Journal of Computers Vol. 37 No. 5, October 2026, pp. 43-58
https://doi.org/10.63367/19911599202610370500 4

Credit Evaluation Model Based on Adaptive Synthetic and Grey Wolf Optimization Improved XGBoost Algorithm

Zhanjiang Li* and Zheng Chong

College of Economics and Management, Inner Mongolia Agricultural University, Hohhot, China
{lizhanjiang582, chongz1999}@163.com

Received 1 May 2025; Revised 25 July 2025; Accepted 31 July 2025

Abstract. Effective default risk prediction is a critical guarantee for the smooth development and operation of financial institutions, with sound mathematical models serving as the foundation. This paper employs an adaptive synthesis method to balance sample sets and uses the Grey Wolf Optimizer XGBoost to establish a default risk prediction model. Consequently, the ADASYN-GWO-XGBoost credit evaluation model was generated, achieving an accuracy rate of 98% in predicting creditworthiness for Chinese family farms. The study further introduces financial data of Chinese listed companies and compares the prediction results of different models across two samples based on multiple indicators. It demonstrates the effectiveness and scientific nature of the ADASYN-GWO-XGBoost in risk prediction, as well as its improved performance in evaluation capabilities, providing verifiable insights into the method.

Keywords: credit evaluation model, default risk, ADASYN, Grey Wolf Optimization algorithm, XGBoost

1 Introduction

Credit is the foundation for the smooth operation of a market economy and has significant implications for modern socio-economic development. In recent years, the Chinese government has issued several positive agricultural support policies, including various preferential policies for agricultural loans, aimed at making it easier for farmers to obtain credit. However, in practice, these policies have increased the probability of default risks for financial institutions. Default risk not only affects the profits of financial institutions but also investors' decision-making, making it a critical factor in financial stability. Antony and Suresh [1] argue that Preventing default risk is essential for maintaining the cash flow of financial institutions. To address the credit risk issues faced by farmers, having effective credit rating tools—developing default rating tools that can assess loan applicants' creditworthiness from multiple perspectives to avoid default risks—is a fundamental task for financial institutions in risk management.

Bo-Wen Chi and Chiun-Chieh Hsu [2] pointed out in their research that the establishment of a prediction model with the ability to identify default risk can assist in the completion of credit rating, make the evaluation results more scientific and accurate, and provide a certain guarantee for the steady operation and long-term development of financial institutions. Lyn C. Thomas [3] also mentioned in her research that the New Basel Accord emphasizes the importance of establishing credit evaluation models. When scholars study credit evaluation models, they use different ways to process and analyze the data, and the methods they use are often determined by the characteristics of the data under study. The default sample data is a typical dichotomous sample set with diversified characteristics, i.e., a variety of influencing factors as well as indicators need to be taken into account when analyzing, and the models established based on different indicators also have differences in performance. Secondly, according to the experience of real data collection, the sample set of default risk data is balanced is relatively rare, and the number of default samples is significantly different from the number of non-default samples. Also, the quality and completeness of the data are important impact factors when exploring modeling.

In summary, the model needs to be built according to the characteristics of the sample data, using algorithms with higher applicability, and targeted adjustment and optimization according to its advantages and disadvantages.

* Corresponding Author

2 Research Methodology

2.1 Balanced Data Sample Method

When building credit evaluation models, scholars commonly use sample data to train and test the effectiveness of the model. However, due to the influence of practical factors, the number of default samples and non-default samples usually has a great gap, which has a certain degree of influence on the accuracy of the established model. The methods of sample equalization in contemporary research fall into two main categories:

The first is the over-sampling method, that is, for the samples with less data volume, according to the existing index data, to generate simulated new data to fill the sample size. Huosong Xia et al. [4] used the SMS-TE over-sampling method to handle the unbalanced data samples and established the financial risk detection model based on XGBoost. Xingyun Li et al. [5] adopted the adaptive synthetic sampling method to balance the sample size and integrated three models of Logistic Regression, Random Forest and CatBoost to improve the prediction accuracy of loan default risk. When the data set is small or the number of a few samples is insufficient, the oversampling method has good applicability, but it is easy to cause the overfitting of the data samples.

The second is undersampling, which refers to clustering towards many samples, and balancing the actual sample set by lowering the total number of categories in the sample set that are more numerous. Ying Zhou and Xiaoting Su [6] applied the stochastic undersampling method to determine the optimal indicator combinations and optimal training samples using AUC and G-means as indicators, respectively, and finally established a risk prediction model based on linear support vector machine. Under-sampling can reduce the sample size while avoiding data overfitting, but there is a little requirement on the number of samples, and there is also the possibility of mistakenly deleting important data information.

The default risk data sample as a typical representative dichotomous sample set, this paper intends to use Adaptive synthetic sampling approach (ADASYN) oversampling equalization sample method for the sample data used.ADASYN is a sample generation method to deal with the category imbalance problem proposed by Haibo He and other scholars in 2008 [7], which is applicable to the classification task. The method has good practical results in the dualized sample set equalization problem. The application of ADASYN in research is more common recently. Based on the data of cross-border e-commerce, Zhiqiang Li and Hyunpark Yu [8] optimized Bp neural network, SVM and K-nearest neighbor model with ADASYN balanced samples, respectively, and made a comparison before and after optimization, which illustrated that ADASYN has a more obvious effect on optimization. Lian Yang and Baofeng Shi [9] constructed a credit evaluation model based on ADASYN optimized neural network and introduced the focal loss FocalLoss function to construct ADASYN-BPNN-FocalLoss credit risk evaluation model. Xingyun Li et al. [5] used the ADASYN balanced network loan data and fused three models, Logistic Regression, Random Forest and CatBoost, and verified the effectiveness of the model's predictive ability by comparison.

The advantage of ADASYN lies in its ability to dynamically adjust the degree of oversampling according to the distribution of the data, and to be more flexible in dealing with a small number of category distributions of different densities and shapes. Compared to traditional SMOTE methods, ADASYN can dynamically adjust the number of synthetic samples based on the distribution density of minority class samples, avoiding excessive generation of noise data in boundary ambiguous areas. In contrast, SMOTE may exacerbate local overfitting due to uniform sampling, leading to a decrease in model generalization. Therefore, ADASYN is more suitable for highly imbalanced and complexly distributed data in financial scenarios.

