Hindawi
Journal of Mathematics
Volume 2021, Article ID 8541929, 10 pages
https://doi.org/10.1155/2021/8541929

Research Article

Analysis of Internet Financial Risk Control Model Based on Machine Learning Algorithms

Mingjin Liu,[1] **Ruijie Gao**,[2] **and Wei Fu**[3]

[1]*Information Technology Center, Sichuan Water Conservancy College, Chengdu 611231, China*
[2]*School of Marxism, Sichuan Water Conservancy College, Chengdu 611231, China*
[3]*Organization Department, Sichuan Water Conservancy College, Chengdu 611231, China*

Correspondence should be addressed to Mingjin Liu; liumingjin@swcvc.edu.cn

Received 5 November 2021; Revised 30 November 2021; Accepted 1 December 2021; Published 20 December 2021

Academic Editor: Miaochao Chen

On the basis of traditional credit risk control, this paper proposes the demand and direction of a new credit risk control strategy based on machine learning and relying on big data. First, on the basis of introducing the basic algorithmic principles of machine learning, we give reasons for choosing machine learning models and build a machine learning-based Internet consumer finance credit risk control strategy model to provide theoretical support for the empirical analysis later. Second, we take the test data of Internet consumer finance S company as the research sample and carry out empirical analysis according to the machine learning-based Internet consumer finance credit risk control strategy model. The comparison of the training results is based on the comprehensive consideration of training time, validation set accuracy, TPR evaluation indicators, and interpretability of the results; it verifies the advantages of the machine learning model in screening the key influencing factors that cause the overdue performance of credit customers. According to the optimized credit risk control strategy, corresponding strategy suggestions are provided for the credit risk control of S company. The research results show that the prediction effect of the classification model based on traditional linear regression is generally lower than that of the model based on the classification algorithm based on machine learning, and there is a complex nonlinear relationship between platform default and its related influencing factors. The accuracy of classification and early warning results of the random forest algorithm is relatively high, and the detection rate of the decision tree model is relatively high, but the cost is also the highest. In addition, the accuracy of the four types of early warning models is relatively stable, reaching an average of 80%. This paper proposes a machine learning-based Internet consumer finance credit risk control strategy model. Its system, timeliness, and risk prediction capabilities provide new ideas and suggestions for Internet consumer finance companies to design risk control strategies.

1. Introduction

Under the influence of concepts such as “Internet+” and “inclusive finance,” the traditional financial industry is rapidly integrated with the current advanced Internet technology, and the financial industry is also continuously being segmented. Among them, based on real consumption scenarios, Internet consumer finance has achieved rapid development in recent years and has broad market prospects [1]. However, due to the widespread moral hazard, adverse selection, shortcomings in personal credit investigation, and insufficient collection of multidimensional data before lending in the Internet finance industry, the rapid development of Internet consumer finance will bring various risks. Establishing a strong risk model and having a scientific risk control strategy constitute an urgent problem to be solved in the development of consumer credit business under the background of the Internet [2–5].

The role of Internet financial platforms in the development of the real economy is self-evident, but the chaos in the online lending industry not only makes investors miserable, but also disrupts the stability of the financial market and the real economy. There are many factors that aggravate the risk of platform issues, which may include external factors such as an unsound legal environment, lack of a social credit system, and macroeconomic and industrial economic downturns and internal factors such as financial problems, risk control and mismanagement, fraud, and self-financing of funds [6]. How to autonomously control the suitable sensor to move to the appropriate position in time to deal with the dynamic and uncertain environment will be an extremely important part of tracking. The deterioration of the online lending platform problem not only hinders the virtuous development of the online lending industry, but also disrupts the financial order and social and economic stability. Therefore, it is necessary to formulate and improve online lending laws and regulations, strengthen the supervision of the online lending industry, regulate market behavior, and control the deterioration of the problem. Grasping the key influencing factors of platform risk and establishing an efficient platform risk early warning model have become two important tasks for strengthening supervision [7–9].

Based on Internet information collection and processing technology, this paper initially established a dynamic early warning system for Internet financial platforms with classification algorithms as the core, using different classification algorithms and horizontally comparing their learning efficiency and accuracy. Classification algorithms have unique advantages when studying risk warning problems with structural changes. One does not need to make any assumptions about the data. The analysis is completely based on the data itself, and it has good adaptability to changes in the data structure. At the same time, using crawler data to establish a systematic and dynamic early warning system based on the classification algorithm model is an important supplement to the previous platform risk research literature. Combining various indicators of model performance, the logistic regression model is more effective in identifying default risks than the decision tree model and the naive Bayes model, and the accuracy rate is relatively higher when using the test set for testing. In addition, in terms of practice, this article can provide regulatory agencies with prompt warning signals with high prediction accuracy. It also helps ordinary investors avoid problem platforms and promotes the healthy development of the online lending industry.

2. Related Work

The evaluation model selected in the context of big data generally has the characteristics of a large number of indicators and poor distinguishing ability of individual indicator variables. However, existing research generally shows that the number of indicators is small and the number of strong classification indicators is large. In addition, the research on the application of big data is more about the business model under the background of big data. Most of this research uses the method of qualitative description to discuss the application of big data in all walks of life and seldom chooses quantitative analysis [10].

