Hindawi
Complexity
Volume 2020, Article ID 8706285, 14 pages
https://doi.org/10.1155/2020/8706285

Research Article

Improved ML-Based Technique for Credit Card Scoring in Internet Financial Risk Control

Shuangshuang Fan,[1] **Yanbo Shen**,[2] **and Shengnan Peng**[1]

[1]*School of Management, China University of Mining and Technology-Beijing, Beijing, CO 100080, China*
[2]*Dahua Certified Public Accountants, Beijing, CO 100080, China*

Correspondence should be addressed to Shuangshuang Fan; 934042440@qq.com

Received 17 July 2020; Revised 16 September 2020; Accepted 17 October 2020; Published 4 November 2020

Academic Editor: Min Xia

With the rapid development of China's Internet finance industry and the continuous growth of transaction amount in recent years, a variety of financial risks have increased, especially credit risk in the financial industry. Also, the credit risk evaluation is usually made by using the application card scoring model, which has the shortcomings of strict data assumption and inability to process complex data. In order to overcome the limitations of the credit card scoring model and evaluate credit risk better, this paper proposes a credit evaluation model based on extreme gradient boosting tree (XGBoost) machine learning (ML) algorithm to construct a credit risk assessment model for Internet financial institutions. At the same time, an Internet lending company in China is taken as a case study to compare the performance of the traditional credit card scoring model and the proposed machine learning (ML) algorithm model. The results show that ML algorithm has a very significant advantage in the field of Internet financial risk control, it has more accurate prediction results and has no particularly strict assumptions and restrictions on data, and the process of processing data is more convenient and reliable. We should increase the application of ML in the field of financial risk control. The value of this paper lies in enriching the related research of financial technology and providing a new reference for the practice of financial risk control.

1. Introduction

Since the 1970s, human society has entered the industrial 3.0 era marked by the application of electronic information technology, and computer technology and Internet have been widely used in various fields and integrated with traditional industries, giving birth to new business models and formats [1]. With the rapid development of China's economy and the popularity of network technology, the traditional financial industry and Internet technology are integrated and derived into a series of network-based financial products [2]. However, due to the imperfection of the trading system and the lack of convenience of operation, Internet finance did not enter the public's attention until "Yu'E Bao" was launched by a financial service company in 2013, leading to the vigorous development stage of Internet finance [3]. Relying on big data and cloud computing technology, Internet finance forms functional financial formats and services in the open Internet platform, including Internet innovation and e-commerce innovation of traditional financial institutions, APP software, e-commerce enterprises of nonfinancial institutions using Internet technology for financial operation, P2P network credit platform, crowd-funding network investment platform, and financial resource mode of mobile financing APP and third-party payment platform [4]. At present, Internet finance has been on a healthy development track in the strategic environment of "green finance" and "science and technology power" advocated by the state and the Chinese government [5].

Due to the late start of China's Internet finance, the regulatory system needs to be improved. Internet finance not only brings vitality to financial enterprises and social financing and investment activities but also causes various

potential risks and challenges. From 2016 to 2018, more than 200 Internet financial platforms in China have defaulted. The credit risk of the involved platforms leads to huge losses, such as operators' fraud or loss of money with them, overdue repayment by borrowers, and collapse of P2P platforms [6]. Based on the increasing negative effect of Internet financial risk on society, it is urgent to establish an effective risk control system. In the traditional financial industry, the credit scoring card model is usually established to deal with credit risk. It uses a large number of historical credit data to describe the customer's income status, credit history, payment level, and other indicators and gives different weights. The indicators are divided into several levels and scored according to the historical data of customers to obtain the relevant credit rating [7].

However, due to the complexity of the modelling process and the limited accuracy of processing a large number of highly complex information, the traditional credit score card model is prone to bias and has some limitations in Internet financial risk management [8]. In this paper, the ML model is proposed to predict credit risk by collecting and mining Internet data, repeatedly calculating, and verifying. Through case study and empirical study, it is concluded that under the same data sources, the ML model has higher accuracy and recall rate than the traditional credit scoring model, and it plays an important role in the Internet financial risk control system. The main contributions of this paper are as follows.

1.1. Contribution in Theory. This paper enriches the theory of ML in the field of financial risk control. The theoretical research on the application of ML in Internet financial risk control in China has not yet formed a perfect system. At the same time, most of the foreign research focuses on the financial market system risk early warning, anti-money laundering of financial institutions, and other aspects, and research focusing on the content of Internet financial risk control is relatively small. In this paper, the application of ML algorithm in the credit risk management of Internet finance is first proposed, which has strong innovation.

1.2. Contribution in Practice. In view of the serious credit risk in China's Internet financial industry, this paper proposes a financial risk control method based on ML algorithm. At the same time, the case study verifies the superiority of the proposed method. Therefore, this study provides valuable and meaningful guidance for the risk management of the actual Internet financial industry and helps to reduce the risk of China's Internet financial industry.

This paper proposes that universities, scientific research institutions, and Internet financial industry should cooperate and communicate with each other. It promotes the latest research results in ML algorithm of scientific research institutions, which can well transfer the value of its practice, that is, to serve the Internet financial industry. The application of science and technology is emphasized. It promotes the close relationship between industry and academia, thus contributing to the strategy of "rejuvenating the country through science and education" advocated by China.

1.3. Contribution for Further Research. This paper presents the application of ML algorithm in Internet financial risk control. Because the traditional risk assessment method has been widely used and has strong interpreted ability, the ideal situation is that the two methods are effectively combined. Then, the proposed method provides a preliminary reference for the future combination of traditional credit scoring model and ML algorithm model. With the deepening of future research, we will explore how to effectively combine different advanced methods.

The remainder of this paper is organized as follows.

Section 2 gives the background and related work. This section reviews the related research results of financial industry risk control at home and abroad and points out the shortcomings of these achievements and the basic ideas of this paper. Section 3 presents the technical model. This section describes the credit card scoring theory model and XGBoost ML method model and points out the evaluation model-related indicators. Section 4 is devoted to case study and empirical analysis. We take a P2P enterprise in China as an example and analyze the advantages of the proposed model. In Section 5, we draw a conclusion.

2. Background and Related Work

This section will systematically introduce the research background and related work. It lists the relevant research on the credit evaluation method by international scholars and Chinese scholars through reading literature. The specific work in this section is divided into academic research on credit scoring, international scholars' research on ML in the field of financial risk control, and Chinese scholars' research on financial risk control.

At the end of this section, it is pointed out that the traditional credit scoring model for financial risk control has limitations, that is, the data have strict assumptions, and it must be linear and cannot process large-scale data. The main research content of this paper is the application of ML algorithm in the field of Internet financial risk control.

2.1. Traditional Credit Scoring Model. The history of credit scoring in the world can be traced back to the 1950s. Mathematician Earl Isaac and engineer Bill Fair first established the world's first commercial credit scoring system FICO and extended it to the financial system [9]. After that, financial institutions make credit decisions through the 5C credit discrimination method [10]. 5C discriminant analysis is composed of five evaluation factors, such as the lender's role, capital, collateral, capacity, and environment. It comprehensively forecasts the performance of borrowers. The limitation of this method is that it is inefficient in processing large-scale data.