2.2 Construction Method of Default Risk Prediction Model

In the existing related research, the methods of building default risk prediction models can be broadly classified into three kinds: one is the mathematical and statistical method of finding the optimal linear combinations for the explanatory variables to analyze the default risk of customers. Guotai Chi et al. [10] use stacked models with different combinations of optimization features to discriminate and analyze the default risk problem of small businesses. This research method has good research ability for modeling low-dimensional linear data, but its final effect will be weakened to some extent when facing high-dimensional nonlinear data.

The second is the credit research on default risk based on experts' experience, i.e., the process of establishing the indicator weighting system based on experts' subjective evaluation of the importance of each indicator, and further establishing the credit evaluation model, and the research algorithms often used include the hierarchical

analysis method and the point estimation of value method, and some other subjective empowerment methods. This method can emphasize the key indicators in a set of data and has some flexibility. Genglan Wang [11] used the hierarchical analysis method to construct the evaluation index system of venture capital projects from the perspective of four elements: entrepreneur quality, technology, market and management, and designed a hierarchical analysis evaluation model. However, due to its excessive subjectivity, it may be affected by expert bias, and scholars often need to consider this aspect.

Thirdly, the machine learning method which is more advantageous to deal with nonlinear loan data, which includes shallow machine learning such as support vector machine and decision tree and deep learning represented by neural network, which uses artificial intelligence algorithms to build a model to study credit related problems. Koresh Galil et al. [12] utilized machine learning techniques, in particular Classification and Regression Tree (CART) and Support Vector Regression (SVR) to predict the credit rating of a company. Yufei Xia [13] proposed a hyper-parameter optimization technique based on TPE algorithm to enhance the decision tree method for credit rating. When using machine learning to construct credit evaluation models, it is often the case that multiple parameters need to be adjusted, thus deteriorating the accuracy of the prediction data, so scholars often introduce optimization functions to improve the algorithm's prediction ability. XGBoost exhibits high accuracy and high efficiency when dealing with both classification and regression problems, and it has a certain advantage as the basic algorithm for default risk prediction models.

Grey Wolf Optimization Algorithm. Grey Wolf optimization algorithm (Grey Wolf Optimizer, GWO) is a group intelligent optimization algorithm, which was first proposed in 2014 by the scholar Mirjalili and others [14] from Griffith University in Australia. The basic idea of the algorithm evolved from the behavioral characteristics of grey wolves when hunting, which has advantages in convergence performance, and has the characteristics of fewer parameters and relatively convenient operation.

In recent years, scholars have paid extensive attention to this algorithm and successfully applied it to the fields of shop floor scheduling, parameter optimization, image classification, etc., and it has played an equally active role in finance-related research. Mohammed Tayebi and Said El Kafhali [15] cited meta-heuristic algorithms, such as GWO, to optimize hyper-parameters to validate that meta-heuristically derived hyper-parameters positive effect in optimizing the performance of fraudulent transaction detection. Min Zhou [16] proposed an improved Grey Wolf Optimization algorithm IGWO for optimizing LSSVM parameters and verifying its effectiveness as well as scientific validity in credit assessment. Feng Zhao and Nyonya Li [17] used the GWO algorithm for optimization of LightGBM parameters to improve the recognition rate of detecting credit card fraud. Anqi Chu and Zhijun Ding [18] applied the Grey Wolf optimization algorithm to synchronize sample undersampling and feature selection for massive sample data, reducing the feature space dimension and improving the classification accuracy.

The Grey Wolf Optimizer algorithm simulates the social hierarchy and cooperative hunting mechanisms of wolf packs. By adjusting only the convergence factor, it can balance global exploration with local development, requiring fewer parameters and converging quickly. In contrast, PSO relies on inertia weights and individual historical best solutions, making it prone to getting trapped in local optima and having higher computational costs. Additionally, the group collaboration mechanism of GWO makes it more robust in high-dimensional parameter spaces of finance, suitable for rapid deployment.

XGBoost. XGBoost (Extreme Gradient Boosting) is an integrated learning algorithm based on gradient boosting decision trees, proposed by T. Chen and C. Guestrin [19] in 2016. It performs well on large-scale, high-dimensional datasets and has achieved excellent results in many machine learning competitions. The core idea is to train multiple weak prediction models by training them to set into one strong prediction model. The basic process is to optimize the boost for the loss function and build a decision number model, while the concept of regularization is introduced to avoid the data overfitting problem. In addition, XGBoost combines feature subsampling and column sampling techniques to enhance the generalization ability of the model by randomly selecting features and samples.

Due to the excellent features of XGBoost, it is widely used in research by scholars. Guotai Chi and Shanshan Wang [20] used particle swarm optimization algorithm to construct an XGBoost-based default prediction model and verified the robustness of the constructed model by using Friedman's test with three public credit datasets in the UCI database. Yixuan Li et al. [21] utilized XGBoost to automatically select the credit risk assessment of credit risk in a digital SCF with the important features in digital SCF and then use MLP for the final assessment to build an XGBoost-MLP model to assess the credit risk of SCF in two stages. Congjun Rao et al. [22] proposed

an integrated Smote-Tomek Link algorithm and an improved Filter-Wrapper for data preprocessing, combining PSO particle swarm and XGBoost to establish a credit risk assessment model for loans. Yuntao Zheng [23] conducted a relevant study on default prediction using a mixture of XGBoost and LightGBM using matched data published by American Express in Kaggle.

XGBoost has been widely used in various real-world problems because it exhibits high accuracy and efficiency when dealing with both classification and regression problems. In contrast, models like random forests suffer from reduced interpretability due to feature redundancy. Additionally, XGBoost's automatic handling of missing values and its parallel computing capabilities for large-scale data make it both accurate and feasible in practical financial scenarios.

2.3 Research Ideas

In view of the two major problems of existing research, namely the number of sample imbalance and the complexity of machine learning parameter adjustment, this paper adopts the adaptive synthetic sampling method (Adaptive synthetic sampling approach, ADASYN), to automatically determine each minority sample, to solve the problem of sample imbalance. At the same time, the grey Wolf optimization algorithm (GWO) is introduced to solve the complicated problem of machine learning parameter adjustment and finally build the default risk prediction model based on ADASYN-GWO-XGBoost. The main model construction process of this study is shown in Fig. 1.

