ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20230302000395
Original scientific paper

Default Prediction of Internet Finance Users Based on Imbalance-XGBoost

Wenlong LAI

Abstract: Fast and accurate identification of financial fraud is a challenge in Internet finance. Based on the characteristics of imbalanced distribution of Internet financial data, this paper integrates machine learning methods and Internet financial data to propose a prediction model for loan defaults, and proves its effectiveness and generalizability through empirical research. In this paper, we introduce a processing method (link processing method) for imbalance data based on the traditional early warning model. In this paper, we conduct experiments using the financial dataset of Lending Club platform and prove that our model is superior to XGBoost, NGBoost, Ada Boost, and GBDT in the prediction of default risk.

Keywords: imbalanced data; internet finance; P2P lending; XGBoost

1 INTRODUCTION

Finance is the general hub of social capital movement, an important regulator of the national economy, and plays an indispensable role in the process of national economic construction and social development. With the continuous vigorous development of information technology, the Internet has injected new vitality into the financial industry. Mobile banking, third-party payment, online insurance sales and other Internet financial services have been integrated into people's daily lives and become part of their lifestyles. The development of information technology has reduced the cost of financial services, expanded the market scope and created new business models [1]. Under the Internet financial model, MSMEs (Micro, Small, and Medium Enterprises) have more financing opportunities, and their financing channels have changed from the traditional single, centralized channels of banks and financial institutions to more and more decentralized peer-to-peer financing channels.

But opportunities and risks exist side by side. Compared with the traditional financial industry, Internet finance has a more complex credit risk problem. Lending platforms "run away", borrowers "escape", investors "run away" has become a social chaos, market confidence suffered a serious blow. A quick and accurate assessment of the repayment ability of borrowers can help reduce the financial risk of lenders, reduce credit fraud, and ensure the safety of Internet financial transactions [2].

In the field of financial credit risk assessment and prediction, it has been studied by many experts and scholars. From traditional machine learning models to sophisticated deep learning models, the accuracy of default prediction has been increasing. However, financial data are inherently complex, highly noisy, dynamic, nonlinear, nonparametric, and chaotic [3]. The lending data is typically imbalanced data, which shows that there are far more normal lending samples than default lending samples, which leads to a model that may have high training accuracy but poor prediction for minority class, which requires us to deal with the imbalance of the data. Existing methods often deal with imbalanced data by under-sampling in the majority class and over-sampling in the minority class, whose division between majority and minority classes depends on the data labels [16-23]. In the real training process, some of the data will interfere with the majority class sample training and have an impact on the minority class sample training, therefore, the simple use of labels to divide the majority class and minority class will lose some useful information and reduce the model accuracy.

Therefore, considering the above problems, this paper introduces the iForest anomaly detection method into the process of imbalanced data processing and constructs an I-XGBoost (Imbalance XGBoost) model based on data balancing processing and XGBoost. The model mitigates data noise and solves the data imbalance problem through feature selection, imbalanced data category detection and imbalanced data processing, and finally uses multiple XGBoost components to complete the prediction in the form of voting.

The main contributions of this paper can be summarized as follows.

(1) We propose a loan default prediction method based on data balancing processing with XGBoost to solve the problem of noise and imbalance in the data.

(2) For the noise in the data, we use three ways to feature select the data separately, and analyze and process the loan data according to the contribution, stability and importance of the data to the prediction, and extract the appropriate features from them.

(3) In response to the unreasonable practice of using data labels to classify data into majority and minority, this paper introduces the iForest detection algorithm to pre-classify data, mix sampling of imbalanced data, and defining "outliers" as minority data, which reduces the learning difficulty of the classifier for minority data while removing noise from majority data. The iForest detection algorithm is used to pre-classify the data.

(4) The proposed method is evaluated on a real dataset in this paper. The experimental results show that the I-XGBoost model effectively improves the recall rate while ensuring the prediction accuracy compared to XGBoost, NGBoost, AdaBoost, and GBDT.

The rest of this paper is described as follows: In section 2, we review the existing methods for assessing financial credit risk. In section 3, we analyze financial data in detail, and in section 4, we propose our I-XGBoost model and introduce its internal structure in detail. In section 5, we conduct prediction experiments and analyze the results. At the end, we summarize the work of this paper.

2 RELATED WORKS

With the rapid growth of the total number of Internet platforms, the associated Internet financial defaults are becoming more frequent, causing investors to suffer huge financial losses, and therefore the proper assessment of credit risk is attracting more and more attention. Risk assessment includes bankruptcy prediction, credit scoring, credit rating, loan/insurance default prediction, bond rating, loan application, consumer credit determination, corporate credit rating, mortgage selection decision, financial distress prediction, tax evasion prediction, etc. [4]. Currently, the widely used methods for financial credit risk assessment and prediction include two main categories: traditional machine learning and deep learning.

