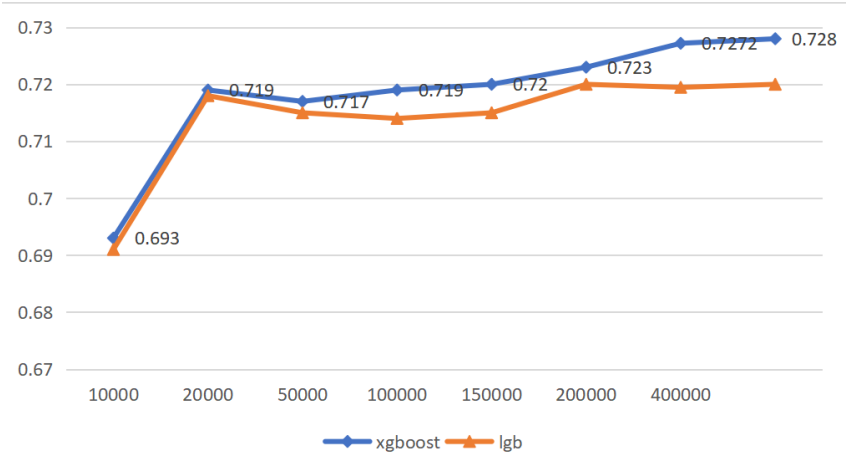

Figure 8: The AUC of the prediction models XGBoost and LightGBM with respect to the change in data volume.

The dataset includes a lot of attributes, and the correlation distribution between the attributes is complex and diverse. LightGBM's leaf-wise growth strategy is not as effective as XGBoost's level-wise growth strategy, which can be optimized for multi-threading, in identifying attributes and parameters that enhance classification performance on a global scale. Although LightGBM effectively reduces time and space consumption in computational engineering, its performance is compromised when constructing and predicting models for datasets with large data volume and complex relationships between attributes. On the other hand, XGBoost exhibits good generalization ability for such complex financial datasets.Furthermore, based on economic considerations, it is possible to control the dataset to around 200,000 records since the model's performance does not show significant changes when the dataset exceeds that size. This would still yield good experimental results.

4. Conclusion and Future Outlook

The development of big data in the financial field has brought about rapid advancements in technological capabilities for financial risk control. Although most risk control models still rely on structured data, their utilization of large-scale data is limited. This article proposes a framework for constructing risk control models using machine learning and conducts experiments in practical business scenarios to verify its feasibility.

From the experiments, it can be observed that the XGBoost model has significant advantages over the LightGB learning algorithm. Specifically, it can handle large-scale datasets, exhibit higher accuracy and robustness, handle highly imbalanced datasets, and demonstrate good interpretability in feature selection and weight allocation.

However, the XGBoost model still faces challenges and areas for improvement in financial risk control. Due to the complexity and diversity of financial data, the XGBoost model may encounter overfitting issues. To address this problem, techniques such as cross-validation and regularization can be employed to improve the model's generalization ability.

In the future, we can expect further applications and developments of the XGBoost model in the field of financial risk control. With the continuous advancement of financial technology and deep learning, providing more application scenarios for the XGBoost model, its advantages in interpretability will also offer better decision support for financial institutions, thereby enhancing the effectiveness of risk control.

In conclusion, the XGBoost model is a powerful machine learning algorithm that has achieved significant success in the field of financial risk control[10]. Through analysis and modeling of extensive financial data, the XGBoost model can assist financial institutions in more accurately assessing customer

ISSN 2616-5902 Vol. 5, Issue 19: 77-84, DOI: 10.25236/AJBM.2023.051912

credit risks and identifying potential fraudulent activities. Although there is still room for improvement in the application of the XGBoost model in financial risk control, with the constant evolution of technology and the expansion of application scenarios, we can expect it to play a more crucial role in its future development.

References

[1] He Peiyu. Research on Credit Anti-Fraud Model Prediction Based on Machine Learning (in Chinese) [D]. Shanghai Normal University, 2021. DOI: 10.27312/d.cnki.gshsu. 2021.001455.
[2] Hao Guanghao. Digital Fraud and the Application of Financial Technology Anti-fraud (in Chinese) [J]. Taxation and Economy, 2019(06): 40-47.
[3] Huang Zhangjie. Research on Credit Default Prediction Based on Machine Learning (in Chinese) [D]. Chongqing Technology and Business University, 2022. DOI:10.27713/d.cnki.gcqgs. 2022.000246.
[4] Xia Pingfan. Research on Intelligent Risk Prediction Methods for Digital Financial Fraud (in Chinese) [D]. Hefei University of Technology, 2022. DOI:10.27101/d.cnki.ghfgu. 2022.001159.
[5] Wu Qi. Research on Credit Default Prediction Based on Supervised Learning (in Chinese) [D]. Zhejiang Financial University, 2022. DOI:10.27766/d.cnki.gzjcj. 2022.000299.
[6] Cao Hanping, Zhang Xiaojing, Zhu Ruijie, et al. Application and Practice of Machine Learning Models in Real-time Anti-fraud in the Digital Financial Era (in Chinese)[J]. Journal of Intelligent Science and Technology, 2019, 1(04): 342-351.
[7] Zhu Yidong. Research on Credit Risk Control Models and Algorithms Based on Machine Learning (in Chinese) [D]. Xiamen University of Technology, 2022. DOI:10.27866/d.cnki.gxlxy. 2022.000133.
[8] Li Kaiheng. Research on Financial Risk Control Models Based on Machine Learning (in Chinese) [D]. University of Electronic Science and Technology of China, 2022. DOI:10.27005/d.cnki. gdzku. 2022. 003102.
[9] Wu Weiqiang, Hou Qilin. Consumer Finance Anti-Fraud Models and Methods based on Machine Learning Models (in Chinese) [J]. Modern Management Science, 2018(10): 51-54.
[10] Zhang Xu, Zhao Yi. "Efficient Real-Time Anti-Fraud through Online Real-Time Decision Making + Offline Machine Learning" (in Chinese) [J]. Financial Electronicization, 2016(12): 89.