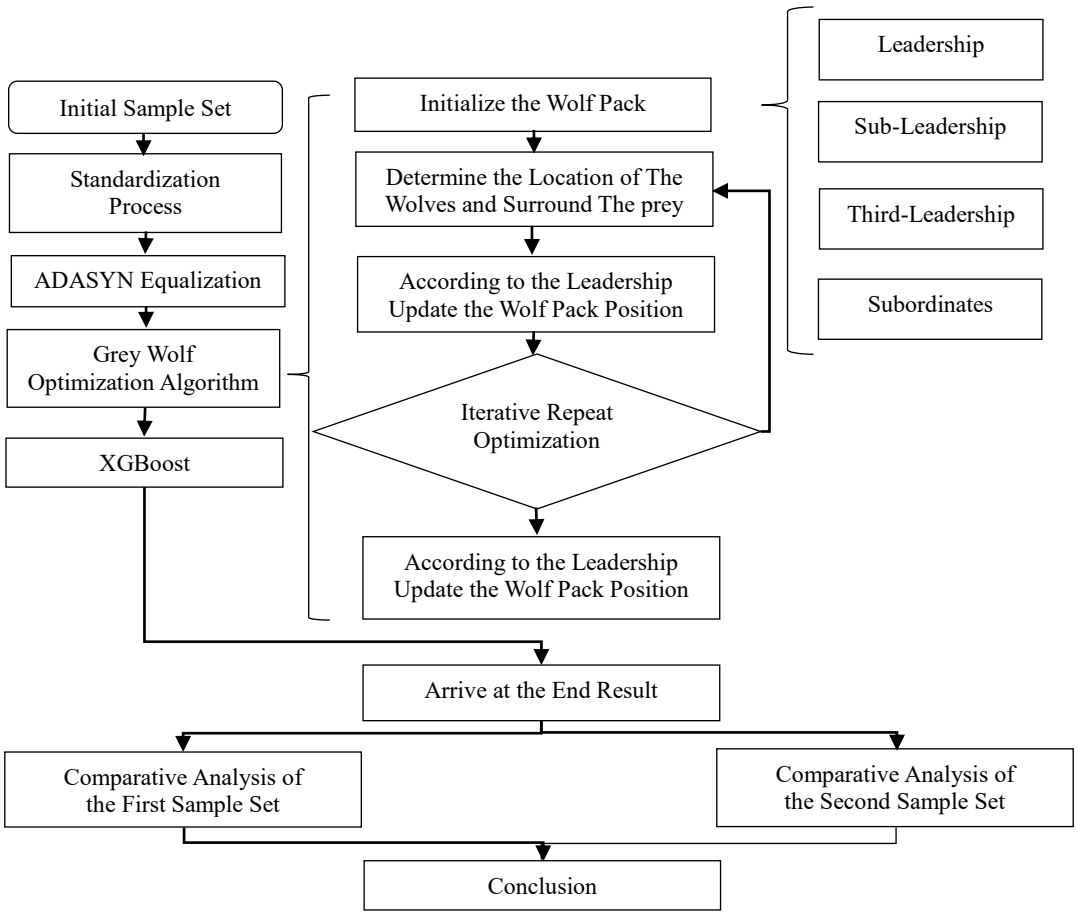


Fig. 1. Technical flow chart

3 Modeling Principles

3.1 ADASYN Principles

First of all it is necessary to set a training set containing m samples D_{tr}, and define the instances in the sample set as x_i, the class label y_i, the number of minority class samples m_s, the number of majority class samples m_l, and from the definition, $m_s + m_l = m$, and $m_l > m_s$.

The algorithmic process is as follows.

Step 1: Calculate the imbalance degree d for the majority and minority classes.

$$d = \frac{m_s}{m_l} \qquad d \in (0,1] \tag{1}$$

Step 2: Preset the parameter α for sample class balance after synthesis $\alpha \in [0,1]$, when α = 1, represents complete balance and calculate the number of minority class samples to be synthesized.

$$N = (m_l - m_s) \times \alpha \tag{2}$$

Step 3: Get the K-nearest neighbor of each sample x_i in the minority class based on the Euclidean distance of the n-dimensional space and calculate the ratio r_i.

$$r_i = \frac{\Delta i}{K} \qquad \Delta i = 1,2\ldots \quad m_s \tag{3}$$

Δi represents the K nearest neighbor ratio of the majority class in x_i, $r_i \in [0,1]$.

Step 4: Regularize r_i.

$$\hat{r}_i = r_i / \sum_{i=1}^{m_s} r_i \tag{4}$$

where r_i conforms to the probability distribution, that is $\sum \hat{r}_i = 1$.

Step 5: Based on the results of Eq. (2) and Eq. (4), the number of samples n_i to be synthesized for each minority class sample x_i.

$$n_i = \hat{r}_i \times N \tag{5}$$

Step 6: For each minority class sample x_i, a random minority sample x_{zi} is proposed from its K-nearest neighbors, and the SMOTE algorithm formula is introduced to synthesize the sample.

$$S_i = x_i + (x_{zi} - x_i) \times l \tag{6}$$

Where (x_{zi} x_i) is the difference vector in n-dimensional space and l is a random number in $[0,1]$, this step will be looped from 1 to n_i several times until the samples are synthesized and the initially set equilibrium parameter level α is reached.

3.2 Principles of Grey Wolf Optimization Algorithm GWO

The GWO optimization process simulates the hunting behavior of grey wolves and builds a mathematical model that includes several stages such as the social hierarchy of the pack, stalking the prey, encirclement, and attacking the target. The social hierarchy is divided into leadership β , secondary leadership γ , tertiary leadership δ , namely, optimal solution, second optimal solution, third optimal solution, and the rest are defined under ω . It works as follows.

Stage 1: Surround the target as you find it.

$$D = \left| C \times X_p(t) - X(t) \right| \tag{7}$$

$$X(t+1) = X(t) - A \times D \tag{8}$$

Equation (7) the result is the distance between the target and the individual, and Equation (8) is the calculation and adjustment of the position of the grey Wolf according to the result of Equation (7). Where t corresponds to the number of iterations of the current algorithm, X_p representing the position vector of the prey, X is the position vector of the grey Wolf, and A and C are the synergy coefficient vector, calculated from the following equation.

$$A = 2a \times f_1 - a \tag{9}$$

$$C = 2 \times f_2 \tag{10}$$

Equation (9) where a is the convergence factor, taking values from 2 to 0 in a linearly decreasing form as the number of iterations of the algorithm rises, and f_1, f_2 is a vector of random numbers with values in the range [0,1].

