Effectiveness Evaluation of Big Data Risk Control Models in Consumer Credit: A Comparative Study of Default Prediction Based on XGBoost Algorithm

Jiazhen Lv[1, *]

[1]Zhengzhou University of Aeronautics, Zhengzhou, China

[*]Corresponding author: 2724106623@qq.com

Abstract. With the rapid development of financial technology and the continuous expansion of the consumer credit market, the accurate identification and prediction of credit risks have become crucial for financial institutions in risk management. This paper focuses on credit risk prediction in the consumer credit market, constructing an XGBoost model based on 168,000 pieces of data from Internet finance company S, and comparing it with models such as logistic regression, SVM, and Gaussian Naive Bayes. The results show that the XGBoost model achieves an accuracy rate of 97% on both the training and validation sets, takes only 0.12 minutes, and has a true positive rate (TPR) of 86.04%, with overall performance significantly superior to logistic regression, SVM, and Gaussian Naive Bayes models. Variable contribution analysis further reveals that "new users", "Sesame Credit", "number of shutdown days", and "Ant Credit Pay Limit" are key factors reflecting overdue behaviors in consumer credit. Based on these core influencing factors, this paper designs corresponding risk control strategies to balance risk control and business efficiency. The research provides a theoretical basis and practical guidance for financial institutions to optimize risk control models, helping to reduce credit risks and promote the sustainable development of consumer credit business.

Keywords: Big Data Risk Control; Consumer Credit; Default Prediction; Machine Learning.

1. Introduction

With the rapid development of financial technology, the consumer credit market has shown vigorous growth momentum and has become an important force driving consumption growth and economic development. However, the rapid expansion of consumer credit business has also brought many risks, among which credit risk is particularly prominent[1]. Accurately identifying and predicting borrowers' default behaviors is crucial for financial institutions' risk management. The emergence of big data technology has provided new ideas and means to solve this problem. By collecting and analyzing massive amounts of customer data, financial institutions can more comprehensively understand borrowers' credit status, thereby achieving accurate risk assessment and control[2]. Big data risk control models have emerged as the times require. They use advanced data analysis algorithms to mine valuable credit risk information from massive data, providing support for consumer credit decision-making.

In recent years, the application of big data risk control models in the consumer credit field has gradually received widespread attention. Many scholars and financial institutions have invested in related research to explore how to build efficient and accurate risk control models[3]. In terms of algorithm selection, logistic regression was widely used in the early stage due to its strong interpretability, but it performs poorly in handling nonlinear relationships and high-dimensional data; Support Vector Machine (SVM) has advantages in small sample learning but has high computational costs in large-scale data scenarios; Ensemble learning algorithms such as random forest and Gradient Boosting Decision Tree (GBDT) improve prediction performance by combining multiple weak classifiers and have gradually become one of the mainstream methods in the risk control field[4]. Among them, the XGBoost algorithm, as an improved version of GBDT, has shown significant advantages in prediction accuracy and computational efficiency through optimization strategies such as introducing regularization terms and supporting sparse data processing[5]. It has been applied to

risk prediction in consumer credit scenarios such as credit card default and small loan overdue, and has achieved certain results. Although existing research has achieved some results, existing risk control models still have room for improvement when facing the increasingly complex and changeable consumer credit market and massive unstructured data[6,7].

This paper focuses on the effectiveness evaluation of big data risk control models in consumer credit, especially default prediction based on the XGBoost algorithm. It aims to explore its application value and feasibility in actual consumer credit business, provide a theoretical basis and practical guidance for financial institutions to optimize risk control models, so as to better cope with credit risk challenges in the consumer credit market and promote the healthy and sustainable development of consumer credit business.

2. Research Methods and Model Construction

2.1. Data Set Source and Preprocessing

The data used in this paper comes from the internal team test data of Internet consumer finance company S. The total number of data samples for this analysis is 168,000, with 30 indicators, namely Id, mobile phone number, gender, age, education, registration time, whether it is an old user, Sesame score, Ant Credit Pay Limit, average monthly phone bill, the number of calls between call details and the user's first contact, the number of calls between call details and the user's second contact, number of shutdown days, mobile phone silence, number of mutual calls, the number of days from the mobile phone account opening time to the current day, the number of direct contacts in the blacklist, the number of indirect contacts in the blacklist, the number of numbers in call details matching the mobile phone address book, installments, the proportion of 91 Credit reference rejected orders, 91 Credit reference total repaid divided by total loan amount, source, FPD10, FPD30, SPD10, SPD30, TPD10, TPD30, CPD10.