However, with the expansion of the loan scale of financial institutions and the increase of the number of borrowers, the above credit evaluation method is not

applicable, and a new method, that is, the credit scoring method, has been adopted. Financial institutions build data-driven models based on quantifiable characteristics of borrowers to manage credit risk [11]. The credit score is based on the historical credit data of the borrower, and the credit score is calculated by the model, and the credit granting person determines whether to grant credit or not and the credit line according to the credit score. Hand and Henley pointed out that the statistical techniques and quantitative methods in the construction of scorecards have been extended from discriminant analysis and linear regression methods widely used in the early stage to logical regression, probit regression, nonparametric smoothing method, Markov chain model, recursive segmentation, expert system, and genetic algorithm [12].

Subsequently, Lee and other scholars empirically studied the effectiveness of using multiple adaptive regression spine (MARS) and classified regression tree (CART) for credit scoring. The two methods are superior to the traditional discriminant analysis and logical regression methods in the accuracy of credit scoring [13]. According to Bee et al. [14], with the development of current data mining technology, the process of establishing a credit scoring model is more convenient, and various new technologies have been developed. However, in the practical application of financial institutions, the commonly used technologies are still logical regression and decision tree because such technologies are more convenient in identifying important input variables, interpreting results, and building models.

2.2. Application of ML in Financial Risk Control. With the development of big data and data mining technology, international scholars have formed rich research results on ML in credit risk prediction and evaluation. Because the goal of the credit management of financial institutions is to optimize the business performance and minimize the risk, decision rules should be established to make credit decisions. Therefore, clustering algorithm is widely used in the credit scoring system in the early stage. For example, William and Huang combined the K-means clustering method with the supervision method for insurance risk identification [15].

Furthermore, Yeo et al. [16], used hierarchical clustering technology to predict the risk of automobile insurance industry. Different customer risk levels are identified by clustering technology to make operational decisions on credit limit. With the increase of data scale, scholars try to build more complex models, such as Khandani, Kim and so on. They use ML algorithm and statistical model to predict consumer default risk with massive customer transaction records and credit management agency data. Their research results show that ML technology reduces the prediction error of 6% to 25% compared with the traditional linear regression model [17]. Chakrabort and Joseph trained a set of financial distress prediction model based on ML and proposed that the ML method was better than the statistical models such as logical regression. In terms of the receiver operating characteristic area discrimination, there is about 10% significant improvement [18]. Ticknor proposed to use neural network algorithm to predict financial market behavior. Empirical results show that the model constructed by this algorithm has the same prediction effect as the advanced model without data preprocessing [19]. Gogas and Agrapetidou constructed a prediction model of financial institutions bankruptcy based on support vector machine, analyzed the data of financial statements publicly disclosed by banks, and predicted the number of bankruptcies of American financial institutions from 2007 to 2013. The model shows 99.2% prediction accuracy [20].

Rtayli and Enneya proposed an enhanced credit card risk identification method based on random forest classifier and support vector machine feature selection algorithm to predict fraud risk. Experimental results show that the classification performance of the algorithm is better than that of local outlier factor, isolated forest, and decision tree algorithms on large datasets [21]. Plawiak et al. proposed deep genetic hierarchical learner network (DGHLN) algorithm, which is an excellent learner training method based on genetic hierarchical training. 21% of the credit rate in Germany was verified by cross validation [22].

2.3. Current Situation and Background of Related Research in China. China's credit reporting system has not yet entered a mature stage. At present, less than 50% of the population in China can generate credit report in the People's Bank of China, which limits the accuracy of the traditional credit scoring card model in assessing the credit risk of the lender. With the advent of the Internet era in recent years, big data and artificial intelligence technologies have gradually developed and spread in domestic financial market risk control, making up for the lack of credit data. By analyzing the borrower's Internet information and converting it into feature vector, ML algorithm is used to predict the potential default risk. The success of this model in Internet financial risk control has attracted domestic scholars' in-depth research. For example, Hou and Liu applied the support vector machine nonlinear classifier to the bank credit risk assessment and analyzed and compared the experimental results with different kernel functions and parameters [23].

Subsequently, Hou and Xue used the approximate support vector machine (PSVM) model in ML principle to conduct an empirical analysis on the personal housing loan data of a commercial bank in Xi'an market. The results show that the accuracy of the model in predicting the credit risk of individual housing loans of commercial banks reaches 87.5% [24]. Hu et al. established the credit risk assessment model under the supply chain finance mode by using support vector machine. By comparing with the model established by principal component analysis and logical regression method, it is confirmed that the credit risk assessment system based on SVM is more effective and superior [25].

Based on the idea of data mining, Zhao and Chen used customer credit consumption behavior data and rough set theory to reduce the condition attributes in the decision table, constructed a decision tree algorithm based on variable precision weighted average roughness and Gini index, and predicted the default repayment of customers according

to the decision attribute value. The experimental results show that the improved dynamic early warning model of credit card consumption credit risk based on rough set and decision tree algorithm is often better than the basic statistical model and ML algorithm in terms of accuracy and stability [26]. Liu and Tang used the area under the ROC curve AUC value as the classification performance index of the binary classification algorithm, constructed a feature selection algorithm AUCRF based on random forest algorithm, and made an empirical analysis of Australian credit data in UCI ML database. The results show that the model based on AUCRF algorithm can obtain higher classification performance with smaller feature subset, AUC = 0.9346 [27].

2.4. ML in Credit Scoring for Internet Financial Risk Management. Note that many methods and technologies for risk control in the financial field have been proposed in the existing literature, including traditional methods for credit risk management in Internet finance. However, this paper first mainly introduces ML algorithm into Internet finance credit risk management. We can verify the innovation of this paper by comparing the existing research results of credit risk management with the contents of this paper.

This study uses "Internet financial credit scoring," "ML in Credit Scoring," and "application ML and Internet financial risk control" as keywords to search. The search scope is review articles on financial risk management published from 2010 to 2020. The study selected peer-reviewed journals and conference articles because of their high quality. We choose the article by reading the conclusion and abstract, and sometimes we need to read the whole article. All unpublished work and dissertations are not included in this current study. Other existing literatures include systematic research on bankruptcy forecasting or the use of credit scoring models, as well as the application of ML in traditional financial field. Table 1 lists the literature investigated and does not mention the application of ML algorithm in the field of Internet financial risk management.

Through literature review, it can be seen that focusing on the research of ML algorithm applied in traditional credit scoring model in the field of Internet financial credit risk management research is insufficient. However, the traditional credit evaluation methods have limitations in multidimensional and large-scale data analysis, and the model method has strict limitations in distribution hypothesis and linearity. It is difficult for Internet credit data to meet the requirements of the traditional model. The ML algorithm based on big data and artificial intelligence can make accurate analysis and prediction of multisource and multitype data and has developed rapidly. Traditional risk measurement methods predict the future default risk based on the borrower's historical data and personal characteristics, while ML algorithm has extensive expansion in the dimension of obtaining information, which can deeply analyze the correlation between such information and default risk based on behavioral information, soft information, and hard information. In the current research, there are few literatures comparing the traditional credit evaluation model and ML model, and the research on the integration of the two methods to evaluate the credit risk of Internet finance is relatively rare. Therefore, on the basis of reading the relevant literature at home and abroad, this paper uses ML algorithm to construct the credit risk model, verifies the performance of ML model better than the traditional credit score card model through empirical verification, makes a deep discussion on how to convert the ML model into the score card model, and puts forward suggestions on the construction of risk control system of Internet financial industry by ML.

3. Model and Evaluation Metric

In this part, the algorithms of credit scoring model and ML model will be discussed. In addition, some evaluation indexes about the performance of the model are introduced. The function of this part is to lay the foundation for the case study and empirical analysis in the next section.