Machine learning and data mining methods have been widely used in financial credit risk assessment in recent years, for example, Altman et al. [5] used traditional logistic c models to analyze the relationship between overall default rates and default losses on bank loans and corporate bonds; Mue et al. [6] tested the impact of social capital on borrower default risk through a logistic algorithm and concluded that by increasing the network lending platform will be able to effectively reduce the overall default rate of the industry; Brown and Mues [7] used logistic regression, neural network, and decision tree models to study loan defaults, and also used GBDT, least squares support vector machine, random forest, and gradient boosting methods to build credit scoring models to predict loan defaults; Jin and Zhu [8] classified the loan status on Lending Club into default, near default, and no default, and conducted a comparative study using different machine learning models. Wang and Rong et al. [31] selected macroeconomic-level and firm micro-level indicators based on the characteristics of the Chinese bond market and used the XGBoost model for default risk prediction.

Considering the powerful characterization capabilities of deep learning, many researchers have applied deep learning to financial risk assessment to improve accuracy, mainly including single model-based approaches and hybrid model-based approaches [4]. Most of the studies on deep learning for risk assessment focus on credit scoring to detect the presence of fraudulent behavior of users. For example, Luo et al. [9] used DBN for credit scoring and company default swap data for company credit rating and output the corresponding credit rating; Neagoe et al. [10] used DMLP and CNN to classify customer credit based on customer characteristics, respectively. There have also been some studies dedicated to fraud detection in transactions. For example, Roy et al. [11] used LSTM to classify a dataset containing 80 million credit card transactions to detect the presence of credit card fraud; Jurgovsky et al. [12] used LSTM to detect credit card fraud from a sequence of credit card transactions; Heryadi et al. [13] investigated CNN, stacked LSTM, and hybrid CNN-LSTM models for credit card fraud detection in Indonesian banks and analyzed the impact of data imbalance problem between fraud and non-fraud data on performance metrics. Many integrated methods have also been used for credit risk assessment. In 2018, Zhu et al. [14] made the first attempt to apply CNNs to credit scoring and proposed a hybrid model of Relief-CNN based on CNNs. Kvamme et al. [15] introduced a new mortgage risk assessment model that uses a combination of CNN and random forest classifier based on consumer transaction data model to predict whether a customer will default or not.

It is worth noting that in loan scenarios, the defaulted ones are often in the minority, which makes the distribution of default data show more than obvious bias, which can also be called imbalance. Both the above machine learning and deep learning models are greatly limited in the accuracy of risk assessment under the data imbalance problem. Studies [16-18] have shown that the distribution imbalance problem can be solved at the data level or at the algorithm level, such as resampling methods [19] and LR-SMOTE methods [20]. From Al-Shabi et al. [21], AE models were found to be very effective in dealing with imbalanced datasets, Yu et al. [22] studied DBN and SVM cascade hybrid models for customer credit risk classification, using resampling techniques to cope with imbalanced data, and the final results showed that the method could effectively improve the classification performance, Liu [23] combined resampling, feature engineering and NGBoost algorithm to build a network loan prediction model and improve the model prediction accuracy.

Throughout the development of credit risk warning in the past, researchers kept exploring and started to study default rates and build loan credit scoring models, from traditional machine learning models to deep learning models, from single models to integrated models, people gradually realized the impact of data imbalance problem on the performance of credit evaluation models and made a lot of exploration, including resampling at the data level and algorithm level cost-sensitive learning. However, there are few hybrid algorithms that integrate the two. There is still a lack of sufficiently accurate credit evaluation models that can handle imbalances while extracting data features.

Based on the above analysis, this paper uses the financial loan dataset of Lending Club to establish a loan default prediction method based on feature selection, data balancing processing and XGBoost, which can automatically and effectively extract valuable features, solve data noise and data imbalance problems, and make more accurate analysis and prediction.

3 FINANCIAL DATA ANALYTICS

3.1 Dataset

In this paper, we use data from Lending Club, one of the largest P2P lenders in the world. The large amount of data and the large number of variables in the Lending Club dataset make it an ideal dataset for conducting experiments in the area of financial risk control. In this paper, we use loan data from the Lending Club dataset from 2007 to Q4 (fourth quarter) of 2018, with 2260,701 samples and 151 features. The main information of the original dataset is described below.

The 151 features of the original dataset contain the following three types of information: loan information, borrower credit information, and borrower personal information. Specifically, loan information includes loan status, loan purpose, application amount, and loan term, etc.; credit information includes risk level, debt-to-income ratio, and number of overdue times within two years, etc.;

and personal information includes job title, years of work, and annual income.

3.2 Default Influencing Factors

(1) Borrower Factors

(i) Soft Factors

Soft factors refer to factors that are difficult to substantiate based on objective facts, including the personal characteristics of the borrower, the purpose of the loaning and social capital relationships.