On the basis of this original data set, in accordance with the principle of objectivity, appropriate data cleaning methods are used to provide effective data support for subsequent model training and testing.

(1) Delete indicators with constant values.

(2) Data logical verification. Check the basic logical rationality of each indicator data, such as age should be greater than 0.

(3) Convert indicator variables. For example, convert the registration time into the number of days from the registration time to the current time.

(4) Missing value handling.

(5) Convert categorical indicator variables. Set categorical variables as dummy variables. Among them, unordered categorical variables: such as gender and new/old users, are split into 2 dummy variables respectively. To avoid information duplication, only 1 dummy variable is retained; in addition, ordered categorical variables: such as education, are set as dummy variables. If a doctoral degree is 1, then junior high school to master's degree are also 1; if it is a master's degree, then master's degree to junior high school are 1, and doctoral degree is 0, and so on to show the order.

(6) Convert the dependent variable, overdue performance, into a binary variable. Convert 7 indicators related to post-loan performance, namely "FPD10, FPD30, SPD10, SPD30, TPD10, TPD30, CPD10", into 1 and 0, where 1 represents acceptable overdue performance, regarded as "normal", and 0 represents unacceptable overdue performance, regarded as "overdue".

After data preprocessing, 24 feature indicators (including dummy variables) and 1 dependent variable indicator are retained, with a total of 163,312 pieces of data. The subsequent model establishment and result analysis in this paper will be based on the data after preprocessing.

Among them, to achieve data logical verification, checking the basic logical rationality of each indicator data found that three indicator data have logical rationality problems, namely: average monthly phone bill, number of mobile phone shutdown days, and total mobile phone usage time.

These three indicators have samples with values less than 0, which are obviously abnormal samples, so they are treated as invalid samples and eliminated in subsequent analysis.

2.2. XGBoost Model Construction

XGBoost (Extreme Gradient Boosting) is a machine learning algorithm improved based on GBDT (Gradient Boosting Decision Tree). Its main improvements include second-order Taylor expansion of the loss function, support for multi-core and parallel computing on a single machine, thereby improving training speed.

(1) Model objective function.

The objective function of XGBoost consists of a loss function and a regularization term, and the specific formula is:

$$Obj^{(t)} = \sum_{i=1}^{n} \ell\left(y_i, y_i^{(t)}\right) + \Omega(f_t)$$

Among them, $\ell\left(y_i, y_i^{(t)}\right)$ is the loss function, which measures the difference between the model's predicted value $y_i^{(t)}$ and the true value y_i.$\Omega(f_t)$ is the regularization term, which is used to control the model complexity and prevent overfitting.

(2) Taylor expansion of the loss function.

XGBoost performs second-order Taylor expansion on the loss function, and the specific formula is:

$$\mathrm{Obj}^{(t)} \approx \sum_{i=1}^{n} \left[\ell\left(y_i, y_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t)$$

Among them, $g_i = \frac{\partial \ell\left(y_i, y_i^{(t-1)}\right)}{\partial y_i^{(t-1)}}$ is the first derivative of the loss function.$h_i = \frac{\partial^2 \ell\left(y_i, y_i^{(t-1)}\right)}{\partial \left(y_i^{(t-1)}\right)^2}$ is the second derivative of the loss function.

(3) Regularization term

The specific form of the regularization term $\Omega(f_t)$ is:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$$

Among them, T is the number of leaf nodes.ω_j is the score of the j th leaf node.γ and λ are regularization parameters, which are used to control the complexity of the model.

(4) Tree generation and splitting

XGBoost decides whether to split a node by calculating the split gain. The formula for the split gain is:

$$\mathrm{Gain} = \frac{1}{2}\left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right) - \gamma$$

Among them, G_L and H_L are the sum of gradients and second derivatives of the left child node, respectively.G_R and H_R are the sum of gradients and second derivatives of the right child node, respectively.γ is the split threshold. When the split gain is less than γ, the node no longer splits.