3.1. Credit Scoring Model. Credit scoring is a supervised learning method, which is essentially a binary classification. According to the historical data characteristics of customers of various categories, a mathematical model is established to predict the default risk of lenders according to "good borrowers" and "bad borrowers" [36]. Because of its strong interpreting ability, logistic regression (LR) is the most commonly used model in credit scoring. The formula of logistic regression model is as follows:

$$P(y=-1|x) = 1 - P(y=+1|x) = \frac{\exp\left(a_0 + a^T x\right)}{1+\exp\left(a_0 + a^T x\right)},$$

$$P(y=+1|x) = \frac{1}{1+\exp\left(a_0 + a^T x\right)}, \tag{1}$$

where $x \in R$ is feature vector; $p(y=+1|x)$ is the probability that the eigenvector borrower x is classified as a nondefaulting customer; and $p(y=-1|x)$ is the probability that the eigenvector borrower x is classified as a defaulting customer. $\{a_0, a\}$ represents whether the model parameters are estimated by using, for example, the maximum likelihood estimation of the training dataset [37]. Once the model parameters are estimated, the decision on the eigenvector x is recorded as $\widehat{y} = +1, \quad$ if

$$P(y=+1|x) \geq P(y=-1|x). \tag{2}$$

According to the above calculation of customer credit evaluation process, credit decision rules can be summarized as follows:

$$\widehat{y} = \begin{cases} +1, & \text{for } 1 \geq \exp\left(a_0 + a^T x\right), \\ -1, & \text{otherwise.} \end{cases} \tag{3}$$

3.2. XGBoost Integrated Learning Method. Qi de et al. proposed the XGBoost algorithm [38] solving real-world

Table 1: Existing literature surveys on financial risk management and their differences from this survey paper.

Survey paper	Articles searched	Objective	Difference from this paper
[28]	165	Research on the bankruptcy risk of financial institutions Investigation of model type by decade Compare model performance	Internet financial risk management was not mentioned The application of ML algorithm was not involved Credit risk was not considered
[29]	214	Research on the application of traditional credit card scoring model Comparing models based on performance	Deep learning models were not covered
[30]	130	ML in financial crisis prediction Focus on private enterprise	Credit risk was not covered Not related to financial industry
[31]	Not specified	Comparing models based on credit rating	Internet financial risk management was not mentioned Lack of evidence to prove that it is advisable to introduce ML algorithm into credit card scoring model
[32]	187	Proposing a new method for scoring Compare traditional techniques Conceptual discussion	Deep learning models were not covered Internet financial risk management was not included
[33]	6	Focus on bankruptcy prediction Models are compared based on design, datasets, and baselines	Internet financial risk management was not included Credit risk was not covered
[34, 35]	49	The search for models to predict the prices of financial markets	Internet financial risk management was not included Credit risk was not covered

classification problem. They posit that XGBoost is an optimized version of gradient boosting machine. The main improvement on GBDT is the normalization of the loss function to mitigate model variances. This also reduces the complexities of modelling and hence the likelihood of model overfitting [39]. Meanwhile, the conventional method uses decision trees as a classification basis. In contrast, XGBoost supports linear classifiers, applicable not only to classifications but also to linear regressions. The traditional approach only deals with the first derivative in learning but XGBoost improves the loss function with Taylor expansion. While the level of complexities increases for the learning of trees, the normalization prevents the problems associated with overfitting [40].

The algorithm has unique advantages in sparse data processing, approximate tree building, and parallel computing, which makes ML technology widely used in mechanical engineering, rail transit, automation technology, and other fields [41]. XGBoost is a gradient lifting ensemble algorithm based on decision tree and linear model. Its basic idea is to combine some decision tree models to form a model with high accuracy. If we give the data as $(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \quad i = 1, 2, \ldots, n, x_i$ represents the independent variable, and y_i represents the a dependent variable. The calculation steps are as follows:

$$\widehat{y}_i^t = \sum_{k=1}^{t} f_k(x_i) = \widehat{y}_i^{(t-1)} + f_t(x_i), \tag{4}$$

where $\widehat{y}_i^t$ is the predicted value of the model in the round t and XGBoost model algorithm is formed by continuous iteration, and each iteration is trained by adding a lesson of decision tree to the prediction value $\widehat{y}_i^t$ of the previous round. In general, the formula of the objective function is as follows:

$$\text{obj}(w) = L(w) + \Omega(w), \tag{5}$$

where w is the parameter to be estimated, $L(w)$ is the loss function, and $\Omega(w)$ is the regularization term. Therefore, minimizing $\text{obj}(w)$ is the criterion for selecting $f(x)$.

$$\begin{aligned}\text{obj}^{(t)} &= \sum_{i=1}^{n} l\left(y_i, \widehat{y}^{(t)}\right) + \sum_{i=1}^{t} \Omega(f_i) \\ &= \sum_{i=1}^{n} l\left(y_i, \widehat{y}^{(t-1)} + f(x)\right) + \Omega(f_i) + \text{constant}.\end{aligned} \tag{6}$$

Taylor expansion is used to expand the approximate objective function and remove the constant term. The final objective function is as follows:

$$\begin{aligned} g_i &= \partial_{\widehat{y}(t-1)} l\left(y_i, \widehat{y}^{(t-1)}\right), \\ h_i &= \partial^2_{\widehat{y}(t-1)} l\left(y_i, \widehat{y}^{(t-1)}\right), \\ \text{obj}^{(t)} &= \sum_{t=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f)_t. \end{aligned} \tag{7}$$

In XGBoost algorithm, the following improvements will be made: the decision tree is divided into the structure part Q of the tree and the weight (fraction) part w of the leaf node.

$$f_t(x) = w_{q(x)}. \tag{8}$$

Moreover, the complexity of the tree is redefined as

$$\Omega(f)_t = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2, \tag{9}$$

where T represents the number of leaf nodes. Under these new definitions, the new form of objective function is

$$\begin{aligned}\text{obj}^{(t)} &= \sum_{t=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f)_t,\\ &= \sum_{t=1}^{n}\left[g_i w_{q(x)} + \frac{1}{2}h_i w_q^2(x)\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2,\\ &= \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i w_j + \frac{1}{2}\left(\left(\sum_{i\in I_j} h_i\right) + \lambda\right)\right)w_j^2\right] + \gamma T.\end{aligned} \tag{10}$$

If $G_j = \sum_{i\in I_j} g_i$, , $H_j \sum_{i\in I_j} h_i$, the objective function can be further rewritten as

$$\text{obj}^{(t)} = \sum_{j=1}^{T}\left[G_i w_j + \frac{2}{2}(H_i + \lambda)w_j^2\right] + \gamma T. \tag{11}$$

After the objective function is obtained, the optimal value of w_j can be obtained by finding the reciprocal of w_j and making it equal to zero:

$$w_j^* = -\frac{G_j}{H_j + \lambda}. \tag{12}$$

Substituting equation (12) into the objective function, we can get

$$\text{obj} = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T. \tag{13}$$

3.3. Evaluation Metric. For the traditional credit scoring model, in order to improve the speed and accuracy of calculation, we need to select variables before establishing the model. The choice of variables is based on their information value, which is abbreviated as IV. Information value describes the importance of the contribution of variables to the prediction results of the model. We choose to add variables with high IV value to the model, while variables with too small IV value will not be added to the model. If we want to calculate IV, firstly need to calculate the WOE, that is, the weight of evidence. WOE is a form of encoding the original independent variable. If you want to code a variable, you need to first group the variable (also known as discretization, boxing, etc.); after grouping, the calculation formula of WOE for group I is as follows:

$$\begin{aligned}\text{WOE}_i &= \ln\left(\frac{py_i}{pn_i}\right)\\ \text{IV}_i &= (py_i - pn_i)\times\text{WOE}_i,\\ &= (py_i - pn_i)\times\ln\left(\frac{py_i}{pn_i}\right),\\ &= \left(\frac{y_i}{y_T} - \frac{n_i}{n_T}\right)\times\left(\frac{y_i/y_T}{n_i/n_T}\right),\\ \text{IV} &= \sum_{i}^{n}\text{IV}_i,\end{aligned} \tag{14}$$

where Py_i is the proportion of bad samples to all bad samples in this group, Pn_i is the proportion of good samples to all good samples in this group, y_i is the number of bad samples in this group, n_i is the number of good samples in this group, y_T is the number of all good samples in the sample, and n_T is the number of all bad samples in the sample.