Leo et al. [11] conducted related research on Internet finance related concepts. They believe that Internet finance is a brand-new financial service model, which is a financial activity that exists under the background of communication technologies such as electronics and computers. They believe that there is a certain degree of pre-lending credit risk. The above comes from the financial platform's insufficient grasp of various dimensions of credit customers' data. They also point out that multidimensional and multidepth data should be collected before lending to judge the credit risk of credit customers. Jaroszewski et al. [12] believe that the risk control of consumer credit in the context of Internet finance can learn from the risk control model of traditional financial institutions. For example, in the control of moral hazard, the group joint responsibility system adopted, through mutual supervision between members, promotes the timely repayment of group members and plays a certain role in reducing the default rate of credit customers. Brătăşanu [13] believes that the credit risk of Internet finance can be controlled and points out that information asymmetry is an important source of credit risk and that the level of credit risk can be reduced by effectively solving this problem. Based on the empirical analysis of the logistic regression model, among the many influencing factors that affect customers' overdue behavior, the level of interest rates has an important impact on whether customers have overdue behavior, and the two show a positive relationship. Gupta et al. [14] believe that the symmetry of information between the platform and credit customers will affect the results of lending and point out that the more sufficient the information provided by the credit customers, the greater the probability that the platform will grant loans. Ghoddusi et al. [15] explained the reasons leading to personal consumer credit risks from three aspects: credit customers, banks, and social system environment. They pointed out that the credit scoring model is insufficient in credit customer credit risk assessment and believed that the ROC curve can be borrowed as an evaluation classification model.

Researchers used the binary logistic regression model to empirically analyze the overdue behavior of credit users and found that the factors that have a significant impact on the overdue behavior of credit customers are credit customer's credit rating, historical overdue times, residence, income, etc. They believe that Internet finance is a general term for various financial activities carried out on the Internet as a carrier. It is different from the traditional paper-based carrier and is more conducive to breaking through the limitations of time and space, diversifying the options of various parties, and reducing transaction costs [16]. In response to the country's strategic goal of "developing inclusive finance," it builds an inclusive financial system with national characteristics and summarizes three basic propositions, namely, the inclusiveness, tolerance, and diversity of subjects of inclusive finance. Scholars summarized the current seven characteristics of Internet finance, such as companion business platform, active cross-border, extensive connection, and intensive account farming, and pointed out that Internet finance will have more new features in the future. Some scholars believe that Internet consumer finance is a kind of nondeposit credit business carried out by traditional financial institutions or mutual finance companies. At the same time, they point out that consumer finance and rapid development play an important role in national economic development, and a series of suggestions are given on how to achieve rapid development, such as improving the system, innovating products, and preventing and controlling risks [17–20].

3. Internet Financial Risk Control Model Construction Based on Machine Learning Algorithms

3.1. Machine Learning Algorithm Architecture. In the field of machine learning, confusion matrix is widely used, which can also be called possibility table or error matrix. It is a specific matrix, which is mainly used to evaluate whether the model training performance is ideal or not. The rows of the matrix represent the actual categories, and the columns represent the predicted categories given by the model. Its advantage lies in the fact that it can intuitively reflect whether there is confusion in the category prediction result of the classification model; that is, an object that originally belongs to a category is classified into another category. Figure 1 illustrates the machine learning algorithm architecture.

For a classification model, in order to obtain an optimal result, it is often necessary to find out the best parameter setting corresponding to the model through multiple tests, then each parameter setting will correspond to a different result, this result defines it as the objective function, and the optimal result we hope to achieve is to make the objective function obtain the optimal value.

$$\begin{gathered} y(n) = A \cdot X(n-1) - t \cdot x(n-1), \\ \nabla^2 \frac{\alpha - 1}{a(a+1)} \frac{\nabla^2 \alpha}{\nabla t^2} + \frac{\rho}{\varepsilon} = 0. \end{gathered} \tag{1}$$

It seems that the loss function is an indicator that can be used to measure the quality of a model. However, if only this indicator is used, there will be a problem that there is a risk of overfitting; the robustness of the model is reduced in the process of training, so that the model has a very bad performance when predicting a new data set. In view of this, we consider adding a second auxiliary indicator, namely, regularization.

$$h(x) = \begin{cases} \varepsilon(i,n) \times x(t), & 0 < t \leq 1, \\ \delta(i,n) \times x(t), & t > 1, \\ y(n) \times x(t), & t \leq 0. \end{cases} \tag{2}$$

TP is the number of samples where the model prediction result is positive and the actual result is also positive; FP is the sample number where the model prediction result is positive but the actual result is negative; TN represents the sample size that the predicted results of the model are consistent with the actual results; FN represents the sample size of the model predicted result contrary to the actual result. Obviously, FP and FN belong to the performance errors of model prediction.

$$\begin{gathered} \min\left\{\lim_{N \longrightarrow \infty} \sum_{i=1}^{N} a_i\right\} + \frac{1}{2} \lim_{N \longrightarrow \infty} \sum_{i=1}^{N} \sum_{j}^{N} a_i y_j a_j y_i k\left(x_i^2, x_j^2\right) = 0, \\ f(x_i, x_2, x_3) = \prod_{i=1}^{3} G(F(x_i); W). \end{gathered} \tag{3}$$

Among them, the true positive rate is also called sensitivity. Its calculation formula is the number of positive examples correctly predicted by the classification model divided by the total number of actual positive examples, that is, the proportion of positive samples correctly predicted by the classification model to all positive samples. The ultimate goal is to measure the prediction level of the classification model for the positive samples. The specificity calculation formula is the number of negative cases correctly predicted by the classification model divided by the total number of negative cases; its ultimate purpose is to measure the prediction level of the classification model for negative samples.

3.2. Internet Finance Pattern Recognition. Based on the Internet financial model, a flaw of the dynamic training set data in this paper is that the "time stamp" information of a few features related to the platform cannot be reliably obtained, such as the disclosure of negative news on the platform and the external capital increase of the platform. These indicators may only be announced after negative news on the platform or capital increase. In addition, investors' impression scores on various platforms are also dynamically changing. We cannot obtain the specific time of such information, and it is difficult to realize the implementation of such information. If the model simply and rudely eliminates such indicators, it will affect the algorithm's impact on the platform. The variable importance measurement is proven to be a robust statistic. This article first establishes a random forest model based on platform data, then evaluates the classification effect of the random forest model, and finally compares it with the linear probability model and stacked noise reduction autoencoders described later. At the same time, the random forest model variable important measurement method is used to grasp the key influencing factors of the platform problem and provide variable screening for the model later. Figure 2 shows the layout of the Internet finance model.