Through the above principles, XGBoost can maintain efficient training speed when processing large-scale data, and effectively prevent overfitting through mechanisms such as regularization and split gain.

2.3. Construction of Consumer Credit Risk Control Strategy Model Based on XGBoost

(1) Selection of key influencing factors. Using PYTHON, the preprocessed sample data set is put into 6 models for training, namely: XGBoost model without parameter tuning, XGBoost model with

parameter tuning, logistic regression model without parameter tuning, logistic regression model with parameter tuning, SVM model, and Gaussian Naive Bayes model. The comparative analysis of training results is mainly carried out from two aspects. On the one hand, from the perspective of parameter adjustment, observe the effect of the XGBoost model before and after parameter adjustment; on the other hand, from the perspective of different models, compare the training effect of other models with that of the XGBoost model. To evaluate the quality of model effects, combined with the background knowledge of credit business, four aspects are selected: training time, accuracy of the training set, true positive rate (TPR) indicator, and interpretability of results. Because predicting a person who should be rejected for a loan as eligible for a loan will cause greater losses to financial institutions and lending platforms, the larger the TPR, the better, that is: the higher the probability of correctly predicting the actual 0 (should be rejected) as 0.

Based on the above four selected evaluation indicators, verify the advantages of the XGBoost model in selecting key influencing factors. And according to the training results of the model, select the key indicator factors affecting the overdue performance of credit customers according to the level of contribution of each indicator.

(2) Strategy design and optimization. Based on the selected key influencing factors and combined with actual business knowledge, design credit risk control strategies (i.e., rule sets). The rules are divided into three categories: first, strict rejection rules, for situations where risks are completely intolerable or violate the regulations of the China Banking and Insurance Regulatory Commission, such as "if the indicator = a specific condition, then reject the loan", such as not handling credit business for minors and students; second, scoring rules, use a relatively stable credit scoring model (such as logistic regression model, financial institutions can also choose internal mature and effective models) to obtain scoring results, and then set rules for different scoring levels according to business needs, such as different levels corresponding to full loan, reduced loan, and rejected loan; third, variable rules, adjusted due to new risk characteristics in the business, with a high frequency of changes, such as setting rules for high-risk platforms and customers hitting the rule set combined with scoring conditions, such as loan rejection and strict manual review. At the same time, strategy design requires business background knowledge and industry sensitivity, and must be verified by actual business data to achieve a low overdue rate and high pass rate, bringing maximum benefits to financial institutions.

After that, find out whether there are strongly relevant indicator factors in the initial strategy that can be used alone to optimize the initial credit risk control strategy. If they exist, the initial risk control strategy is optimized for the second time to design a new credit risk control strategy.

3. Results

70% of the preprocessed samples are used as the training set, and 30% as the validation set. The results are shown in Table.1.The XGBoost model has the highest accuracy, the least time consumption, not very high requirements for input data, and can also give the contribution of each variable, with good interpretability. The confusion matrix of XGBoost is shown in Figure 1. Overall, XGBoost is one of the relatively optimal models. From the test results, there is no significant fluctuation in accuracy before and after parameter tuning, but there is a large difference in time consumption. Therefore, it is recommended to use XGBoost without parameter tuning.

Logistic regression can be used as an alternative method, with relatively good performance in accuracy, time consumption, and interpretability. Moreover, in terms of interpretability, it can give the coefficients and positive/negative signs of each variable, with strong interpretability. In terms of interpretability, SVM and Gaussian Naive Bayes cannot give the coefficients or contribution rates of each variable, so their interpretability is poor.