The borrower's personal characteristics mainly include the borrower's age, gender, educational background, marital status and other information, which can reflect the borrower's character traits to a certain extent, which in turn affects the borrower's default risk. It is generally believed that older borrowers have a lower risk of default than younger borrowers; female borrowers have a lower risk of default; married borrowers have a lower risk of default than unmarried borrowers; borrowers with higher education have a lower risk of default than those with lower education, and the influence of these personal characteristics on default risk has been confirmed in previous relevant studies [2, 33, 34].

Loaning purposes mainly include loan repayment, business start-up, medical treatment, etc. They affect the default risk to different degrees, such as higher default risk for borrowing to repay previous loans, and higher default risk for initial business start-up compared to business expansion.

Social capital relationships refer to the borrower's family, friends and other social relationships. It is generally believed that a good social network helps the borrower repay the loan and reduces the risk of default.

(ii) Hard Factors

Hard factors are factors that can be substantiated by objective facts, including the borrower's economic and credit characteristics.

Economic characteristics mainly include information about the borrower's income and property status, which can be used to assess the risk of default, such as borrowers who own more properties and vehicles have a lower risk of default.

Credit characteristics can assess the risk of default based on information such as the borrower's historical lending behavior and credit history, such as a credit rating based on the borrower's historical behavior; the higher the borrower's credit rating, the lower the risk of default.

(2) Loaning Factors

Loaning Factors refer to the characteristic information of this loan behavior, mainly including the loaning amount, term, interest rate, etc. It is generally believed that the larger the borrowing amount and the higher the interest rate, the higher the default risk; the longer the borrowing term means, the more uncertain and the higher the total interest amount, which also causes the higher default risk.

Borrower Factors and loaning Factors together constitute the influencing factors of loaning default. In reality, it is often necessary to analyze and evaluate these two types of Factors and fully consider various factors that may affect the borrower's repayment in order to effectively avoid the risk of default.

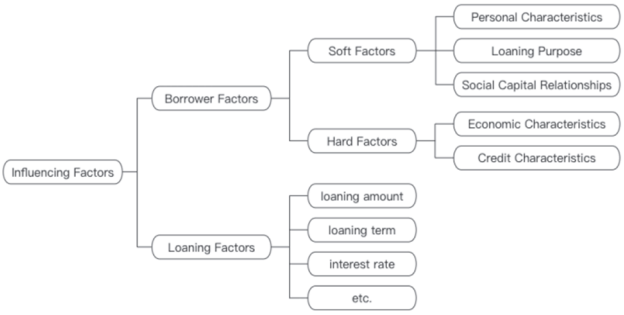

Figure 1 Influencing factors of loaning default

3.3 Imbalance Analysis

(1) Distribution of the Target Variable

In the dataset used in this paper, the target variable is the loan status variable, which indicates the lending status and has nine different values. The nine lending statuses can be classified into three types: Paid, Default, and Late, with each loan status and its corresponding type as follows.

Table 1 Loan status and type

variable	type
Fully Paid	Paid
Current	Paid
Charged Off	Default
Late (31-120 days)	Late
In Grace Period	Late
Late (16-30 days)	Late
Does not meet the credit policy. status: Fully Paid	Paid
Does not meet the credit policy. status: Charged Off	Default
Default	Default
NaN	

Table 2 Loan Status Distribution

status	amount
Paid	1957056
Default	269360
Charged Off	Default

Two of the loan types, Paid and Default, were taken as the target variables for the subsequent classification experiments and were distributed as follows.

It can be seen that the number of samples in the Paid and Default status differs significantly, with a ratio of about 7:1, which is called "imbalanced data".

(2) Imbalanced data

Imbalanced data refers to data where the samples are very unevenly distributed across categories. Traditional classification methods usually assume a balanced distribution of data categories, but in reality, data usually have imbalanced characteristics, the number of samples in one category is smaller or even much smaller than the number of samples in other categories.

When dealing with the binary classification problem, if the number of positive and negative samples in the data is extremely imbalanced, the traditional machine learning methods tend to overly favour the majority class in training and ignore the minority class, resulting in lower classification accuracy for the minority class. If we use the traditional machine learning method, the model will tend to predict the samples as "Paid" because it can improve the accuracy of the model, but this cannot effectively prevent the risk of loan default. In practice, we need to design

algorithms for positive and negative sample distributions, and ensure the classification accuracy of majority and minority classes in imbalanced data.

4 METHOD

4.1 I-XGBoost Model Architecture

As mentioned above, the imbalance problem of lending data seriously affects the performance of default account detection models. For this purpose, this paper proposes an I-XGBoost model based on data balancing processing with XGBoost, and the model framework is shown in Fig. 2.