We can repeat the process to get K models, and the final prediction can use the combined results of K models (for example, for classification problems, the classification result

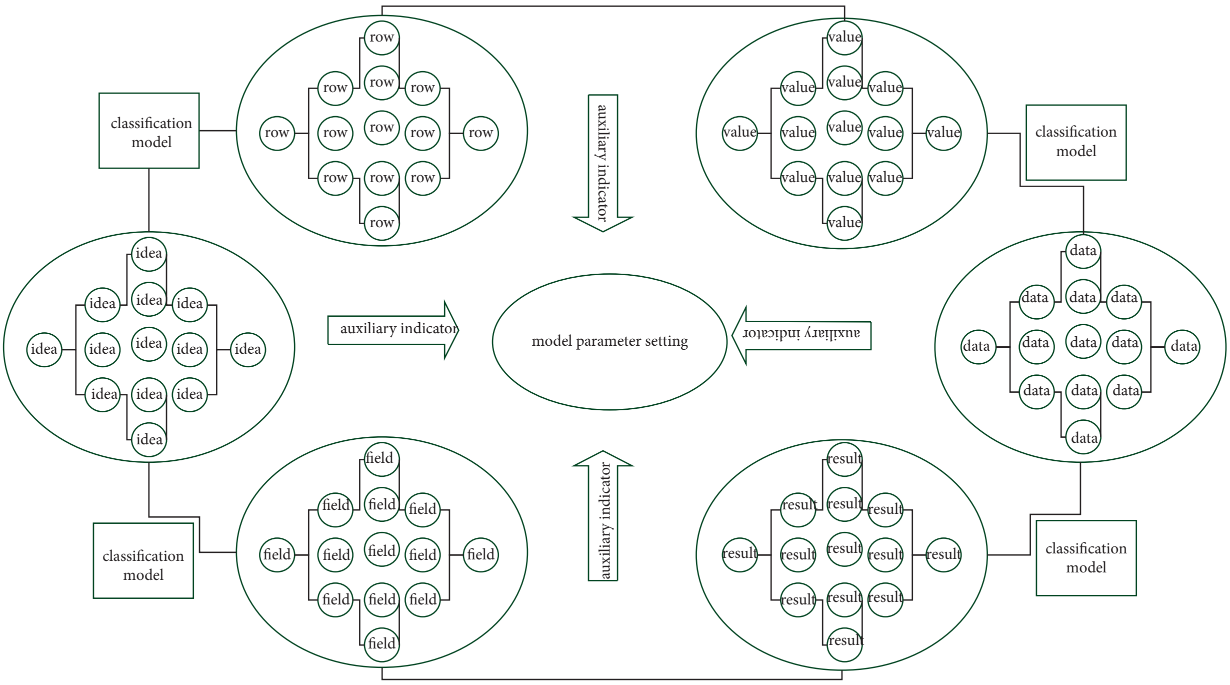

Figure 1: Machine learning algorithm architecture.

with the most votes can be used, and for regression problems, the average of K prediction results can be used). Compared with the limited sensing ability of single sensor, k-model multiple sensors can perform tasks in parallel. The final prediction result depends on the comprehensive judgment of K doctors, so the random forest method has natural advantages over other methods. The specific process of the random forest algorithm is as follows: We use the sample data to construct a classification tree, randomly select z features from all the features of the platform sample to produce a decision tree, and then split them to generate K independent decision trees. There is no need to split the decision tree. Then, combined into a random forest, the classification trees generated by the model are random and independent of each other, so when combining them, you can assume the same weights. When the random forest algorithm model deals with problems similar to the division of online lending platforms, it obtains the prediction results through voting by all decision tree classifiers in the model.

3.3. Financial Risk Control Evaluation Indicators. The results of the financial risk control evaluation model have a good analytical nature. The dependent variable of the model can be a binary variable, and there is no restriction on the measurement of the independent variable. It can be a categorical variable or a numerical variable. The model can be used to select indicators that affect the interpreted scalar from a large number of explanatory variables for in-sample regression, and it can also achieve classification for out-of-sample prediction. The advantage of the logistic regression method is that the model itself is relatively simple and interpretable and has a certain degree of scalability, but it has the assumption that the features are independent of each other.

However, the specific use of funds is difficult to control. The pure credit Internet consumer finance operation model transfers consumer loans in the form of future accounts receivable to the Internet wealth management platform. The wealth management platform packages them into Internet wealth management products to provide investors with investment. The credit platform in this mode is an intermediary platform. It provides investors' funds to consumers and allows them to pay in installments. Finally, the consumer uses the interest-bearing loan to repay the borrower according to the stipulated time. Judging from the current development of pure credit Internet consumer finance, the risks of consumer credit are mainly borne by Internet wealth management companies. Internet finance companies mainly use big data, cloud computing, and other Internet technologies to innovate financial solutions to maximize the control of bad debt risks. Figure 3 shows the distribution of the feature node index for optimal matching.