Stage 2: After the prey is identified and surrounded, the grey Wolf will track the prey, and γ, δ and ω will adjust the position with β as the standard to optimize the enclosure.

The positional variability of individuals with respect to β, γ and δ was first determined.

$$D_\beta = \left| C_1 \times X_\beta - X \right| \tag{11}$$

$$D_\gamma = \left| C_2 \times X_\gamma - X \right| \tag{12}$$

$$D_\delta = \left| C_3 \times X_\delta - X \right| \tag{13}$$

Among them, D_β, D_γ and D_δ are the distances between β, γ and δ and other grey wolves, respectively. X_β, X_γ and X_δ are the current positions of β, γ, δ, and X is the current position of the grey wolf. C_1, C_2 and C_3 are random vectors.

Then adjust the position.

$$X_1 = X_\beta - A_1 \times D_\beta \tag{14}$$

$$X_2 = X_\gamma - A_2 \times D_\gamma \tag{15}$$

$$X_3 = X_\delta - A_3 \times D_\delta \tag{16}$$

$$X(t+1) = \frac{(X_1 + X_2 + X_3)}{3} \tag{17}$$

Eqs. (14), (15) and (16) represent the adjustment of the position of other values ω to optimal solution β, second optimal solution γ and third optimal solution δ, $X(t + 1)$ respectively represents the final position of ω after adjustment. A_1, A_2 and A_3 are random vectors.

Finally, the above process will be completed by several iterations until the maximum number of iterations is reached, and the result obtained after determining all the grey wolf positions is the optimal parameter.

3.3 XGBoost Principle

After the ADASYN equilibrium data and GWO optimization parameters, the data sample can be used for XGBoost machine learning to establish the final default risk prediction model.

The specific calculation steps are as follows:

Stage 1: Set the dataset $D = \{(x_i, y_i)\}$ $(i = 1,2,3\ldots, n)$, where x_i represents the attribute set of the i^{th} sample, and y_i represents the category to which the i sample belongs, then there will be:

$$\tilde{y}_l = \sum_{k=1}^{K} f_k(x_i) \tag{18}$$

Equation (18) represents the prediction for the l^{th} tree; K is the total number of trees, where is the prediction for the k^{th} tree.

Stage 2: Construct the objective function according to the result of equation (18) and the actual value of y_i.

$$loss = \sum_{i=1}^{n} l(\tilde{y}_l, y_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{19}$$

$$\Omega(f_k) = \lambda T + \tfrac{1}{2}\mu \left\| \tau \right\|^2 \tag{20}$$

Equation (19), that is, the objective function consists of differentiable loss function and regularization function, which is the loss function, which represents the difference between the predicted value and the true value, is the added regularization, and its function is to avoid overfitting the function result. In Equation (20), T represents the number of leaf nodes corresponding to the k tree; the fraction of each leaf, the weight, and the value determines the complexity level of the model.

Stage 3: Then train the XGBoost, and the calculation process is as follows:

$$\tilde{y}_l^{(0)} = 0 \tag{21}$$

$$\tilde{y}_l^{(1)} = f_1(x_i) = \tilde{y}_l^{(0)} + f_1(x_i) \tag{22}$$

$$\tilde{y}_l^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \tilde{y}_l^{(t-1)} + f_t(x_i) \tag{23}$$

In equation (23), $\tilde{y}_l^{(t)}$ represents the model prediction result of round t.

Stage 4: A new function is added to keep the result $\tilde{y}_l^{(t-1)}$ of round *t*-1, and the objective function of the round *t* is:

$$loss^{(t)} = \sum_{i=1}^{n} l\left[y_i, \tilde{y}_l^{(t-1)} + f_t(x_i) \right] + \Omega(f_t) \tag{24}$$

Stage 5: Through the Taylor expansion, retain the first three items of formula (24), and remove the minimum items, which can be converted into:

$$loss^{(t)}(q) = \sum_{t=1}^{T} \left[g_i f_i(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \sum_{k=1}^{K} \Omega(f_k) \tag{25}$$

$$g_i = \psi_{\tilde{y}_l^{(t-1)}} l(y_i, \tilde{y}_l^{(t-1)}) \tag{26}$$

$$h_i = \psi^2_{\tilde{y}_l^{(t-1)}} l(y_i, \tilde{y}_l^{(t-1)}) \tag{27}$$

Stage 6: Replacing the optimal result of leaf node into formula (25), because credit evaluation is a two-classification problem, the logarithmic loss function can be used, and the final objective function can be obtained after simplified operation:

$$loss^{(t)}(q) = \sum_{i=1}^{T} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \sum_{k=1}^{K} \Omega(f_k) \tag{28}$$

3.4 Standardization of Indicators

Quantitative indicators can be classified as positive, negative and interval-type indicators and are standardized as follows:

Positive indicator, i.e., the larger the value of this indicator the better the sample is, is given by the formula:

$$y_{ij} = \frac{x_{ij} - \min\limits_{1 \le i \le s}(x_{ij})}{\max\limits_{1 \le i \le s}(x_{ij}) - \min\limits_{1 \le i \le s}(x_{ij})} \tag{29}$$

Negative indicators, i.e., the smaller the value of this indicator the better the sample is doing, are given by the formula:

$$y_{ij} = \frac{\max\limits_{1 \le i \le s}(x_{ij}) - x_{ij}}{\max\limits_{1 \le i \le s}(x_{ij}) - \min\limits_{1 \le i \le s}(x_{ij})} \tag{30}$$

Interval-type indicators, the value of the indicator in a certain range of values indicates that the sample is in good condition, set the value range of this region for [a,b], then the formula is:

$$y=\begin{cases} 1-\dfrac{a-x_{ij}}{\max(\max\limits_{1\le i\le s}(x_{ij})-b,a-\min\limits_{1\le i\le s}(x_{ij}))} & x_{ij}<a \\ 1-\dfrac{x_{ij}-b}{\max(\max\limits_{1\le i\le s}(x_{ij})-b,a-\min\limits_{1\le i\le s}(x_{ij}))} & x_{ij}>b \\ 1 & a\le x_{ij}\le b \end{cases} \tag{31}$$

Where y_{ij} is the standardized value of the *jth* indicator of the *ith* sample after processing, x_{ij} denotes the initial unprocessed data of the j^{th} indicator of the ith sample; s is the total number of samples.