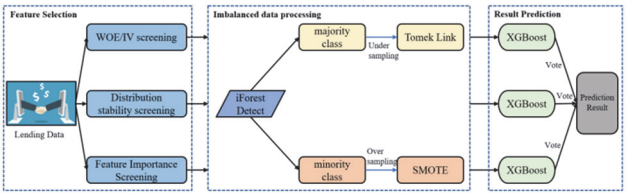

Figure 2 I-XGBoost model framework

The model consists of three modules: feature selection module, imbalanced data processing module and result prediction module. The processing flow is as follows:

(1) First, in the feature selection module, the loan data are fed into the WOE/IV screening, distribution stability screening and feature importance screening algorithms, respectively, to obtain the feature-selected data.

(2) Second, the feature-selected data will be processed by the imbalanced data processing module for data balancing, and the original data is divided into majority class samples and minority class sample sets by iForest detection, and the majority class samples and minority class samples are under-sampled and over-sampled by Tomek Link algorithm and SMOTE algorithm, respectively.

(3) Finally, each set of data is fed into the XGBoost model and trained. For each new sample, three sets of XGBoost predictions are obtained, and the final prediction is determined by voting.

Next, this article will introduce the above three modules in detail.

4.2 Feature Selection Module

A large amount of data, many data dimensions, and many data features do not mean that a better model can be trained; therefore, feature selection is necessary. In this paper, we use three ways to feature select the data separately for screening the contribution, stability and importance of the data to the prediction.

First, a WOE/IV statistical value screening is performed, which can be used to measure the contribution of the independent variable to the model. To code WOE for a variable, it is necessary to first process this variable in groups. For the i-th group, the WOE is calculated as follows.

$$WOE_i = \ln\left(\frac{p_{y_i}}{p_{n_i}}\right) = \ln\left(\frac{y_i / y_T}{n_i / n_r}\right) = \ln\left(\frac{y_i / n_i}{y_T / n_T}\right) \tag{1}$$

where p_{y_i} is the proportion of defaulting users in group i to the defaulting users in all samples, and p_{n_i} is the proportion of normal users in group i to the proportion of normal users in all samples, and y_i is the number of defaulting users in group i, and n_i is the number of normal users in group i, and y_T is the number of all non-compliant users in the sample, and n_T is the number of all normal users in the sample.

Next, the IV values of the variables in each grouping are calculated using the WOE_i. The IV values of the variables in each grouping are calculated with the following equation.

$$IV_i = \left(p_{y_i} - p_{n_i}\right) * WOE_i \tag{2}$$

Thus, the IV value of the entire variable is the sum of the IV values of the variables in each subgroup.

$$IV = \sum_{i}^{n} IV_i \tag{3}$$

where n is the number of variable groups. Based on the range of IV values, the strong and weak predictive power of the corresponding feature can be determined, and thus the feature selection is carried out, and the data after the feature selection is L_a.

Secondly, the mean and standard deviation are used to filter the data with strong distribution stability, and the stable data are selected as the training and test sets, and the features with less fluctuation in distribution are used for prediction. For each feature F_i, the data mean $\overline{F_i}$ and variance are calculated as follows, and the data after this feature selection is noted as L_b.

$$\overline{F_i} = \left(F_{x_1} + F_{x_2} + \ldots + F_{x_n}\right) / n \tag{4}$$

$$F_{\sigma^2} = \frac{\sum_{i=1}^{N}\left(F_{x_i} - \overline{F_i}\right)^2}{n} \tag{5}$$

Finally, the data is pre-trained using the LightGBM algorithm, and the feature importance function of the model is invoked to rank the importance of features and perform feature selection, and the data after this feature selection is noted as L_c.

4.3 Imbalanced Data Processing Module

First, for the lending dataset $L_\alpha\left(\alpha \in [a,b,c]\right)$, to perform iForest sample detection, for the non-equilibrium features in the lending data, many studies [16-19] classify majority and minority classes based on labels; however, in the borrowing scenario, normal and default borrowing samples have some similarity in borrowing patterns, and some data have important roles in the prediction of different classes or both, so it is under considered to simply classify classes based on labels. In this paper, we use the iForest [24] method to mix and sample the imbalanced

data, and define the "outliers" as the minority class data to reduce the learning difficulty of the classifier for the minority class while removing the noise data of the majority class.

The iForest is composed of multiple independent trees iTree, each of which is constructed through the following process.

(1) Randomly select an attribute.

(2) Randomly select a value of this attribute value.

(3) Classify each record according to the value taken from (2), and put the records less than value on the left subtree and those greater than or equal to on the right subtree.

(4) Recursive steps 1, 2, 3 until a condition is satisfied that the incoming data has only one or more identical records, or the height of the tree reaches the height threshold.