For the machine learning model, there are many evaluation indexes, and the commonly used indexes are accuracy rate, true positive rate, false positive rate, accuracy rate, F1 score, etc., which are shown in Table 2. We can also construct confusion matrix based on these indicators, which is shown in Table 3. In addition, we also draw the receiver operating characteristic (ROC) curve and the Kolmogorov–Smirnov (KS) curve of the subjects to reflect the performance of the model more vividly.

AR measures the overall predictive effectiveness of model; however, it is not a reliable parameter as it yields misleading results if the dataset is not balanced. The parameters mentioned above are calculated based on the confusion matrix shown in Table 1. True positive (TP) refers the number of defaults that are correctly predicted as defaults; false positive (FP) refers the number of nondefaults that are mistakenly predicted as defaults; true negative (TN) refers the number of nondefaults that are correctly predicted as nondefault; false negative (FN) refers the number of defaults that are mistakenly predicted as nondefaults. In addition to these evaluation indexes, there are two very important ML model prediction performance indicators, such as AUC and KS curve.

3.3.1. ROC and AUC. When the output of the model classifier is continuous, the AUC value can be used as the evaluation standard, and its value range is AUC∈[0, 1]. If we use f to represent a classifier, "$x\_$ "to represent negative samples and" X_+" to represent positive samples, the output result of f is $(x\_) < f(x_+)$, the ROC curve of the classifier passes through the point (0, 1), and the corresponding AUC value is 1. The AUC value of the normal classifier is between 0.5 and 1; if the AUC value of a classifier is lower than 0.5, it means that it is not as good as random guess.

The AUC value is defined as the whole area value under the ROC curve (shown in Figure 1). ROC curve can be obtained by confusing TPR and FPR of matrix. With FPR as the horizontal axis and TPR as the vertical axis, we can obtain the corresponding sensitivity and specificity by giving thresholds. Sensitivity means the probability of a major classification determined as a major classification while specificity means the probability of a minor classification determined as a minor. Assuming that we have a large number of adjustable thresholds, we can get a sensitivity-specificity correlation diagram. That is to say, ROC curve is the trajectory of sensitivity and specificity under different thresholds. The closer the inflection point of ROC curve is to the upper left corner, the larger the area under the curve is, indicating that the model has better effect. On the contrary, the closer the inflection point is to the diagonal line from the upper right to the lower left, the smaller the area under the curve is, indicating that the model is less effective.

Table 2: Evaluation index of ML model.

Evaluation index	Formula
Accuracy rate (AR)	(TP + TN)/(TP + TN + FN + FP)
Recall rate (Recall)	TP/(TP + FN)
Or TPR	—
FPR	FP/(FP + FN)
Precision rate	TP/(TP + FP)
F1 score	2PR/(P + R)

Table 3: Confusion matrix.

		Predicted value		Total
		1	0	
Actual value	1	True positive rate (TP)	False negative rate (FN)	TP + FN
	0	False positive rate (FP)	True negative rate (TN)	FP + TN
Total		TP + FP	FN + TN	TP + FN + FP + TN

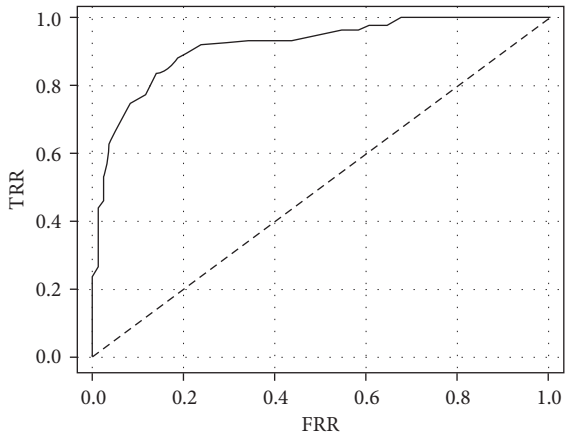

Figure 1: ROC curve graph.

Generally, the AUC value is a probability value to judge whether a model is good or bad. In this paper, we will judge the advantages and disadvantages of the binary classification prediction model with the help of AUC evaluation value. Its evaluation ability is shown in Table 4.

3.3.2. KS Curve. KS curves (shown in Figure 2) are TPR and FPR curves formed under different threshold levels, which are mainly used to verify the distinguishing ability of the model. In the financial risk control, the credit scoring system is constructed, and KS value is often used to measure the performance of the risk control model. Through KS value, we can measure the distinguishing ability of the model from the maximum distance between the cumulative percentage function curve between the correctly predicted borrowers who have not defaulted and the incorrectly predicted borrowers who have overdue. The discrimination ability of KS value is shown in Table 5.

Table 4: AUC evaluation ability classification.

AUC value	Evaluation ability
0.9–1	High accuracy
0.7–0.9	Some accuracy
0.5–0.7	Low accuracy
0–0.5	Not in conformity with the actual situation

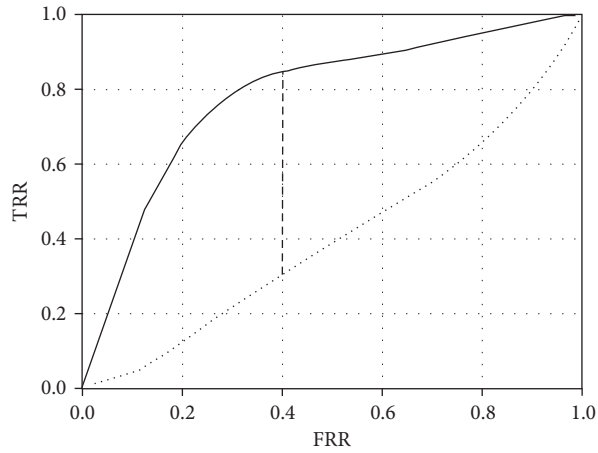

Figure 2: KS curve graph.

Table 5: KS value and model customer identification ability.

KS	The ability to identify risk
<0.2	Null
0.2–0.4	A little
0.41–0.5	Strong
0.51–0.6	Stronger
0.61–0.75	Strongest
>0.75	Abnormal

4. Case Study

In this study, we chose a large P2P Internet lending platform in China as the research case. We analyze the data of 30225 short-term loans from August to December 2018. According to the different performance of the borrowers, they are divided into different categories: D0 represents the borrowers who are not overdue, that is, to repay the principal and interest within the loan term; D1 is overdue less than one month; although they did not repay the loan on time, the overdue time was not too long; if a borrower is overdue for more than one month, D2 is used. For short-term loans, the overdue days are considered as serious overdue because once the customer exceeds these time, the possibility of reperformance is relatively small. The reason for the above classification of borrowers is to better carry out the following analysis and model construction.