We divide all the indicators according to the target classification in the training set and find the optimal feature of the best match, that is, the root node (the first level indicator). According to the "root node," the training set is divided into two categories, and the best matching second-level nodes (second-level indicators) are searched again and recursively until the data is divided and the specified conditions are met. After the data samples of the test set enter the decision tree, the decision tree is traversed according to the value of each index of the sample, so as to achieve the final classification of the prediction. The advantage of the decision tree method is that the principle is simple, the calculation speed is fast, and it can generate a classification path that is in line with people's common sense and is easy to understand. The model is not sensitive to missing values; the disadvantage is the instability

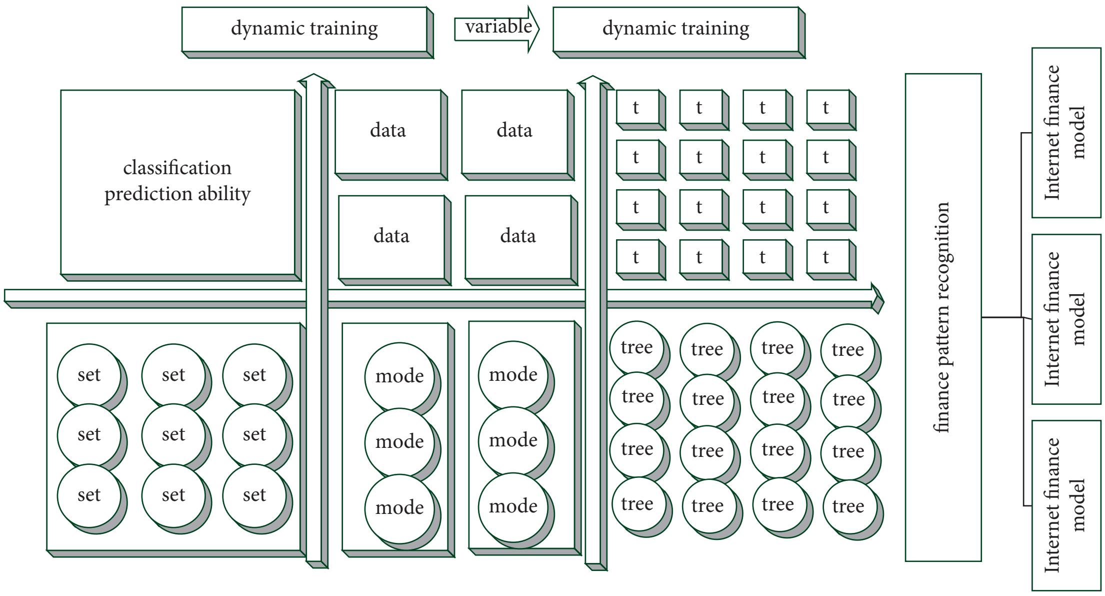

Figure 2: Layout of Internet finance mode.

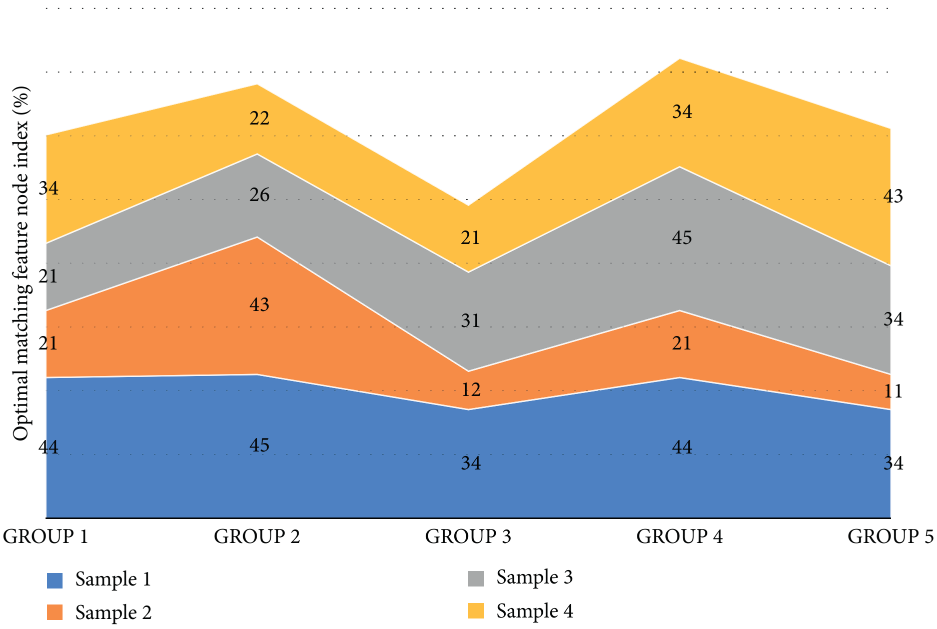

Figure 3: The distribution of feature node indicators for optimal matching.

of the decision tree itself, and the shape of the tree may have some influence on sample selection. Strong sensitivity, according to different sampling samples, may produce different decision tree forms.

3.4. The Weight Iteration of the Risk Control Model. In the wind control model, multiple input information is input to a "neuron," and a value between 0 and 1 is calculated through a certain weighted formula. The input information of the neurons in the next layer is calculated again until the final layer, and the predicted input is obtained. The neural network can be single-layered or multilayered, and there can be multiple nodes in each layer. The initial weight is given randomly, and the weight is continuously adjusted by comparison with the correct result to minimize the error of the result.

The BP neural network does not need to give the specific function relationship between the input layer and the output layer in advance; that is, it does not need to give the specific mathematical expression of the activation function in advance. BP neural network mainly uses external input samples and constantly changes the link weight of the network layer to reduce errors and improve accuracy, thereby automatically learning a model, explaining the internal connection between the input layer and the output layer, and obtaining a desired output for all inputs values. The BP neural network also overcomes the limitations of some traditional statistical methods in dealing with nonlinear problems, such as normal distribution characteristics for variables, and a clear linear function relationship between independent variables and dependent variables. This model is particularly suitable for dealing with complex internal mechanisms. Figure 4 shows the weight distribution of the risk control model. In addition, the five pie shapes in the figure benefit from the distributed sensing characteristics of MSN, and the largest proportion of MSN has a wider sensing range than the other four single-sensor nodes.