The process of prediction is to search the test records from the iTree root node to determine which leaf the predicted points fall on. iTree considers the minority class points to be rare, so they will be assigned to the leaf nodes soon, so in iTree, the minority class points behave as the path from the leaf nodes to the root node $h(x)$ is very short, so use $h(x)$ to determine whether a record belongs to a minority class $S(x, n)$:

$$S(x,n) = 2^{\left(-\frac{h(x)}{c(n)}\right)} \tag{6}$$

$$c(n) = 2H(n-1) - \left(\frac{2(n-1)}{n}\right) \tag{7}$$

$$H(k) = \ln(k) + \xi \tag{8}$$

where n is the sample size, and $h(x)$ is the height of the sample on the iTree, and $S(x,n) \in [0,1]$, the closer the value is to 1 means the sample belongs to the minority class, the closer it is to 0 means the sample belongs to the majority class, and ξ is the Euler constant.

We combine multiple iTrees to see and form an independent forest iForest. The method of constructing an iForest is similar to that of a random forest in that a part of the dataset is randomly sampled to construct each tree to ensure the difference between different trees, but the difference is that we need to limit the size of the samples. By the above process, we get the minority class sample set $L_\alpha^{\min}$ and the set of majority class samples $L_\alpha^{\max}$.

Second, we undersample and oversample the majority class samples and minority class samples by the Tomek Link algorithm [25] and SMOTE algorithm [26], respectively.

The central idea of the Tomek Link algorithm is that:

(1) In the minority class set $L_\alpha^{\min}$ and the majority class set $L_\alpha^{\max}$ select a minority class sample point in x_i, select a sample point in the majority class x_j, and $d(x_i, x_j)$ denotes x_i and x_j the Euclidean distance between.

(2) If for any x_k, there are x_i and x_j between $d(x_i, x_j)$ are minimal, it is considered that x_i and x_j are a pair of Tomek Link pairs.

(3) Then x_i and x_j of the boundary of the data can be considered as noise, delete x_j and write down the sampling result as L_α^1 .

The central idea of the SMOTE algorithm is that:

(1) For minority class sets $L_\alpha^{\min}$ and the majority class set $L_\alpha^{\max}$, the number of minority class samples is t and the number of majority class samples is p. Using Euclidean distance to find samples from the t samples of the minority class x_i of k nearest neighbors.

(2) Generate a random number while selecting a random sample among k nearest neighbors, which is between 0 and 1, i.e. $x_{\text{new}} = x + \text{rand}(0,1) \times (x_n - x)$, to obtain a new sample.

(3) Repeat (2) n times, then n new samples can be generated, and the generated samples are L_α^2 .

4.4 Result Prediction Module

After the above processing, we obtained three sets of sampled data, i.e. $L_\alpha^{1,2} = L_\alpha^1 + L_\alpha^2 \left(\alpha \in [a,b,c]\right)$, respectively, using the XGBoost model for training, and for a new sample, the prediction results of each of the three XGBoost models are noted as $\hat{x}_{L_a^{1,2}}$, $\hat{x}_{L_b^{1,2}}$, $\hat{x}_{L_c^{1,2}}$, then the final prediction results are

$$\hat{x} = Ticket\left(\hat{x}_{L_a^{1,2}}, \hat{x}_{L_b^{1,2}}, \hat{x}_{L_c^{1,2}}\right) \tag{9}$$

where the *Ticket* function is the voting function, i.e., the majority of the three XGBoost classifier predictions (normal users/defaulting users) is chosen as the final prediction result.

5 EXPERIMENT

5.1 Dataset

The missing data in the original dataset is relatively serious, and the missing rate of some variables is even close to 100%. To ensure the quality of the input data, we delete the variables with a missing rate of 50% or more. In addition, for some repetitive variables, we only keep a column of variables with higher quality, such as Title and Purpose both describe the purpose of loaning, then delete the Title variable.

The refined data have higher quality, which helps the model learn the valid information, effectively avoid the influence of redundant information on the model, and can significantly save the computational cost and improve the training efficiency. The subsequent experiments in this paper are based on the processed data, which contains 312,997 samples and 80 features. 20% of the data set is divided as the training set and 20% as the test set.

5.2 Experimental Setup

We implemented the I-XGBoost model using scikit-learn, and in the model training phase, the parameters of

XGBoost for the result prediction module were tuned and the optimal parameters were shown in Tab. 3.

Table 3 Experimental setup of result prediction module

Hyper parameters	value
eta	0.3
max_depth	3
gamma	0.4
lr	0.3

To demonstrate the excellence of the I-XGBoost model, the following method shown by Tab. 4 is chosen as the baseline model in this paper.

Table 4 Baseline model and description

model	description
XGBoost	eXtreme Gradient Boosting
NGBoost	Natural Gradient Boosting
AdaBoost	Adaptive Boosting
GBDT	Gradient Boost Decision Tree

Among them, XGBoost conducted experiments with four sets of parameters, and the parameter settings for each set of experiments are listed in the following Tab. 5.