In order to facilitate the construction of the evaluation model, a total of 24180 borrowers were classified as D0 and D1 ("good borrowers") and 6045 borrowers were classified

as D2 ("bad borrowers"), which accounts for nearly 80% and 20% of the total sample, respectively. The sample information includes six dimensions including basic information of borrowers such as education, income, age, gender, and so on; credit card transaction records such as billing information and repayment information; debit card payment information; associated loan information; e-commerce platform transaction information; and telecomputer operator information. According to the classification of the six dimensions, we subdivide the variables. However, not all the variable information values meet the minimum threshold set by us, that is, the IV value is greater than 0.02, so we finally selected 372 variables to prepare for modelling.

4.1. Credit Scoring Model

4.1.1. Variable Filtering. For the credit card scoring model, generally only 10–15 variables need to be selected to build the model. Then, you need to filter the variables in advance. The standard for selecting variables is the size of their IV values, and variables with too small IV values are not suitable for selection into the model. In this study, the variables whose IV value is greater than 0.05 and WOE trend is monotonous are selected, and the variables whose correlation coefficient is too high are removed. For example, if the label of the variable is the bank number of the borrower and the IV value is 0.056 through calculation, then this variable will be selected to be added to the credit scoring model. However, the IV value calculated by the average amount of each consumption of the borrower in the last 90 days is 0.026, so the IV value of this variable is too small, and this variable will be eliminated when constructing the model. In addition, the calculated number of transactions consumed in the last 30 days shows that the IV value is 0.082, but the WOE trend is inconsistent, so this variable will also be eliminated.

Following the above ideas, according to IV value, WOE trend, correlation coefficient, and business logic principle, 16 variables are finally selected from different dimensions to consider establishing the model, as shown in Table 6.

4.1.2. Credit Scoring Model. We use the method of logistic regression to build the model because it is easy to monitor and deploy, which is a common method to build credit scoring model. Firstly, we check the coefficients of each variable in the logistic regression method, and it is valid only when the coefficients of variables are positive and variables with negative coefficients will be deleted. Secondly, we set the threshold of p value as 0.05, and if the p value of the variable is greater than this significance level, it will be deleted. Finally, using the programming software, we get the credit scoring model and convert the test set samples, and then we get the scores, as shown in Table 7.

In Table 7, the score represents a range of points. GS is the abbreviation for the number of good samples, followed by the rate represents the ratio of good samples. BS is the abbreviation for the number of bad samples, followed by the ratio of bad samples. TS represents the number of all samples and TR represents the ratio of all samples, and BR refers that bad debt rate.

We can see from Table 7 that with the increase of credit score, the number of good samples in the overall sample shows an upward trend, except for slight decrease in individual intervals, while the number of bad samples is decreasing in general. It shows that good samples should get higher credit scores, while bad samples have lower credit scores. In addition, good samples correspond to lower bad debt rate, while bad samples have higher bad debt rate. From the credit score results, the traditional model has some functions in the credit risk of borrowers.

4.1.3. Model Performance Evaluation. The evaluation of credit scoring model is mainly reflected by KS, AUC, GINI, and other indicators. KS evaluates the model's ability to distinguish customers by calculating the maximum difference between the cumulative percentage of bad customers and good customers; AUC is the standard for judging the advantages and disadvantages of classifiers; GINI coefficient is used to evaluate the risk differentiation ability of the model (Table 8).

As can be seen from Table 8, the KS scores of different types of datasets in the credit scoring model are between 0.3 and 0.4, indicating that the ability of the model to identify customers is not satisfactory and AUC value is between 0.7 and 0.8, indicating that the classifier is better than random guess, and if the model threshold is set properly, there is a certain predictive value; when the GINI value is about 0.5, it indicates that the risk differentiation ability of the model is acceptable.

4.2. ML Model.

In this section, the methods related to this work are presented in the following four aspects: data cleaning and feature selection, processing of imbalanced dataset, ML algorithm model setting, and analysis of the result yielded by the proposed ML model.

Step 1. Data cleaning and feature selection.

In data cleaning, we focus on two issues: the processing of empty points and the arrangement of outliers. There are usually four methods to deal with empty points: case deletion method, missing data calculation method, machine learning method, and model-based process [42]. In this study, we mainly deal with empty values based on experience. Specifically, we will delete the features that more than 95% of borrowers did not fill in. At the same time, we will add new features to describe the remaining features. If it is empty, use "1″"; otherwise, use "0." In addition, the average values of these features are calculated to fill the empty points. As far as outliers are concerned, it has been proved that using filters on outliers can improve model performance [43]. By referring to other studies [8], we detect outliers manually and keep reasonable outliers. At the same time, the upper and lower values of the box chart are used to replace the abnormal values. In addition, the feature values are standardized and scaled so that they fall within the specified range of [0, 1].

Table 6: Variables selected into the model.

Bank of deposit	0.053
Number of credit cards	0.052
Maximum credit card limit in recent 1 month	0.084
Maximum overdue days of short-term loans	0.286
The salary per month	0.249
The standard deviation of the number of SMS messages sent at night in the last three months	0.072
The standard deviation of the frequency of answering unlabeled numbers at night in recent two months	0.075
Debit card ratio	0.073
Bill number	0.069
Amount to be paid under credit products	0.065
Average consumption in recent 30 days	0.063
Total data months	0.068
The proportion of credit cards with bills in the last 60 days	0.066
Balance of credit products	0.062
Percentage standard deviation of dialing all numbers at night in recent 60 days	0.060
Bank of deposit	0.061

Table 7: Test set sample score.

Score:	GS	GS rate (%)	BS	BS rate (%)	TS	TR (%)	BR (%)
[low, 575]	366	6.01	542	24.63	908	10.02	51.28
[576, 585]	518	8.31	396	15.32	914	9.65	32.32
[586, 595]	722	9.35	345	15.2	1067	10.86	30.19
[595, 601]	699	9.35	263	10.36	962	9.52	21.25
[602, 608]	885	10.52	132	9.52	1017	10.27	18.77
[609, 615]	887	10.36	106	7.02	988	9.65	14.38
[616, 622]	896	10.89	96	5.99	992	9.68	12.35
[623, 630]	902	11.96	90	5.26	992	10.05	10.37
[631, 641]	932	11.68	68	3.88	1000	10.01	7.85
[642, high]	941	12.05	35	1.38	976	9.68	2.73

Table 8: Result evaluation of credit scoring model.

Data classification	KS	AUC	GINI
Training set	0.3628	0.7269	0.4696
Validation set	0.3265	0.7225	0.4426
Testing set	0.3269	0.7244	0.4535

The next work is feature selection which could improve the operation efficiency and the prediction result of classifier. Generally, subset selection can be used to improve the performance of feature selection process, such as wrappers, filters, and embedding method [44, 45]. According to reference [8], this study adopts a tree-based feature selection method, which is an embedded method, namely, feature selection based on random forest model [46]. Random forest can be used not only to calculate the importance of different features but also to delete irrelevant features.

Step 2. Processing of imbalanced data.

Most of the studies on credit risk assessment models for Internet financial institutions are based on imbalanced data, which means the number of nondefault cases is usually larger than the default ones; if we ignore the class imbalance problem to buid a classification model, we might obtain a model that has high accuracy for the determination of nondefaults but extremely low accuracy for default. To solve this issue, this paper tries to deal with SMOTE algorithm.