The new type of Internet consumer finance credit risk control is based on the use of machine learning technology to achieve risk control in the context of big data. It needs to collect more dimensional and in-depth data variables and identify strongly related risk factors. Because the current way of thinking of humans has certain difficulties in understanding nonlinear relationships, machine learning theory can help humans solve such difficulties. Therefore, in the current actual risk control scenarios, computer technology can be used to grasp these relationships at relatively low cost. These information data can accurately reflect consumers' credit level and risk-bearing ability, before financial institutions and consumers have financial business relations, that is, in the pre-lending link, use this as a reference and decision-making basis. How to achieve risk control by relying on machine learning in the context of big data requires the use of machine learning techniques to obtain and process Internet user behavior data, and then these super-large samples of nontraditional variables are reviewed from the credit risk control review. The causality is extended to the correlation between variables.

4. Application and Analysis of Internet Financial Risk Control Model Based on Machine Learning Algorithms

4.1. Machine Learning Algorithm Data Feature Extraction. Since there are 66 training sets and test sets in the dynamic early warning process of this article, the sample data is dynamically changing, and the number of variable indicators is too large to be dynamically descriptive. AIC data has higher data redundancy and stronger adaptability and robustness, that is, the failure of a single sensor does not affect the detection task of the entire system. When the model becomes more complicated as the number of parameters increases (k increases), the likelihood function will also increase, and the calculated AIC value may become smaller.

At this time, the model may be overfitting and the AIC becomes larger. The principle of the AIC information criterion is to select the model with the smallest AIC. While paying attention to the degree of fit, the number of parameters (penalties) should be reduced as much as possible to reduce the possibility of overfitting of the model. For the sample training of the Internet financial platform training set, the logistic model finally selected 8 features. This paper uses the vif() function in the car package in the R language to test the model and finds that the logistic model does not have collinearity of the independent variables. The prediction of the logistic model can output the probability of the category to which the record belongs. This article has conducted multiple tests, applied the principle of conservativeness, and set the model threshold to 0.2. When the classification probability is greater than 0.2, the platform is classified as category 1; otherwise, it is classified as category 0 (normal platform). Figure 5 indicates the data feature division of machine learning algorithms. The data line and the histogram in the figure interact and cooperate with each other to complete tasks that a single sensor cannot do.

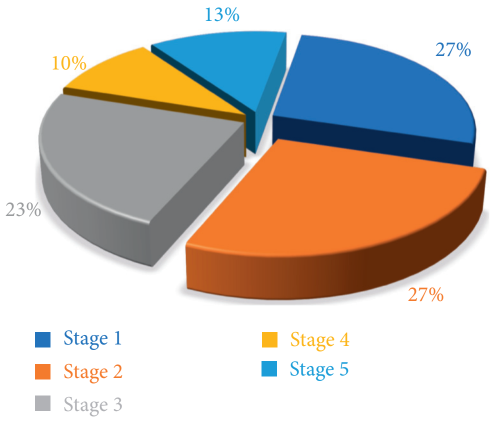

Figure 4: The distribution of the weight of the risk control model.

It can be seen that in the 200 test sets, 108 defaulters were correctly predicted not to default, and 14 people did not default, but at the same time, 75 nondefaulters were misclassified as defaulters, and 3 defaulters were misclassified as nondefaulters. For defaulters, the model prediction is unsatisfactory. It can be seen that each coefficient of the new model is very significant. At the same time, the new model L2 also passed the chi-square test, indicating that the fit of the new model is very good, and the new fitted model is easier to interpret due to the reduction of variables. Since the prediction result of the logistic regression model is a specific probability value, we use the nature of the log probability function itself to convert the probability values greater than 0.5 into 1, and the probability values less than 0.5 are converted to 0, so that the results of the binary classification can be compared with the test set. It can be seen from the results that in the 200 test sets, 182 defaulters were correctly predicted not to default, and 7 people did default. However, at the same time, 10 defaulters were misclassified as nondefaulters, and 1 nondefaulter was misclassified as defaulter. Misclassification is a defaulter, the accuracy of the model is $A = 94.5\%$, the precision rate is $P = 94.9\%$, and the recall rate is $R = 99.5\%$. It can be seen that the prediction effect of the model is relatively good.

4.2. Internet Finance Risk Control Model Simulation. Based on the customer data of a credit card center of a commercial bank A, the application data of 200,000 real users were selected, the bank's own risk assessment model was used for evaluation and screening, and then the CreditNet model was used for further evaluation on this

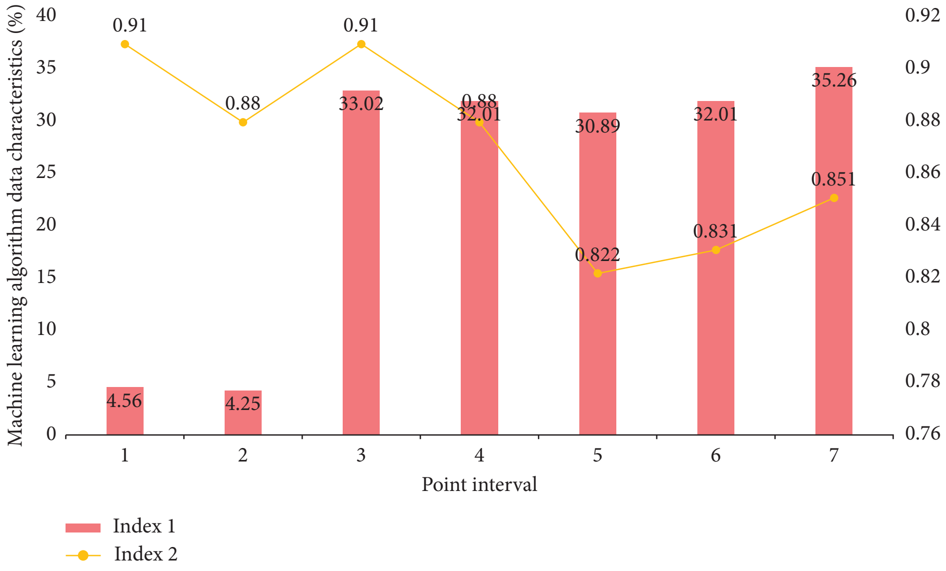

Figure 5: Machine learning algorithm data feature division.

basis. The horizontal and vertical axes of the ROC curve are the false positive rate (FPR) and true positive rate (TPR) mentioned above. As for how the curve is made, it will be explained in detail below. First, we set the output result of the classification model as the probability of belonging to the positive sample.