Table 1. Comparison of results of each model (including new/old user variables)

Model	Training Set	Validation Set	Model Parameters	Time Consumed (minutes)	TPR (Validation Set)
Xgboost	97%	97%	Default parameters	0.12	86.04%
Xgboost with tuning	97%	97%	learning_rate=0.05; n_estimators=500; max_depth=2	7.57	85.90%
SVM	97%	97%	Default parameters	191.16	85.20%
Logistic Regression with tuning	96.52%	96.84%	Penalty=l1, C=0.1	1.78	86.55%
Logistic	96.55%	96.68%	Default parameters	0.16	85.95%
Gaussian Naive Bayes Classification	92.14%	91.89%	Default parameters	0.15	98.08%

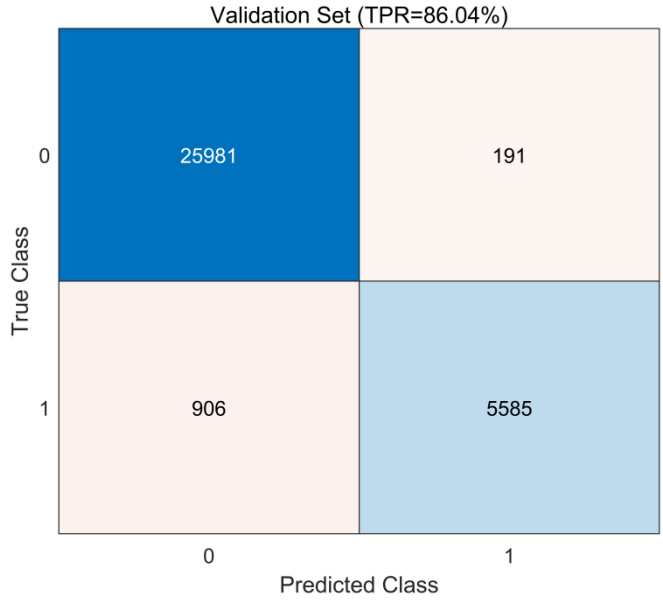

Figure 1. XGBoost confusion matrix

As shown in Figure 2, from the analysis results of the contribution of each variable in the XGBoost model, the importance of variables shows significant differences. "New user" ranks first with a contribution of 0.128. There is a lack of historical interaction data between new users and the platform, and the platform has insufficient information about their behavioral preferences, performance capabilities, etc. This information gap makes the "new user" identity itself a key signal for measuring potential uncertainty, and thus occupies a core weight in the model.It indicates that whether a user is using the product or service for the first time is the core basis for the model's judgment. It may indicate that new users have essential differences from old users in behavioral patterns, consumption habits, or risk characteristics, making this identity attribute a strongly correlated indicator for predicting the target (such as credit risk, purchase intention, etc.). "Sesame Credit" follows closely with a contribution of 0.118. As a credit evaluation indicator, its high weight highlights the key role of credit status in model decision-making. Especially in financial risk control or service access scenarios, this variable directly reflects users' repayment ability and credit reliability. "Number of shutdown days" (0.072) also shows high influence. This variable may be related to users' life stability, consumption ability, or equipment usage habits, indirectly reflecting some potential attributes of users, thereby affecting the model's judgment on the prediction target.