Table 5 Experimental setup of XGBoost

model	n_estimators	booster	gamma	max_depth
XGB1	500	gbtree	0	3
XGB2	500	gbtree	0.3	3
XGB3	600	gbtree	0.3	6
XGB4	500	gblinear	0	None

5.3 Baseline Model

We used four models, XGBoost [27], NGBoost [28], AdaBoost [29], and GBDT [30], as baseline models, which have achieved good prediction performance in past studies. All four models use differentiated classifiers by building multiple classifiers and integrating the prediction results of each classifier with some strategy to improve the prediction ability, and their respective strategies are different.

The core of NGBoost is the use of natural gradients, which introduces probability prediction capability into gradient boosting, solving the technical problem that gradient boosting methods are difficult to handle real-valued probability prediction. The core of the AdaBoost model is Adaption, which automatically strengthens the samples that were misclassified in the previous weak classifier, and the weighted whole samples are used again to train the next basic classifier. And a new weak classifier is added in each round until some predefined sufficiently small error rate is reached or a pre-specified maximum number of iterations is reached. The GBDT model continuously reduces the residuals in the computation by continuously adding new trees to build a new model in the direction of residual reduction (negative gradient), i.e., the loss function reduces the residuals in the direction of the fastest speed. XGBoost uses multiple simple weak classifiers, and the idea of the algorithm is to keep adding weak classifiers to accommodate the deviations of previous classifiers. In contrast to the traditional GBDT, XGBoost adds a penalty term to the objective function, making the model much more generalizable, and supports downsampling of ranks, optimizing the computational speed.

5.4 Model Evaluation

The confusion matrix is a common metric for evaluating classification problems and consists of four components, *TP*, *FP*, *TN*, and *FN*, which have the following meanings.

True Positive, the true class of the sample is positive and the model identifies it as such. False Negative, the true class of the sample is positive but the model identifies it as negative. False Positive, the true class of the sample is negative but the model identifies it as such. True Negative, the true class of the sample is negative and the model identifies it as such.

Based on the confusion matrix, a series of evaluation metrics can be calculated, and the metrics commonly used to evaluate the performance of classifiers are *Accuracy*, *Precision*, *Recall*, and *F1-score*. *Precision* is the percentage of correct samples to the total number of positive samples. *Recall* is the percentage of correct samples to the total number of positive samples. The *F1-score* is an evaluation indicator that integrates the accuracy and recall rate, and a poor result of either metrics will lead to an unsatisfactory *F1-score* result.

In this paper, we conduct a study on the problem of default account detection, and pay more attention to whether we can accurately identify default samples in the sample, and set the sample whose loan status is default as a positive sample. We hope that the model can achieve excellent results in identifying positive samples as a minority class, so we choose three metrics, *Precision*, *Recall*, and *F1-score*, as the evaluation index of each model in this experiment, calculated as follows.

$$precision = \frac{TP}{TP + FP} \tag{10}$$

$$recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1\text{–}score = \frac{2 * precision * recall}{precision + recall} \tag{12}$$

5.3 Results and Analysis

We compared the I-XGBoost model with four baseline models, XGBoost, NGBoost, AdaBoost, and GBDT, on the dataset, where the XGBoost model includes four parameter combinations, and Tab. 6 shows the *Precision*, *Recall*, and *F1-score* metrics of each model.

Overall, the I-XGBoost model has the best performance in all metrics, followed by XGBoost using gradient boosting tree as a weak classifier. Although XGBoost using parameter combination 1 has a slightly higher F1-score, the I-XGBoost model improves Precision by 4.44 percentage points and Recall by 1.40 percentage points over XGBoost. Recall improves by 1.40 percentage points, and since we set the defaulted samples as positive samples, the better Precision and Recall indicate that the I-XGBoost model achieves better results in identifying potential defaulted accounts, which is in line with realistic needs. XGBoost using a linear model as a weak classifier

is the least effective, and our I-XGBoost model Recall improves by 44 percentage points compared to XGBoost using a linear model.

Table 6 Comparison between I-XGBoost and the baseline model

model	metric		
	precision	*recall*	*F1-score*
I-XGBoost	0.94	0.72	0.77
XGB1	0.90	0.71	0.78
XGB2	0.94	0.71	0.77
XGB3	0.94	0.71	0.77
XGB4	0.40	0.50	0.45
NGBoost	0.94	0.71	0.77
AdaBoost	0.92	0.72	0.77
GBDT	0.94	0.72	0.77

To address the problem of data imbalance, the I-XGBoost model addresses and mitigates data noise through feature selection, imbalanced data category detection, and imbalanced data processing. The experimental results show that I-XGBoost effectively solves the data imbalance phenomenon while extracting data features.