Calculate the corresponding false positive rate and true positive rate, and finally take each generated TPR and FPR as a point and connect them in the coordinate axis to get the ROC curve corresponding to the classification model. If the number of hidden layer units is greater than the number of input layer and output layer units, and the activation function is a nonlinear sigmoid function, then SDA can map the input data to a high-dimensional nonlinear separable space, which is more conducive to finding the optimal classification hyperplane. The nonlinear function of SDA can effectively approximate the nonlinear relationship between data. Each influencing factor has not only a linear impact but also a nonlinear impact on the risk of platform problems. SDA can excellently describe the interwoven nonlinear relationship within the variable system to obtain the best fitting effect. Figure 6 shows the risk control factor curve of Internet finance. The three curves represent different meanings. Among them, curve 3 shows the largest, curve 1 the smallest, and curve 2 the medium impact factor. The single target state is represented by lowercase letters (such as x).

It is easy to see from the coordinate axis that there are 3 special points on the ROC curve. The first is when FPR = 0, TPR = 1, and the corresponding point is (0, 1); it represents FN = 0 and FP = 0; this result shows that the classification model is perfect because it divides all samples correctly. The second is when the false positive rate = 1, the true positive rate = 0, and the corresponding point is (1, 0). The classification model at this time classifies all samples incorrectly and has the worst classification effect. The third is when FPR = 1, TPR = 1, and the corresponding point is (1, 1); the classification model will predict all samples. At this time, the classification model predicts all samples into a negative category. Therefore, when using the ROC curve to evaluate the classification effect of the classification model, you can judge the classification effect of the classification model by observing how close the ROC curve is to the point (0, 1) in the upper left corner. When it is not possible to visually determine which classification model has better classification effect by observing the ROC curve, you can use a quantitative indicator, AUC, which is a value between 0 and 1, which represents the lower part of the ROC curve. The value of AUC is usually between 0.5 and 1, and the corresponding classification model equal to 0.5 belongs to a random classifier. Figure 7 is a comparison of the prediction quantization error of the machine learning models.

From the comparison of model prediction errors, it can be seen that the prediction error of the Probit model is as high as 7.6%, the prediction error of the random forest is 3.12%, and the prediction error of the stacked noise reduction autoencoder is the lowest, only 3%. Model comparison results show that stacked noise reduction autoencoding has the best data fitting effect and the highest prediction accuracy. Compared with the Probit model, the stacked noise reduction autoencoder can better fit various influencing factors due to its stronger learning ability and adaptive ability, better robustness and fault tolerance, and better nonlinear characterization ability.

With the increase in the number of hidden layer units, the error of the test set has a tendency to rise, and the error of the training set has a tendency to decrease. This shows that as the number of hidden layer units increases, the stacked noise reduction autoencoder is in danger of overfitting. The number of optimal hidden layer units appears at 20. After multiple training times, the optimal number of hidden layer units is 20, and the optimal number of hidden layers is 1. As

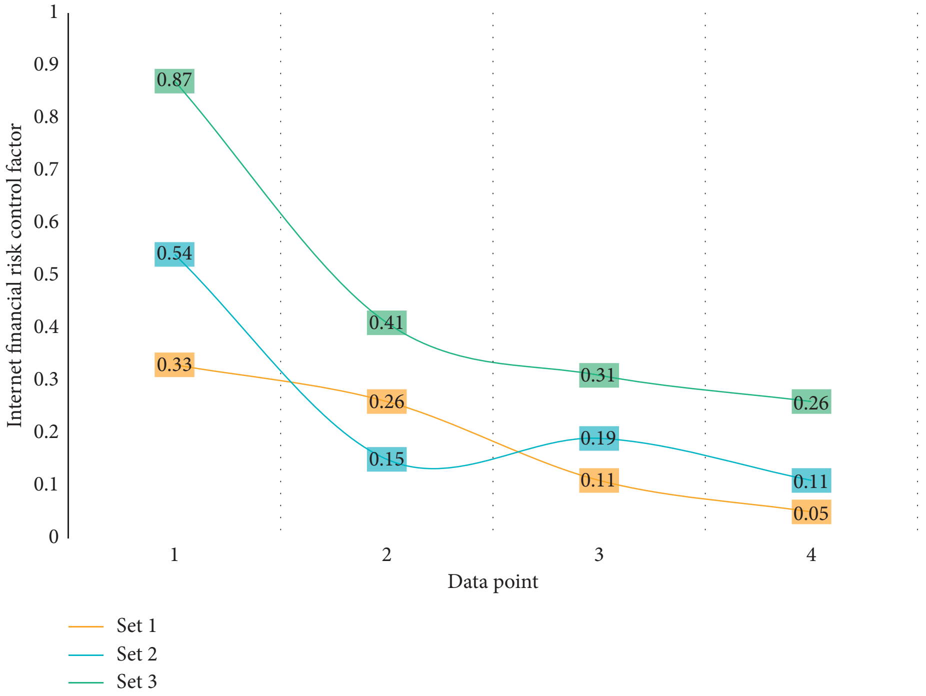

Figure 6: Internet finance risk control factor curve.