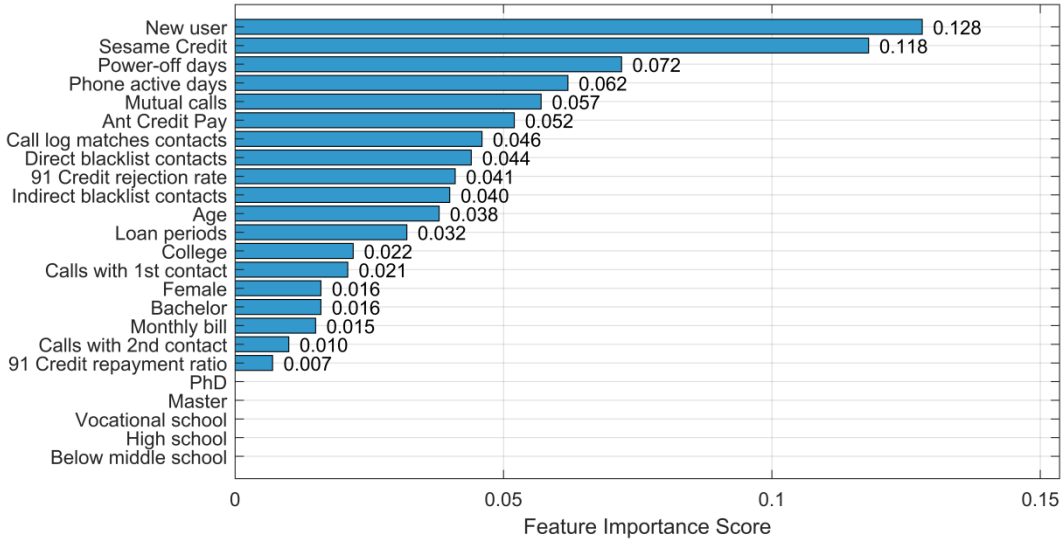

Figure 2. Contribution rates of various variables in XGBoost

In contrast, the contribution of education-related variables such as "doctor" and "master" is 0. This result reveals that in the prediction scenario targeted by the current model, education factors do not have a significant impact on the target variable. This may be because the nature of the business pays more attention to users' actual behavioral data or real-time status rather than educational background. At the same time, variables with low contribution, such as "91 credit reference repayment ratio" (0.007) and "vocational school", although their weights in the model are small, may still provide auxiliary information under specific conditions. For example, when other variables are missing or data is abnormal, these low-contribution variables may play a certain supplementary verification role.

Table 2. Ranking of variable contribution degrees of each model

Logistic Regression	Logistic Regression with Parameter Tuning	XGBoost	XGBoost with Parameter Tuning
New user	New user	New user	New user
91 Credit reference rejection record ratio	Junior high school and below	Sesame Credit	Mutual calls
Female	91 Credit reference rejection record ratio	Number of shutdown days	Sesame Credit
Bachelor	Bachelor	Ant Credit Pay Limit	Number of shutdown days

The results of the contribution rates of variables in the four models are shown in Table.2. Statistical analysis reveals that after training, different models yield different rankings of contribution rates. However, the ranking of the contribution of "new/old users" is basically consistent across the four models, while the rankings of contribution rates of other variables are inconsistent. Key indicators are screened out, and risk control rules are constructed as shown in Figure 3.

The loan application process starts with “Loan Application”. First, it determines whether the user is a New User or an Old User. If it is an Old User, the review will be carried out based on “Sesame Credit” or “Ant Credit Pay Limit”: when the Sesame Credit score is > 650 or the Ant Credit Pay Limit is > 2000, a “Full - amount loan” can be obtained; if the Sesame Credit score is in the range of 560 - 650, or the Ant Credit Pay Limit is in the range of 500 - 2000, the process will enter the “Reduced - amount loan” procedure; if the Sesame Credit score is < 560 and the Ant Credit Pay Limit is < 500 at the same time, it is necessary to enter the “Credit Evaluation” link. If it is a New User, directly go

through “Credit Evaluation”. If the evaluation result is “High”, the loan can be approved; if it is “Low”, “Loan rejection” will be the result, thus completing the determination of the loan application.