This is due to the fact that the feature selection module improves the quality of features used for final prediction in three aspects: contribution of data to prediction, stability and importance. In addition, the imbalanced data processing module uses iForest sample detection to reduce the learning difficulty of the classifier for the minority class while removing the majority class noisy data, and then undersamples and oversamples the majority class samples and the minority class samples to effectively deal with the data imbalance phenomenon in the loan data.

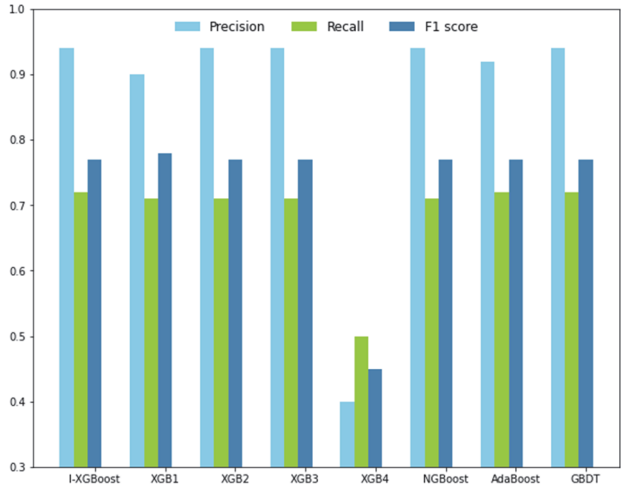

Figure 3 Comparison of prediction results of each model

In terms of evaluation indicators, *Recall* is lower than *Precision*, which indicates that the model is more accurate in identifying accounts with default risk, but there is still some room for improvement in the ability to identify these risky accounts. In future research, we should continue to optimize the model's processing of imbalanced data to improve the predictive ability for default risk.

6 CONCLUSION

In this study, we use the financial loan dataset of Lending Club to validate I-XGBoost, a loan default prediction method based on feature selection, data balancing processing with XGBoost, which can automatically and effectively extract valuable features, solve the data noise and data imbalance problems, and make more accurate analysis and prediction.

This study provides a new method that combines feature extraction with imbalanced data processing, and verifies the effectiveness of this combination. In the comparison with XGBoost, NGBoost, AdaBoost and GBDT models, we found that I-XGBoost has significantly improved the prediction accuracy and recall rate. This case provides a solution to the problem of high noise and imbalance in financial data, and provides a useful reference for further research by other scholars in this field.

Acknowledgment

This work was supported by the Shanghai Sailing Project from Shanghai Science and Technology Committee (Grant No. 20YF1434000).

7 REFERENCES

[1] Xiao, M. (2022). Supervision strategy analysis on price discrimination of e-commerce company in the context of big data based on four-party evolutionary game. *Computational Intelligence and Neuroscience*, 2900286. https://doi.org/10.1155/2022/2900286

[2] Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of default in P2P lending. *PloS one*, *10*(10), e0139427. https://doi.org/10.1371/journal.pone.0139427

[3] Si, Y. W. & Yin, J. (2013). OBST-based segmentation approach to financial time series. *Engineering Applications of Artificial Intelligence*, *26*(10), 2581-2596. https://doi.org/10.1016/j.engappai.2013.08.015

[4] Rajan, N., George, A., Saravanan, S. V., Kavitha, J., & Gopalakrishnan, C. S. (2022). An analysis on customer perception towards fintech adoption. *Journal of Logistics, Informatics and Service Science*, *9*(3), 146-158. https://doi.org/10.33168/LISS.2022.0311

[5] Altman, E. I., Resi, A., & Sironi, A. (2002). The link between default and recovery rates: effects on the procyclicality of regulatory capital ratios.

[6] Miao, L.-Y. &Chen, J.-L. (2014). The impact of social capital on borrowers' default risk in P2P network lending: the case of Prosper. *Financial Forum, 19*(3), 9-15.

[7] Brown, I. & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446. https://doi.org/3453.10.1016/j.eswa.2011.09.033

[8] Jin, Y. & Zhu, Y. (2015). A data-driven approach to predict default risk of loan for online peer-to-peer (P2P) lending. *Fifth international conference on communication systems and network technologies*, *IEEE*, 609-613.

[9] Luo, C., Wu, D., & Wu, D. (2017). A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence*, *65*, 465-470. https://doi.org/10.1016/j.engappai.2016.12.002

[10] Neagoe, V. E., Ciotec, A. D., & Cucu, G. S. (2018). Deep convolutional neural networks versus multilayer perceptron for financial prediction. *2018 International Conference on Communications (COMM)*, 201-206.