In the dataset of this study, because the sample size of defaulting borrowers accounts for less than 10% of the total sample, it belongs to unbalanced sample. If the misjudgment rate is used as the evaluation index of the model, the data in this paper may have a relatively large risk, and it is impossible to get a valuable model. The SMOTE algorithm artificially synthesizes new samples based on a small number of samples, and adds the synthesized new samples to the data set. The basic idea of SMOTE algorithm is to find the distribution space of small class samples according to the partial characteristics of two kinds of samples in p-dimensional space and finally generate new small class samples between small class samples and small class samples. Referring to [47], the algorithm flow is as follows:

(i) Taking Euclidean distance as the standard, for each sample $\$ \times \$$ in a small sample class, the distance from it to all samples in the minority sample set S_min is calculated, and its k-nearest neighbour is obtained.

(ii) The sampling rate is set according to the sample imbalance ratio to determine the sampling rate N.

(iii) For each minority sample $\$ \times \$$, it is randomly selected from its k-nearest neighbours, if the selected nearest neighbour is $/hat{$x$}$.

(iv) For every randomly selected nearest neighbour $/hat{x}$, build new sample by the formula

$$x_{\text{new}} = x + r \,\text{and}\, (0, 1) \times (\tilde{x} - x). \tag{15}$$

The specific idea of the algorithm is shown in Figure 3.

By using SMOTE algorithm, the data distribution in this paper tended to be balanced. The ratio of expected default to expected nondefault is 1 : 1.33, which makes the sample category basically balanced.

Step 3. ML model setting.

In this section, we employ grid search to set a series of hyperparameters, which is a fundamental parameter optimization method. And it will substantially divide the hyperparameter into the grids with same length in the certain rang of coordinate system. Every point in the coordinate system represents a set of hyperparameters, and then we could adopt every point in a certain interval into our model to verify the performance of the algorithm. The point that performs best is called best hyperparameter. In other word, the algorithm of grid search is to traverse the points corresponding to all grids.

Grid search was used to optimize the combination of hyperparameters within 5 cross-validations [47]. Since grid search uses an exhaustive search of predefined hyperparameter space, we provide the search space for these algorithms here: number of iterations was set in the range of 100 to 500, the depth of trees is in the range of 5 to 25, and learning rate is in the mathematical set of (0.001,0.01,0.1,1). The lowest gradient descent of loss function is set to 0. Besides this, SONNIA (2016) was used to generate the SOMs in this work [48]. Parameters were set as default.

Step 4. Results and analysis.

The data processing speed of machine learning is calculated by programming: "start = time. Perf _ counter ( ), End = time. Perf_counter ( ), T = end-start," and we get the speed of the ML model to deal with selected variables is 9 milliseconds which is very fast.

The application of ML model should consider the actual business situation of the organization. For Internet financial institutions, setting up a strict preloan approval system and granting loans only to customers with high credit scores can reduce the credit risk to some extent, but it will lead to a large number of customers unable to carry out transactions due to lack of qualifications, which will affect their business results. Based on this situation, this experiment considers different prediction results of ML model under different preset probabilities, as shown in Table 9.

In Table 9, PP is an abbreviation for the preset probability value. GS represents cumulative good samples, and GB represents cumulative bad samples. The passing rate is

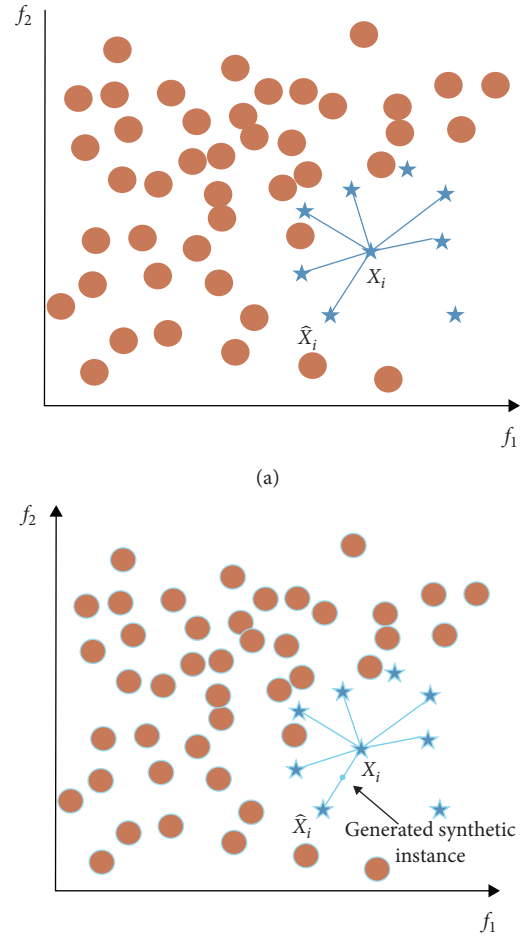

Figure 3: SMOTE algorithm principle.

expressed by PR, and ER stands for the error rate. The values of each item are reserved with four digits after the decimal point. It can be seen from Table 8 that the model can achieve the highest KS value of 0.4936 with the preset probability of 0.6~0.65, which means that a loan customer can pass the screening only when the probability of predicting a good customer is greater than 0.6. Under this standard, 59.98% of the applicants have passed the loan application. However, the error rate of the model is 7.44%, which means that 7.44% of the bad samples are wrongly judged as good samples.

Because the low pass rate will also affect the financial performance, the operators engaged in the Internet finance industry should comprehensively measure and compare the KS value, pass rate, and misplacement rate from the perspective of realizing business and then choose a preset probability threshold that best meets its operating conditions. By comparing Tables 8 and 9, we can see the difference between the traditional credit scoring model and the ML model. Under the preset probability of 0.6, the KS value of traditional credit scoring model is 0.3269, while the result of ML model is 0.4936. This shows that under this preset probability, the prediction ability of ML model is obviously better than that of traditional credit card scoring model.

Table 9: ML model prediction results.

PP	GS	DS	KS	PR	ER
[0.50, 0.55]	0.1444	0.5269	0.4165	0.7023	0.0992
[0.55. 0.60]	0.1818	0.5986	0.4930	0.6032	0.0861
[0.60, 0.65]	0.2635	0.6320	0.4936	0.5998	0.0744
[0.65, 0.70]	0.2895	0.6598	0.4360	0.5366	0.0634
[0.70, 0.75]	0.3265	0.7998	0.3963	0.5183	0.0588
[0.75, 0.80]	0.3984	0.7911	0.3641	0.4880	0.0481
[0.80, 0.85]	0.4698	0.8698	0.3201	0.4609	0.0307
[0.85, 0.90]	0.5024	0.9269	0.3004	0.3504	0.0287

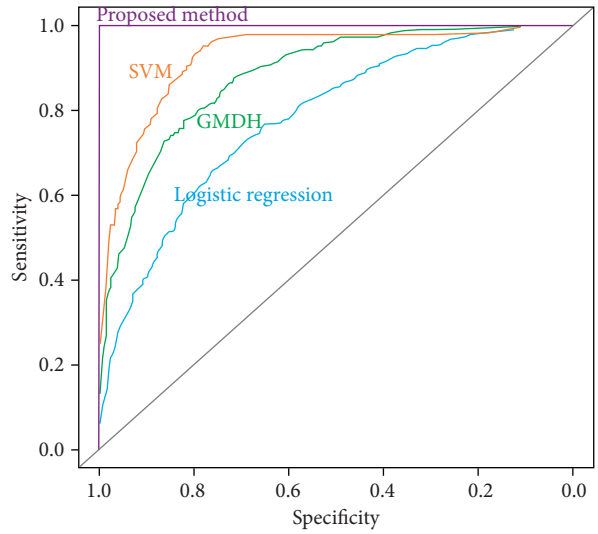

Figure 4: ROC from same training data on different classifiers.