3.5 Comparative Analysis Principle

In order to validate the performance of the ADASYN-GWO-XGBoost default risk prediction model, mathematical evaluation metrics need to be introduced as a way to illustrate its effectiveness and accuracy and to visualize how it compares with other model effects.

Before establishing the comparative evaluation index system, we clarify that TP (True Positive, True Positive) represents the number of correctly identified default samples, FP (False Positive, False Positive) indicates the number of non-default samples incorrectly labeled as defaults, and TN (True Negative) is the number of non-default samples correctly identified as non-defaults, while FN (False Negative, False Negative) is the number of default samples incorrectly labeled as default. Next, a series of evaluation criteria will be used to measure the performance of the model, the evaluation metrics are as follows.

Accuracy is the percentage of the number of correctly classified samples out of all determinations. It is derived by dividing the number of correctly categorized samples by the total number of samples and is often used to measure the overall performance of the model. The formula is expressed as.

$$\text{Accuracy}=\frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})} \tag{32}$$

True Positive Recall Rate (TPR), also known as recall rate, refers to the proportion of samples correctly identified as default by the model in all actual default samples. It is an important measure of the model's ability to detect default samples. The formula can be expressed as:

$$\text{TPR}=\frac{\text{TP}}{(\text{TP}+\text{FN})} \tag{33}$$

True Positive Precision Rate (TPP), also known as precision rate, refers to the proportion of truly defaulted samples in all samples that are predicted to default. The calculation formula follows:

$$\text{TPP}=\frac{\text{TP}}{(\text{TP}+\text{FP})} \tag{34}$$

True Negative Recall Rate (TNR), sometimes called the true negative rate, is the proportion of samples correctly identified as non-default in all actual non-default samples. The formula is:

$$\text{TNR}=\frac{\text{TN}}{(\text{TN}+\text{FP})} \tag{35}$$

True Negative Precision Rate (TNP) is the proportion of samples non-default in all samples predicted to be non-default. The calculation formula follows:

$$\mathrm{TNP} = \frac{\mathrm{TN}}{\left(\mathrm{TN} + \mathrm{FN}\right)} \tag{36}$$

Geometric mean (G-mean) is a performance measure that comprehensively considers precision and recall and is especially suitable for unbalanced datasets. The calculation formula is as follows:

$$\mathrm{G-mean} = \sqrt{\mathrm{TNP} \times \mathrm{TPP}} \tag{37}$$

AUC (Area Under the Curve) refers to the area from the origin to the bottom of the curve on the receiver operating characteristic curve (ROC curve), and this indicator reflects the ability of the model to distinguish between different categories. Because the ROC curve in some cases cannot clearly show the differences between the models, while the AUC can be more intuitive to illustrate the differences in the ability of each model, the larger the value can indicate that the model for the sample prediction ability is better.

4 Empirical Analysis

4.1 Sample Data Processing

The samples used for modeling in this paper are derived from a comprehensive dataset on the operation of family farms and ranches across various leagues and cities in China's Inner Mongolia Autonomous Region, spanning a total of five years. This dataset was meticulously collected and processed to ensure its suitability for the study's objectives. Initially, the raw data comprised a larger set of indicators and samples, but through rigorous preliminary screening and processing, it was refined to include 785 samples and 56 key indicators. These indicators were selected based on their relevance to credit evaluation and default risk prediction, ensuring that the dataset captures the essential financial and operational characteristics of the family farms and ranches.

The dataset includes both qualitative and quantitative indicators, providing a holistic view of the farms' operations. The selection of these indicators was guided by their potential to influence creditworthiness and default risk. For instance, qualitative indicators like business registration and financial management practices reflect the farms' operational legitimacy and governance quality, while quantitative indicators such as profitability and land use duration provide insights into their financial health and stability.

Among the 56 indicators, 23 are qualitative, such as whether the farm has registered a business license or employs professional financial management staff. These qualitative indicators were scored using predefined criteria, as detailed in Table 1 of the paper. The remaining 33 indicators are quantitative, covering aspects like average years of land circulation and annual profit. These quantitative indicators were standardized using methods tailored to their types (positive, negative or interval-type) ensuring consistency and comparability across the dataset.

Table 1. Scoring criteria for qualitative indicators

(1) No.	(2) Indicator name	(3) Option content	(4) Score
1	Whether to register a business license	Yes	1
		No	0
...	...	...	...
23	Whether the operator's family has government workers	Yes	1
		No	0

The standardization process, as outlined in Equations (29) to (31) of the paper, transformed these indicators into a uniform scale, facilitating their integration into the predictive model. After the sample data were scored by the qualitative indicators and standardized by the quantitative indicators as described above, the results are shown in Table 2.

Table 2. Normalization results for the first sample set

(1) No.	(2) Indicator name	(3) Type	The default sample			Non-default sample		
			(4) Sample 1	…	(110) Sample 107	(111) Sample 108	…	(788) Sample 785
1	Whether to register a business license	Quality	0.000	…	1.000	1.000	…	1.000
2	Availability of professional financial management staff	Quality	0.000	…	0.000	1.000	…	1.000
...	...	...	…	…	…	…	…	…
55	Average number of years of land in circulation	Forward	0.071	…	0.100	0.186		0.100
56	Annual profit	Forward	0.014	…	0.013	0.022	…	0.028

4.2 ADASYN Equalization Process

In the current sample data, the total number of default samples is 107, and non-default samples are 678, with a ratio of 107:678, that is, non-default samples are about six times the number of default samples, which has equalization requirement. The sample data is inserted into formula (1) to formula (6), and the new sample data is obtained, containing 1361 samples. The total number of new samples synthesized by the algorithm is 576 default samples. The ratio of default samples and non-default samples in the equilibrium result data is close to 1:1, and the results after equilibrium are shown in Table 3.