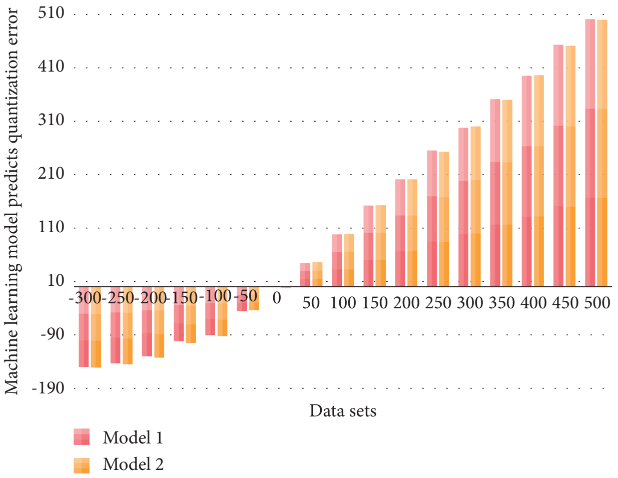

Figure 7: Comparison of prediction and quantization errors of machine learning models.

shown in the text, the stack-type noise reduction autoencoder has insufficient learning of the problem platform due to the imbalance of the target variable data, and the fitting of the problem platform is obviously weaker than the fitting of the normal platform. However, the overall prediction accuracy, recall rate, and F1 score of the model reached 97%, and the overall performance of the model was slightly higher than that of the random forest.

4.3. Example Application and Analysis. The realization of this part of the model algorithm mainly relies on the existing packages of Python. We put the preprocessed sample data set into 6 models for training, namely, machine learning model without parameter adjustment, machine learning model with parameter adjustment, logistic regression model without parameter adjustment, logistic regression model with parameter adjustment, SVM model, and Gaussian Bayes model.

In the case of a stationary target, the only uncertainty that needs to be considered is the observation noise. The comparative analysis of training results is mainly carried out from two aspects. On the one hand, starting from adjusting the parameters, see how the machine learning model performs before and after adjusting the parameters. On the other hand, starting from different models, use the training effects of other models. Because predicting a person who should have refused a loan to be lendable will cause greater losses to financial institutions and lending platforms, the probability of loan rejection should be correctly predicted as 0. Figure 8 shows the regression model parameter training accuracy curve.

Statistics show that there is roughly an inverted U-shaped relationship between the number of customers who have consumer credit behavior and their age. In the processed sample data, credit customers are concentrated between 18 and 35 years old. Among them, there is an increasing trend from 18 to 27 years old. At the age of 27, the number of customers reached a peak of 13,270; then, a downward trend began. At the age of 35, it dropped to 5,289, which is a 60% decrease compared to the peak at the age of 27. It can be explained from two aspects. Why are these consumers of consumer credit behavior mainly concentrated in the young group? On the one hand, the young group is more receptive to new things, and it has more reasons to try new things; on the other hand, the young group's consumer demand is also relatively strong.

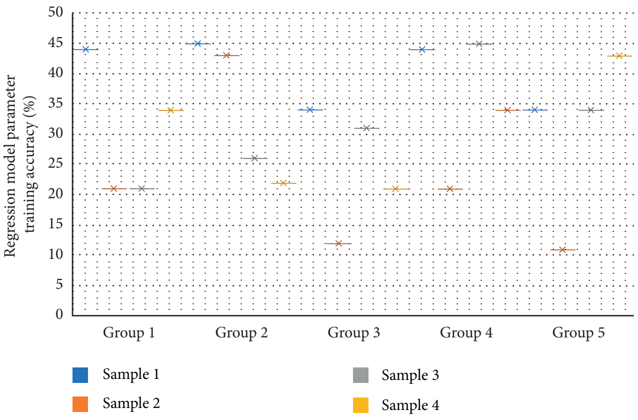

FIGURE 8: Regression model parameter training accuracy curve.

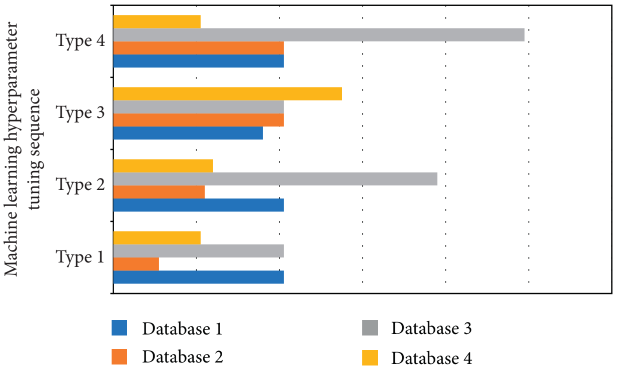

FIGURE 9: Responsiveness distribution of machine learning data series.

Therefore, comprehensively considered, the performance of the sample data is also in line with the actual situation, and young people are more likely to have consumer credit behaviors. Figure 9 is a comparison of responsiveness distribution of machine learning data series.

Due to the complexity of machine learning hyperparameter adjustment, it is necessary for this article to set the hyperparameter value reasonably in combination with actual problems. In this paper, the number of input layer units of the stacked autoencoder is 34, the number of output layer units is 2, the learning rate is 0.01, the sparsity penalty weight is 0.1, the maximum number of iterations is 100, the number of hidden layer units is selected from {20, 25, 30, 35, 40, 45, 50}, and the number of hidden layers is selected from {1, 2, 3, 4}. In this paper, the training set and the test set are divided according to the 3 : 1 random stratified sampling method, the number of samples in the training data set is 642, and the number of samples in the test data set is 213. According to empirical analysis, if the platform-affiliated company is a listed company or a state-owned enterprise, the possibility of platform problems is less than that of a privately owned Internet financial platform.