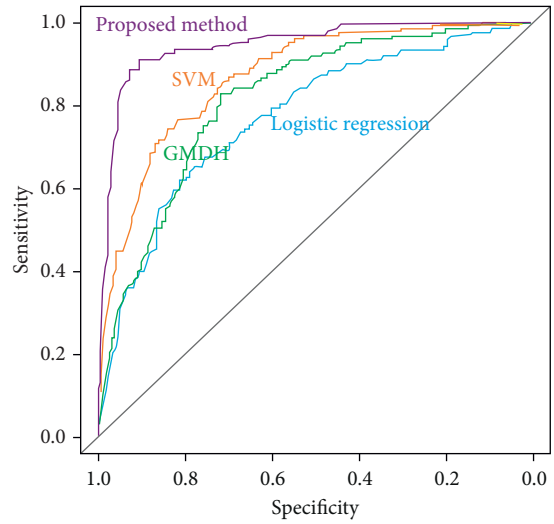

Figure 5: ROC from same test data on different classifiers.

4.3. Validation of the XGBoost ML-Based Model in This Paper. In order to verify the effectiveness of ML algorithm in the credit risk management of Internet financial industry, this paper selects a large Internet financial lending platform in China as a research case and compares the performance of the traditional model and the model proposed in this paper. In order to further enhance the comparability of the model, more methods are introduced to compare the simulation and experimental results. We compare the results of credit scoring model which based on logistic regression, neural network method and support vector machine learning method for data grouping processing [49] with the results of the method proposed in this study [50]. Referring to other research ideas [30], we compare the simulation figures of

Table 10: Comparison of accuracy rates yielded by different classifiers.

Model	Cutoff point	Training data			Test data		
		Classification	Discriminant accuracy	Total accuracy	Classification	Discriminant accuracy (%)	Total accuracy (%)
Logistic regression	0.35	Nondefaults Defaults	72.9 70.1	71.5	Nondefaults Defaults	69.7 70.5	70.1
GMDH	0.41	Nondefaults Defaults	85.1 73.7	79.4	Nondefaults Defaults	75.1 75.1	75.1
SVM	0.26	Nondefaults Defaults	88.3 77.9	83.1	Nondefaults Defaults	78 76.7	77.4
Proposed method	0.6	Nondefaults Defaults	100 100	100	Nondefaults Defaults	89.6 90.6	90.1

different experimental results. Figures 4 and 5 show the ROC based on training set and test set data, respectively, and Table 10 shows the classification accuracy, which comes from the optimal cutoff point 0.6 when default accuracy equates to nondefault accuracy based on test data.

We can see from Figures 4 and 5 that the AUC value of XGBoost classifier is the best based on the same test data. At the same time, Table 10 shows that the overall accuracy rate of the proposed Internet financial risk assessment model is the best (90.1%), which is better than the traditional logistic regression model (70.1%), support vector machine (77.4%), and GMDH (75.1%). When dealing with the same dataset and training set, the performance of this method is better than other classifiers.

5. Conclusions

In this paper, we propose an improved ML-based technique for credit card scoring in Internet financial risk control, which has better performance than the traditional credit scoring modern in Internet financial risk control. Because the traditional credit evaluation model is complicated and has strict research on the selection of variables, it has some limitations. And this method has strict data requirements that in the Internet age, there is a limitation that it cannot analyze the personal credit data with high dimension, high complexity, and nonlinearity. However, with the deep integration of Internet era and traditional financial industry, the vigorous development of Internet financial industry is the inevitable trend of social development. At the same time, financial institutions engaged in Internet financial business will process a large number of customer data, which is more important for the control of credit risk. Therefore, we must consider which method to use to carry out the credit risk of the Internet financial industry, and ML algorithm has become a good alternative. The main contribution of this paper is to propose the application of ML algorithm to financial risk control in the field of Internet finance because it can show better performance than the traditional credit scoring model and better match with the background of big data. Therefore, this paper has a certain reference value for the risk management practice of the Internet financial industry.

The proposed ML model is tested on an Internet financial platform in China. The experimental results indicated that the process of building up model and dealing with data is more efficient. Compared with the traditional credit scoring model, ML algorithm can process a large number of data in a very short time to meet the requirements of Internet financial institutions to process a large number of customer information. In addition, there are no strict restrictions on the data processed by ML algorithms. In order to improve the performance of the model prediction results, we can set model parameters in advance, add variables to the model, and then eliminate the variables that contribute less to the model according to the importance of features. The experimental results show that only when the probability that a loan applicant is predicted to be a good customer is greater than 0.6, can the loan application be screened. At this time, the KS value obtained by the ML model is 0.4936, which exceeds the KS value of the traditional credit scoring model of 0.3269. This indicates that the ML model has certain advantages in the application of Internet financial risk control.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] S. Yu, B. Bing, H. Peikai et al., "The study of the tourism enterprises' financing Capacity under the background of internet, travel and finance commune," *Finance Account*, vol. 11, no. 24, pp. 3-4, 2017.

[2] J. Zhang and Q. Sun, "Research on financing cost of small and medium-sized enterprises by internet finance Open," *Open Journal of Social Sciences*, vol. 48, no. 25, p. 95, 2017.

[3] Y. Lin and C. Chen, "Research on enterprise financial risk evaluation based on association rules," *Friends Account*, vol. 96, no. 2365, pp. 32–35, 2017.

[4] J. LiuJ.A. Xiuyi et al., "Multi-label classification algorithm based on association rules mining," *Journal of Software*, vol. 28, no. 11, pp. 2865–2878, 2017.