Table 3. Equalization results for the first sample set

(1) No.	(2) Indicator name	The default sample			Non-default sample		
		(3) Sample 1	...	(685) Sample 683	(686) Sample 684	...	(1363) Sample 1361
1	Whether to register a business license	0.000	...	0.000	1.000	...	1.000
2	Availability of professional financial management staff	0.000	...	0.000	1.000	...	1.000
...	...	...	...	...	...	...	...
55	Average number of years of land in circulation	0.071	...	0.139	0.186	...	0.100
56	Annual profit	0.014	...	0.201	0.022	...	0.028

4.3 Model Realization

In the balanced treatment dataset, the data were divided into training set and test set according to the ratio of 70% and 30%. First, the training set was used to cultivate the performance of the model, and subsequently the test set was used for prediction and outcome inference. The data is first replaced from equation (7) to equation (17) to calculate the optimal parameters using the GWO algorithm. The parameters of the Grey Wolf Optimization algorithm are shown in Table 4.

Table 4. Parameter setting of the Grey Wolf optimization algorithm

(1) Parameter	(2) Value
Iterations	20
Dimension	2
Minimum limit	0.01
Maximum limit	10

The parameters of GWO are adjusted according to the results in Table 1 and then applied to the XGBoost model to calculate the optimal parameters for the current problem. Then, these parameters are substituted into Equation (18) to Equation (28), and the final prediction of the sample is made by XGBoost. The results are shown in Fig. 2.

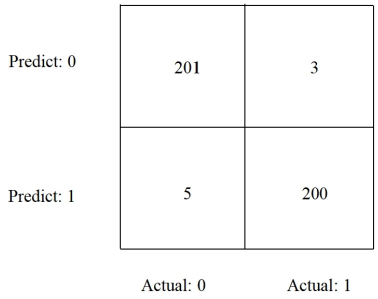


Fig. 2. Heat map of prediction results

4.4 Comparative Analysis

To comprehensively evaluate the performance of the ADASYN-GWO-XGBoost model, this study adopted a multi-dimensional and multi-level comparative analysis method. In terms of experimental design, we carefully selected seven representative comparison models, which cover the mainstream algorithms and technical approaches in the current credit risk assessment field. Specifically, these include: the GWO-XGBoost model based on a single optimization algorithm, the ADASYN-XGBoost model that only uses sample balancing techniques, the unoptimized basic XGBoost model, as well as traditional machine learning algorithms such as K-nearest neighbors, Support Vector Machines, Random Forests and Naive Bayes.

In terms of the evaluation metric system, we adopted a multi-dimensional set of metrics including accuracy (Accuracy), true positive rate (TPP), true positive recall (TPR), true negative rate (TNP), true negative recall (TNR), and geometric mean (G-mean). Specifically, we also calculated the area under the receiver operating characteristic curve (AUC), which provides a comprehensive reflection of the model's overall classification performance across different thresholds. Through systematic calculations from formulas (32) to (37), we obtained the quantitative performance of each model in various metrics, with detailed comparison results summarized in Table 5.

Table 5. Results of the comparative analysis of the first sample set

(1) No.	(2) Arithmetic	(3) Accuracy	(4) TPP	(5) TPR	(6) TNP	(7) TNR	(8) G-mean	(9) AUC
1	A-G-XGBoost	0.980	0.985	0.976	0.976	0.985	0.980	0.980
2	GWO-XGBoost	0.975	0.990	0.981	0.875	0.933	0.931	0.933

3	ADASNY-XGBoost	0.976	0.966	0.985	0.985	0.966	0.976	0.976
4	XGBoost	0.915	0.946	0.955	0.727	0.686	0.829	0.820
5	K-N Neighbor	0.966	0.981	0.981	0.857	0.857	0.917	0.919
6	SVM	0.932	0.985	0.938	0.629	0.880	0.787	0.909
7	Plain Bayes	0.886	0.978	0.886	0.574	0.886	0.749	0.886
8	Random Forest	0.962	0.980	0.975	0.865	0.889	0.921	0.932

In addition, in order to further verify the practicability of the model, this paper introduces the financial data of Chinese listed companies as the second sample set, including 107 indicators, a total of 804 samples, including 239 default samples and 565 non-default samples, with a ratio of about 1:2.4. The results after the standardized treatment are presented in Table 6.

Table 6. Standardized results for the second sample set

(1) No.	(2) Indicator name	The default sample			Non-default sample		
		(3) Sample 1	…	(241) Sample 239	(242) Sample 240	…	(806) Sample 804
1	Net fixed assets	0.000	…	1.000	1.000	…	1.000
2	Financial expenses	0.228	…	0.407	0.118	…	0.159
...	...	…	…	…	…	…	…
106	Cash flow volatility	0.203	…	0.081	0.118	…	0.082
107	Bank advance	0.014	…	0.013	0.022	…	0.028

The same operation as the first sample set was performed on the standardized results of the second data set, and ADASYN-GWO-XGBoost and several other comparison models were introduced to compare and analyze the obtained results. Relevant results are shown in Table 7.

Table 7. Results of the comparative analysis of the second sample set

(1) No.	(2) Arithmetic	(3) Accuracy	(4) TPP	(5) TPR	(6) TNP	(7) TNR	(8) G-mean	(9) AUC
1	A-G-XGBoost	0.926	0.918	0.934	0.935	0.918	0.926	0.926
2	GWO-XGBoost	0.872	0.918	0.902	0.764	0.797	0.837	0.841
3	ADASNY-XGBoost	0.910	0.954	0.864	0.874	0.958	0.913	0.911
4	XGBoost	0.926	0.938	0.959	0.894	0.843	0.915	0.901
5	K-N Neighbor	0.665	0.688	0.892	0.541	0.238	0.610	0.565
6	SVM	0.727	0.722	0.994	0.875	0.097	0.795	0.546
7	Plain Bayes	0.711	0.720	0.971	0.500	0.071	0.600	0.521
8	Random Forest	0.962	0.980	0.975	0.865	0.889	0.921	0.932

The results presented in Table 5 and Table 7 highlight the superior performance of the ADASYN-GWO-XGBoost model in predicting default risk across two distinct datasets: the credit rating sample set of family farms in Inner Mongolia and the financial data of Chinese listed companies.