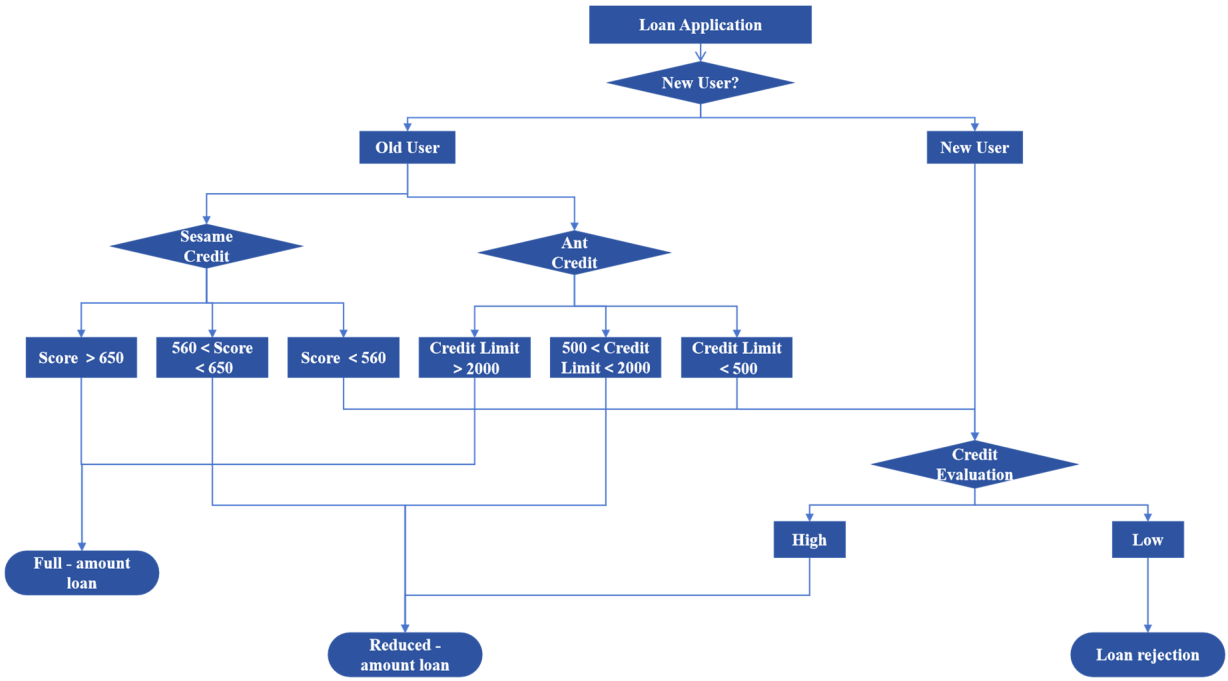

Figure 3. Ecological Benefits of Green Buildings

4. Conclusions

This paper focuses on credit risk prediction in the consumer credit market, constructs an XGBoost model based on 168,000 pieces of data from Internet finance company S, and compares it with logistic regression, SVM, Gaussian Naive Bayes and other models to systematically evaluate the effectiveness of big data risk control models in consumer credit default prediction. The analysis of indicators such as model accuracy, time consumption, and true positive rate (TPR) shows that the XGBoost model achieves an accuracy rate of 97% on both the training set and the validation set, takes only 0.12 minutes, and has a true positive rate of 86.04%. Its overall performance is significantly better than logistic regression, SVM, and Gaussian Naive Bayes models. Variable contribution analysis further reveals that "new user", "Sesame Credit", "number of shutdown days", and "Ant Credit Pay Limit" are key factors affecting overdue behaviors in consumer credit. Based on these core influencing factors, this paper designs corresponding risk control strategies to balance risk control and business efficiency.

In conclusion, this study not only confirms the significant advantages of the XGBoost algorithm in big data risk control of consumer credit, clarifies the core risk factors, but also proposes implementable risk control strategies. The research results provide theoretical support and practical guidance for financial institutions to optimize risk control models and improve credit risk management capabilities, and are of great significance for promoting the sustainable development of consumer credit business. In the future, we can further expand data dimensions (such as incorporating social data, consumption scenario data, etc.) and explore the adaptability of the model in different credit products to enhance the robustness of the risk control system.

References

[1] Bhattacharya A, Biswas S K, Mandal A. Credit risk evaluation: a comprehensive study[J]. Multimedia Tools and Applications, 2023, 82(12): 18217-18267.

[2] Garg A, Maheshwari A, Kapoor R, et al. A Comparative Analysis for Predicting Loan Default Risks using Machine Learning Algorithms[J]. Available at SSRN 5065367, 2024.

[3] Taş C. Comparison of Machine Learning and Standard Credit Risk Models' Performances in Credit Risk Scoring of Buy Now Pay Later Customers[D]. Middle East Technical University (Turkey), 2023.

[4] Niu M, Wang Y, Zhang K, et al. Comparison of different individual credit risk assessment models[C]//Second International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2022). SPIE, 2023, 12597: 664-670.

[5] Uphade D B, Muley A A, Chalwadi S V. Identification of Most Preferable Machine Learning Technique for Prediction of Bank Loan Defaulters[J]. Indian Journal of Science and Technology, 2024, 17(4): 343-351.

[6] Luo Z, Hsu P, Xu N. SME default prediction framework with the effective use of external public credit data[J]. Sustainability, 2020, 12(18): 7575.

[7] Thomas L, Crook J, Edelman D. Credit scoring and its applications[M]. Society for industrial and Applied Mathematics, 2017.