[5] Nonymousm and Liu, "Research on bank product recommendation model based on big data mining in fintech era China," *Finance Computer*, vol. 233, no. 4356, pp. 38–40, 2018.
[6] J. Zhao, "Internet Finance and its risk prevention and control," *Tax and Economy*, vol. 23, no. 2336, pp. 52–63, 2018.
[7] L. C. Thomas, J. Crook, D. Edelman et al., "Credit scoring and its applications," *Society for Industrial and Applied Mathematics*, vol. 579, no. 3487, pp. 221–225, 2012.
[8] B. Wang, L. Ning, Y. Kong et al., "Integration of unsupervised and supervised machine learning algorithms for credit risk assessment," *Expert Systems with Applications*, vol. 128, pp. 301–315, 2019.
[9] S. Firdous and R. Farooqi, "Impact of internet banking service quality on customer satisfaction," *Journal of Internet Banking and Commerce*, vol. 121, no. 2201, pp. 1–17, 2017.
[10] H. R. Khedmatgozar and A. Shahnazi, "The role of dimensions of perceived risk in adoption of corporate internet banking by customers in Iran Electron," *Electronic Commerce Research*, vol. 2, no. 18, pp. 389–412, 2018.
[11] H. A. AbdouGenetic, "programming for credit scoring: the case of Egyptian public sector banks," *Expert Systems with Applications*, vol. 9, no. 36, pp. 11402–11417, 2009.
[12] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring: a review," *Journal of the Royal Statistical Society: Series A*, vol. 3, no. 160, pp. 523–541, 1997.
[13] S. Lee, C.-C. Chiu, Y.-C. Chou, and C.-J. Lu, "Mining the customer credit using classification and regression tree and multivariate adaptive regression splines," *Computational Statistics and Data Analysis*, vol. 163, no. 2569, pp. 1113–1130, 2006.
[14] W. Y. Bee, H. O. Seng, N. Huselina Mohamed Husain et al., "Using data mining to improve assessment of credit worthiness via credit scoring models," *Expert Systems with Applications*, vol. 38, no. 10, pp. 13274–13283, 2011.
[15] G. Williams and Z. Huang, "Mining the knowledge mine: the hot spots methodology for mining large real world databases," in *Proceedings of the 10th Australian Joint Conference on Artificial Intelligence*, Perth, Australia, November 1997.
[16] A. Yeo, K. Smith, R. Willis, M. Brooks et al., "Clustering technique for risk classification and prediction of claim costs in the automobile insurance industry," *Intelligent Systems in Accounting Finance*, vol. 322, no. 2986, pp. 39–50, 2001.
[17] E. Khandani, A. J. Kim, A. W. Lo et al., "Consumer credit-risk models via machine-learning algorithms," *Journal of Banking & Finance*, vol. 344, no. 4356, pp. 2767–2787, 2016.
[18] C. Chakraborty and A. Joseph, "ML at central banks," *Bank of England Staff Working*, vol. 674, 2017.
[19] J. Ticknor, "Bayesian regularized artificial neural network for stock market forecasting," *Expert Systems with Applications*, vol. 14, no. 40, pp. 5501–5506, 2013.
[20] T. P. Gogas and A. Agrapetidou, "Forecasting bank failures and stress testing: a ML approach," *International Journal of Forecasting*, vol. 34, no. 13, pp. 440–455, 2018.
[21] N. Rtayli and N. Enneya, "Selection features and support vector machine for credit card risk identification," *Procedia Manufacturing*, vol. 46, pp. 941–948, 2020.
[22] P. Pławiak, M. Abdar, J. Pławiak et al., "DGHNL: A new deep genetic hierarchical network of learners for prediction of credit scoring," *Information Sciences*, vol. 516, pp. 401–418, 2020.
[23] H. Hou and S. Liu, "Credit risk assessment of commercial banks based on support vector machine," *Computer Engineering and Application*, vol. 31, no. 40, pp. 176–178, 2004.
[24] J. Hou, Q. Xue, P. Xu et al., "Prediction of credit risk of personal housing loan by approximate support vector machine," *Journal of Xi'an University of Technology*, vol. 6, no. 31, pp. 559–565, 2011.
[25] H. Hu, L. Zhang, D. Zhang et al., "Research on credit risk assessment of supply chain finance based on support vector machine," *Soft Science*, vol. 5, no. 25, pp. 26–30, 2011.
[26] T. Zhao and W. Chen, "Credit risk re assessment of retail business of commercial banks based on credit consumption behavior," *Financial Theory and Practice*, vol. 12, no. 37, pp. 75–79, 2016.
[27] X. Liu, J. Tang, Z. Duan et al., "Research on feature selection of AUCRF algorithm in credit risk evaluation," *Computer Applications and Software*, vol. 4, no. 35, pp. 293–295, 2018.
[28] J. L. Bellovary, D. E. Giacomino, M. D. Akers et al., "A review of bankruptcy prediction studies: 1930 to present," *Financial Education*, vol. 42, no. 33, pp. 1–42, 2010.
[29] H. A. Abdou and J. Pointon, "Credit scoring, statistical techniques and evaluation criteria: a review of the literature," *Account and Financial Manage*, vol. 18, no. 3, pp. 59–88, 2011.
[30] Y.-C. Chang, K.-H. Chang, G.-J. Wu et al., "Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions," *Applied Soft Computing*, vol. 73, pp. 914–920, 2018.
[31] W. Lin and Y. Hu, "ML in financial crisis prediction: a survey," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 4, no. 42, pp. 421–436, 2012.
[32] X. Wang, M. Xu, Ö.T. Pusatli et al., "A survey of applying ML techniques for credit rating: existing models and open issues," *Neural Information Processing*, vol. 98, no. 56, pp. 122–132, 2015.
[33] F. Louzada, A. Ara, G. B. Fernandes et al., "Classification methods applied to credit scoring:Systematic review and overall comparison," *Surveys in Operations Research and Management Science*, vol. 2, no. 21, pp. 117–134, 2016.
[34] S. Devi and Y. Radhika, "A survey on ml and statistical techniques in bankruptcy prediction," *Knowledge-Based Systems*, vol. 22, no. 67, pp. 120–127, 2018.
[35] M. BrunoHenrique et al., "Literature review: ML techniques applied to financial market prediction," *Expert Systems with Applications*, vol. 124, pp. 226–251, 2019.
[36] Credit Risk Scorecards, *Developing and Implementing Intelligent Credit Scoring*, SAS Publishing, Cary, NC, USA, 2005.
[37] I. J. Myung, "Tutorial on maximum likelihood estimation," *Mathematical Psychology*, vol. 1, no. 47, pp. 90–100, 2003.
[38] F. A. Qi de, Xu L. Cheng, Z. Zhu et al., "Xgboost recommendation algorithm based on collaborative filtering," *Computer Application Research*, vol. 5, no. 37, pp. 1317–1320, 2020.
[39] L. Song, S. Wang, C. Yang et al., "Application of improved XGBoost in unbalanced data processing," *Computer Science*, vol. 6, no. 47, pp. 98–103, 2020.
[40] Liu and Chen, "Loan risk prediction method based on SMOTE and XGBoost," *Computer and Modernization*, vol. 2, pp. 26–30, 2020.
[41] Y. Cui, W. Qi, H. Pang et al., "Recommendation algorithm combining collaborative filtering and xgboost," *Computer Application Research*, vol. 1, no. 12, pp. 62–65, 2020.
[42] P. J. García-Laencina, J. L. Sancho-Gómez, A. R. Figueiras-Vidal et al., "Pattern classification with missing data: a review," *Neural Computing and Applications*, vol. 2, no. 19, pp. 263–282, 2010.
[43] V. García, A. I. Marqués, J. S. Sánchez et al., "On the use of data filtering tech- niques for credit risk prediction with

instance-based models," *Expert Systems with Applications*, vol. 39, no. 18, pp. 13267–13276.

[44] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2011.

[45] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 19, no. 23, pp. 2507–2517, 2007.

[46] L. Breiman, "Random forest," *Machine Learning*, vol. 46, no. 327, pp. 1–35, 1999.

[47] X. Pan, "Prediction and analysis of credit risk of P2P online loan borrowers," Guizhou University of Finance and Economics, Guiyang, China, 2019pp. 34-35, Master thesis.

[48] Sonnia, *Molecular Networks GmbH: Germany and Altamira*, vol. 2, LLC, New York, NY, USA, 2016, https://www.mn-am.com/products/sonnia.

[49] A. G. Ivakhnenko, "The group method of data handling–a rival of the method of stochastic approximation," *Soviet Automatic Control*, vol. 13, pp. 43–55, 1996.

[50] B. E. Boser, I. M. Guyon, V. N. Vapnik et al., "A training algorithm for optimal margin classifiers," in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pp. 144–152, ACM Press, New York, NY, USA, July 1992.