In the family farm dataset, the model achieved an impressive Accuracy of 0.98, securing the top position among the eight evaluated credit rating tools. This high accuracy underscores the model's ability to correctly classify both default and non-default samples. The True Positive Precision (TPP) of 0.985 ranked second, indicating that the model is highly reliable when it predicts a default, with very few false positives. The True Positive Recall (TPR) of 0.976, ranking fourth, reflects the model's strong capability to identify actual default cases, minimizing the risk of overlooking critical defaults. The True Negative Precision (TNP) of 0.976, ranking second, and the True Negative Recall (TNR) of 0.985, ranking first, demonstrate the model's robustness in correctly identifying non-default cases, which is crucial for minimizing unnecessary financial interventions. The Geometric Mean (G-mean) of 0.98 and the Area Under the Curve (AUC) of 0.98, both ranking first, further validate the model's balanced performance across all classes and its excellent discriminatory power between default and non-default cases.

For the dataset of Chinese listed companies, the model maintained a high Accuracy of 0.98, ranking second, which confirms its adaptability to different financial contexts. The TPP of 0.985 (ranking fourth) and TPR of 0.976 (ranking fifth) indicate consistent performance in identifying defaults, though with slight variations due to the dataset's unique characteristics. The TNP of 0.976 (ranking first) and TNR of 0.985 (ranking second) highlight the model's continued excellence in recognizing non-default cases. The G-mean of 0.98 (ranking first) and AUC of 0.98 (ranking second) reaffirm the model's balanced and reliable performance, even when applied to more complex corporate financial data.

The performance advantages of ADASYN-GWO-XGBoost can be intuitively seen from the default rating results of the two sample sets and the comparison results with other models, which can be mainly attributed to the synergistic effect of the following two core mechanisms:

(1) Effective alleviation of the data imbalance problem by ADASYN

When traditional models face highly imbalanced original data, due to the gap in the number of samples, the models are more likely to be influenced by the majority class samples when receiving information. Finally, the prediction results are more biased towards the majority class samples, resulting in insufficient ability to identify default samples. In the model established in this study, ADASYN is used to dynamically generate synthetic samples. The oversampling ratio is adaptively adjusted according to the distribution density of the minority class samples, making the number of default samples close to a 1:1 ratio with that of non-default samples, which significantly improves the model's sensitivity to the minority class. As a result, the True Positive Rate (TPR) increased to 0.976 and 0.934 respectively in the two data sample test sets. In addition, ADASYN avoids the local overfitting problem that may be caused by oversampling methods such as SMOTE. By using a weighting mechanism to reduce the generation of noisy samples, it further enhances the generalization ability of the model.

(2) The efficiency of GWO in optimizing the parameters of XGBoost

The performance of the XGBoost model highly depends on the selection of hyperparameters. Traditional grid search or random search methods are time-consuming and prone to getting trapped in local optima. The GWO algorithm simulates the cooperative mechanism of wolf pack hunting and quickly converges to the global optimal solution within 20 iterations, significantly reducing the risk of overfitting of the model. At the same time, it improves the feature diversity through feature subsampling. In contrast, the XGBoost model without parameter adjustment by GWO has insufficient ability to capture complex nonlinear relationships due to unreasonable parameter settings. In addition, the dynamic search mechanism of GWO enables it to balance the exploration and exploitation stages, thus generating results with high stability when predicting different sample sets.

5 Conclusion

This research effectively alleviates the problem of sample imbalance in default risk prediction through the ADASYN algorithm and introduces the GWO algorithm to optimize the parameters of XGBoost, constructing an ADASYN-GWO-XGBoost default risk prediction model. The experimental results show that this model outperforms other models in multiple evaluation indicators such as G-mean, Accuracy, and AUC, demonstrating its potential in risk prediction. However, this study still has the following limitations:

(1) Limitations of synthetic data: The model training relies on the synthetic samples generated by ADASYN. Although this method alleviates the class imbalance problem by dynamically adjusting the oversampling degree, the synthetic data is completely generated based on the existing sample distribution, which may lead to the model performing excellently in simulated data but having insufficient generalization ability in real scenarios. Especially in the actual financial environment, the data distribution is complex and dynamically changing. Over-

reliance on synthetic data may cover potential noises or abnormal patterns, thus affecting the actual prediction effect of the model.

(2) Insufficient verification of real data: The current experiment is only based on two types of sample sets, namely the financial data of family farms in Inner Mongolia and listed companies, and the sample size is limited. The diversity and complexity of the financial environment in different regions and industries require the model to be verified in a wider range of real datasets to further test its universality.

In response to the above limitations, future research can be carried out in the following directions:

Firstly, collect more complex real datasets. It is necessary to further collect larger-scale, multi-source heterogeneous real default data, verify the robustness of the model in complex scenarios, and explore its adaptability in different economic cycles or industry backgrounds. Secondly, explore more excellent algorithm combinations. Try to integrate other oversampling techniques or optimization algorithms, further compare and optimize, and reduce the dependence on a single synthetic method. Thirdly, enhance the interpretability of the constructed model. Combine interpretability tools such as SHAP and LIME to analyze the decision-making logic of the model, help financial institutions understand key risk factors, and improve the credibility and operability of the model in practice. Fourthly, improve the model's ability to dynamically capture risks. Introduce an online database or a transfer learning mechanism to enable the model to dynamically adapt to changes in the data distribution and reduce the performance degradation caused by static synthetic data.

In summary, the ADASYN-GWO-XGBoost model provides an efficient framework for default risk prediction, excelling in balancing imbalanced datasets and integrating through intelligent algorithms to optimize predictive accuracy. However, despite its significant potential in theoretical and experimental settings, its practical applicability still requires rigorous validation through extensive testing of diverse real-world financial datasets. Future research should focus on empirical validation across different industries, economic cycles, and regional financial environments to ensure the robustness and generalizability of the model. Additionally, further exploration of hybrid methods such as combining ADASYN with other advanced sampling techniques or integrating alternative optimization algorithms can enhance the model's adaptability and performance. At the same time, improving the interpretability of the model is crucial for financial institutions to adopt it, ensuring transparency in the decision-making process. By bridging the gap between algorithmic innovation and practice, researchers can translate these theoretical advancements into practical tools for financial risk management, ultimately promoting more stable and efficient credit markets. This balanced approach will be key to driving the development of the field and ensuring that such models bring tangible benefits in real-world financial applications.