However, in the case of target motion, we need to consider not only the physical movement of the target, but also the transfer of the target motion mode pair. According to the collected sample data, the private sector, the state-owned sector, and the listed company accounted for 94.040%, 3.311%, and 2.639%, respectively. From the statistical results of the probability of problems, the probability of a problem with the platform of listed companies and state-owned enterprises in the background is 0, and the probability of a problem with the platform of private companies is 0.43. The results of the analysis of variance also explain the platform. The difference in background has a significant impact on the normal and sustainable operation of the platform.

5. Conclusion

Starting from the specific characteristics and relevant determinants of Internet financial platforms, this article uses big data crawling methods to crawl public data of multiple third-party website platforms based on public information on the Internet. The data of the home platform is divided into training set and test set based on the seven aspects of platform strength, product features, safeguard measures, risk control capabilities, network services, information disclosure, and investor impressions, and a dynamic mobile window is set. Next, we use four classification algorithm models, logistic model, decision tree, random forest, and neural network; innovatively adopt dynamic training set and test set methods to study the problem of dynamic screening of national Internet financial platform risks; and compare models horizontally. The experiment uses Python to realize the machine learning model without tuning and the machine learning model training with tuning and lists the basic performance of the model and its interpretability; then, it is compared with the predesigned one. We compare the model without parameter adjustment logistic regression model, parameter adjustment logistic regression model, SVM model, and Gaussian Bayes model, in terms of the training time of the model, the accuracy of the validation set, the TPR index, and the interpretability of the training. The results verify the advantages of the machine learning model in screening the key influencing factors that affect the overdue performance of credit customers. Finally, we combine the machine learning model and the logistic regression model to design a preliminary credit risk control strategy, verify the preliminary strategy to further optimize it, and then design a new credit risk control strategy. Then, we point out the specific research deficiencies and put forward the focus and direction of follow-up research on the application of machine learning models in Internet consumer finance risk control strategies.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

[1] X. Ma and S. Lv, "Financial credit risk prediction in internet finance driven by machine learning," *Neural Computing & Applications*, vol. 31, no. 12, pp. 8359–8367, 2019.

[2] S. Qi, K. Jin, B. Li, and Y. Qian, "The exploration of internet finance by using neural network," *Journal of Computational and Applied Mathematics*, vol. 369, p. 112630, 2020.

[3] X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, and X. Niu, "Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning," *Electronic Commerce Research and Applications*, vol. 31, pp. 24–39, 2018.

[4] M. M. Hasan, J. Popp, and J. Oláh, "Current landscape and influence of big data on finance[J]," *Journal of Big Data*, vol. 7, no. 1, pp. 15–17, 2020.

[5] Y. Liu, J. Peng, and Z. Yu, "Big data platform architecture under the background of financial technology: in the insurance industry as an example," 2018, https://arxiv.org/abs/1901.10527.

[6] J. Zhou, W. Li, J. Wang, S. Ding, and C. Xia, "Default prediction in P2P lending from high-dimensional data based on machine learning," *Physica A: Statistical Mechanics and its Applications*, vol. 534, p. 122370, 2019.

[7] K. Yang, Y. Shi, Y. Zhou, Z. Yang, L. Fu, and W. Chen, "Federated machine learning for intelligent IoT via reconfigurable intelligent surface," *IEEE Network*, vol. 34, no. 5, pp. 16–22, 2020.

[8] S. Gu, B. Kelly, and D. Xiu, "Empirical asset pricing via machine learning," *Review of Financial Studies*, vol. 33, no. 5, pp. 2223–2273, 2020.

[9] L. Munkhdalai, T. Munkhdalai, O.-E. Namsrai, J. Lee, and K. Ryu, "An empirical comparison of machine-learning methods on bank client credit assessments," *Sustainability*, vol. 11, no. 3, p. 699, 2019.

[10] T. M. Ghazal, M. K. Hasan, M. T. Alshurideh et al., "IoT for smart cities: machine learning approaches in smart healthcare-A review," *Future Internet*, vol. 13, no. 8, p. 218, 2021.

[11] M. Leo, S. Sharma, and K. Maddulety, "Machine learning in banking risk management: a literature review," *Risks*, vol. 7, no. 1, p. 29, 2019.

[12] A. C. Jaroszewski, R. R. Morris, and M. K. Nock, "Randomized controlled trial of an online machine learning-driven risk assessment and intervention platform for increasing the use of crisis services," *Journal of Consulting and Clinical Psychology*, vol. 87, no. 4, pp. 370–379, 2019.

[13] V. Brătăşanu, "Digital innovation the new paradigm for financial services industry," *Theoretical & Applied Economics*, vol. 24, 2017.

[14] R. Gupta, S. Tanwar, S. Tyagi, and N. Kumar, "Machine learning models for secure data analytics: a taxonomy and threat model," *Computer Communications*, vol. 153, pp. 406–440, 2020.

[15] H. Ghoddusi, G. G. Creamer, and N. Rafizadeh, "Machine learning in energy economics and finance: a review," *Energy Economics*, vol. 81, pp. 709–727, 2019.

[16] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020.

[17] M. Attaran and P. Deb, "Machine learning: the new 'big thing' for competitive advantage," *International Journal of Knowledge Engineering and Data Mining*, vol. 5, no. 4, pp. 277–305, 2018.

[18] T. A. Tuan, H. V. Long, L. H. Son, R. Kumar, I. Priyadarshini, and N. T. K. Son, "Performance evaluation of Botnet DDoS attack detection using machine learning," *Evolutionary Intelligence*, vol. 13, no. 2, pp. 283–294, 2020.

[19] T. K. Rodrigues, K. Suto, and H. Nishiyama, "Machine learning meets computation and communication control in evolving edge and cloud: challenges and future perspective," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 1, pp. 38–67, 2019.

[20] K. Johnson, F. Pasquale, and J. Chapman, "Artificial intelligence, machine learning, and bias in finance: toward responsible innovations," *Fordham Law Review*, vol. 88, p. 499, 2019.