[11] Roy, A., Sun, J., Mahoney, R., Alonzi, L., Adams, S., & Beling, P. (2018, April). Deep learning detecting fraud in credit card transactions. *2018 Systems and Information Engineering Design Symposium (SIEDS)*, 129-134.

[12] Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P. E., He-Guelton, L., & Caelen, O. (2018). Sequence classification for credit-card fraud detection. *Expert Systems with Applications*, *100*, 234-245. https://doi.org/10.1016/j.eswa.2018.01.037

[13] Heryadi, Y. & Warnars, H. L. H. S. (2017, November). Learning temporal representation of transaction amount for fraudulent transaction recognition using CNN, Stacked LSTM, and CNN-LSTM. *2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, 84-89.

[14] Zemlickienė, V. (2019). Using TOPSIS method for assessing the commercial potential of biotechnologies. *Journal of System and Management Sciences*, *9*(1), 117-140. https://doi.org/10.33168/JSMS.2019.0107

[15] Kvamme, H., Sellereite, N., Aas, K., & Sjursen, S. (2018). Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications*, *102*, 207-217. https://doi.org/10.1016/j.eswa.2018.02.029

[16] Yu, H., Yang, X., Zheng, S., & Sun, C. (2018). Active learning from imbalanced data: A solution of online weighted extreme learning machine. *IEEE transactions on neural networks and learning systems*, *30*(4), 1088-1103. https://doi.org/10.1109/TNNLS.2018.2855446

[17] Sun, J., Lang, J., Fujita, H., & Li, H. (2018). Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. *Information Sciences*, *425*, 76-91. https://doi.org/10.1016/j.ins.2017.10.017

[18] He, H., Zhang, W., & Zhang, S. (2018). A novel ensemble method for credit scoring: Adaption of different imbalance ratios. *Expert Systems with Applications*, *98*, 105-117. https://doi.org/10.1016/j.eswa.2018.01.012

[19] Li, Y. & Vasconcelos, N. (2019). Repair: Removing representation bias by dataset resampling. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* 9572-9581.

[20] Liang, X. W., Jiang, A. P., Li, T., Xue, Y. Y., & Wang, G. T. (2020). LR-SMOTE-An improved unbalanced data set oversampling based on K-means and SVM. *Knowledge-Based Systems*, *196*, 105845. https://doi.org/10.1016/j.knosys.2020.105845

[21] Al-Shabi, M. A. (2019). Credit card fraud detection using autoencoder model in unbalanced datasets. *Journal of Advances in Mathematics and Computer Science*, *33*(5), 1-16. https://doi.org/10.9734/jamcs/2019/v33i530192

[22] Yu, L., Zhou, R., Tang, L., & Chen, R. (2018). A DBN-based resampling SVM ensemble learning paradigm for credit classification with imbalanced data. *Applied Soft Computing*, *69*, 192-202. https://doi.org/10.1016/j.asoc.2018.04.049

[23] Hui, L. (2021). *Research on Internet financial default prediction based on hybrid NGBoost*. Zhongnan University of Economics and Law.

[24] Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. *2008 eighth ieee international conference on data mining*, 413-422.

[25] Tomek I. (1976). Two Modifications of CNN. *IEEE Transactions on Systems Man & Cybernetics*, *SMC-6*(11), 769-772.

[26] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357. https://doi.org/10.1613/jair.953

[27] Chen, T. & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785-794.

[28] Duan, T., Anand, A., Ding, D. Y., Thai, K. K., Basu, S., Ng, A., & Schuler, A. (2020, November). Ngboost: Natural gradient boosting for probabilistic prediction. *International Conference on Machine Learning*, 2690-2700.

[29] Freund, Y. & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*(1), 119-139. https://doi.org/10.1006/jcss.1997.1504

[30] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232. https://doi.org/10.1214/aos/1013203451

[31] Wang, J., Rong, W., Zhang, Z., & Mei, D. (2022). Credit Debt Default Risk Assessment Based on the XGBoost Algorithm: An Empirical Study from China. *Wireless Communications and Mobile Computing*, 8005493. https://doi.org/10.1155/2022/8005493

[32] Ciurea, C., Chiriță, N., & Nica, I. (2022). A Practical Approach to Development and Validation of Credit Risk Models Based on Data Analysis. *Economic Computation & Economic Cybernetics Studies & Research*, *56*(3), 51-67. https://doi.org/10.24818/18423264/56.3.22.04

[33] Sheraz, M., Nasir, I., & Dedu, S. (2021). Extreme Value Analysis and Risk Assessment: A Case of Pakistan Stock Market. *Economic Computation & Economic Cybernetics Studies & Research*, *55*(3), 5-20. https://doi.org/10.24818/18423264/55.3.21.01

Contact information:

Wenlong LAI
Statistics and Information Department,
Shanghai Zheshang Borui Asset Management Research Company,
Shanghai 200023, China
E-mail: laiwenlong@zsamc.com