6 Acknowledgments

The research was supported by the National Natural Science Foundation of China (72161033), Inner Mongolia Natural Science Foundation of China (2025MS07001), Start-up Support Program for Overseas Students in Inner Mongolia for Innovation and Entrepreneurship of China (2025) and Inner Mongolia Rural Pastoral Development Institute of China.

References

[1] T.-M. Antony, G. Suresh, Determinants of credit risk: Empirical evidence from Indian commercial banks, Banks and Bank Systems 18(2)(2023) 88-100.
https://doi.org/10.21511/bbs.18(2).2023.08

[2] B.-W. Chi, C.-C. Hsu, A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model, Expert Systems with Applications 39(3)(2012) 2650-2661.
https://doi.org/10.1016/j.eswa.2011.08.120

[3] L.-C. Thomas, Modelling the credit risk for portfolios of consumer loans: Analogies with corporate loan models, Mathematics and Computers in Simulation 79(8)(2009) 2525-2534.
https://doi.org/10.1016/j.matcom.2008.12.006

[4] H.-S. Xia, W.-Y. An, Z.-P. Zhang, Credit risk models for financial fraud detection: a new outlier feature analysis method of XGBoost With Smote, Journal of Database Management 34(1)(2023) 1-20.
https://doi.org/10.4018/JDM.321739

[5] X.-Y. Li, D.-J. Ergu, D. Zhang, D.-F. Qiu, Y. Cai, B. Ma, Prediction of loan default based on multi-model fusion,

Procedia Computer Science 199(2022) 757-764.
https://doi.org/10.1016/j.procs.2022.01.094
[6] Y. Zhou, X.-T. Su, Credit Risk Prediction of Company Based on Optimal Feature Set, Journal of System & Management 30(5)(2021) 817-838.
https://xtglxb.sjtu.edu.cn/EN/10.3969/j.issn.1005-2542.2021.05.001
[7] H.-B. He, Y. Bai, E.-A. Garcia, S.-T. Li, ADASYN: adaptive synthetic sampling approach for imbalanced learning, in: Proc. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008.
https://doi.org/10.1109/IJCNN.2008.4633969
[8] Z.-Q. Li, X.-P. Yu, Research on Credit Risk Model Optimization of Small and Micro Enterprises in Cross-border E-commerce Based On ADASYN, Journal of Jiangxi Normal University (Philosophy and Social Sciences edition) 56(2)(2023) 118-127.
https://doi.org/10.3969/j.issn.1000-579X.2023.02.015
[9] L. Yang, B.-F. Shi, Credit Risk Evaluation Model and Empirical Research Based on Focal Loss Modified Cross-Entropy Loss Function, Chinese Journal of Management Science 30(5)(2022) 65-75.
https://doi.org/10.16381/j.cnki.issn1003-207x.2020.2188
[10] G.-T. Chi, X. Huang, Y. Zhou, X.-B. Guo, Discriminating the default risk of small enterprises: Stacking model with different optimal feature combinations, Expert Systems with Applications 229(Part A)(2023) 120494.
https://doi.org/10.1016/j.eswa.2023.120494
[11] G.-L. Wang, The evaluate of project risk investment based on AHP, in: Proc. 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), 2010.
https://doi.org/10.1109/ICACTE.2010.5579178
[12] K. Galil, A. Hauptman, R.-L. Rosenboim, Prediction of corporate credit ratings with machine learning: Simple interpretative models, Finance Research Letters 58(Part D)(2023) 104648.
https://doi.org/10.1016/j.frl.2023.104648
[13] Y.-F. Xia, C.-Z. Liu, Y.-Y. Li, N.-N. Liu, A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring, Expert Systems with Applications 78(2017) 225-241.
https://doi.org/10.1016/j.eswa.2017.02.017
[14] S. Mirjalili, S.-M. Mirjalili, A. Lewis, Grey Wolf Optimizer, Advances in Engineering Software 69(2014) 46-61.
https://doi.org/10.1016/j.advengsoft.2013.12.007
[15] M. Tayebi, S.-E. Kafhali, Performance analysis of metaheuristics based hyperparameters optimization for fraud transactions detection, Evolutionary Intelligence 17(2)(2024) 921-939.
https://doi.org/10.1007/s12065-022-00764-5
[16] M. Zhou, Credit evaluation for hybrid grey wolf optimization and least squares support vector machine approach, Journal of Chengdu University of Technology (Science & Technology Edition) 46(4)(2019) 507-512.
https://doi.org/10.3969/j.issn.1671-9727.2019.04.11
[17] F. Zhao, N.-N. Li, Credit card fraud detection method based on VAE-GWO-LightGBM, Journal of Northeast Normal University (Natural Science Edition) 55(4)(2023) 77-84.
https://doi.org/10.16163/j.cnki.dslkxb202209210002
[18] A.-Q. Chu, Z.-J. Ding, Application of Gray Wolf Optimization Algorithm on Synchronous Processing of Sample Equalization and Feature Selection in Credit Evaluation, Computer Science 49(4)(2022) 134-139.
https://doi.org/10.11896/jsjkx.210300075
[19] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proc. the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
https://doi.org/10.1145/2939672.2939785
[20] G.-T. Chi, S.-S. Wang, Default risk prediction model for Chinese listed companies based on XGBoost, Journal of Systems & Management 33(3)(2024) 735-754.
https://doi.org/10.3969/j.issn.1005-2542.2024.03.013
[21] Y.-X. Li, C. Stasinakis, W.-M. Yeo, A hybrid XGBoost-MLP model for credit risk assessment on digital supply chain finance, Forecasting 4(1)(2022) 184-207.
https://doi.org/10.3390/forecast4010011
[22] C.-J. Rao, Y. Liu, M. Goh, Credit risk assessment mechanism of personal auto loan based on PSO-XGBoost Model, Complex & Intelligent systems 9(2)(2023) 1391-1414.
https://doi.org/10.1007/s40747-022-00854-y
[23] Y. Zheng, A default prediction method using XGBoost and LightGBM, in: Proc. 2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), 2022.
https://doi.org/10.1109/ICICML57342.2022.10009823