A study on personal credit risk assessment model based on integrated learning

Zhonghao Dong*

East China University of Science and Technology, 200237, Shanghai, China

*Corresponding author: 19040024@mail.ecust.edu.cn

Abstract. Financial risk control is an important means to control capital security. Under the background of market economy and inclusive finance, the demand for personal credit loans is increasing rapidly. Therefore, it is very important to control financial risks. However, in the financial field, due to the requirements of security, data is scarce and large-scale data cannot be obtained, so many researchers cannot carry out detailed data analysis and technical research, and can only carry out simple theoretical analysis. However, the current financial risk control mostly relies on manual operation, which is inefficient. In order to solve this problem, this paper firstly obtains a large number of personal credit loan information from foreign websites. Then, the feature engineering technology is used to clean the data. Finally, the data mining technology in the field of artificial intelligence is used for improvement, which greatly promotes the automation and intelligent process of financial risk control. In this paper, the data mining algorithm is used to classify the credit loan dataset. The results show that our model achieves good results and can effectively avoid nearly 70% default risk, and the time cost is very low.

Keywords: Credit Risk, Data Mining, Decision Tree, Integrated Learning, Logistic Regression.

1. Introduction

Financial risk control is an important means to control financial risks and ensure capital security. It is widely used in Internet finance and other industries, and plays a key role in maintaining the security of consumer finance, supply chain finance, credit lending, financial management platform, credit investigation platform and payment platform. Under the background of market economy and inclusive finance, the demand for credit loans for ordinary people is increasing rapidly, and the amount of loans for individual users is increasing day by day. Therefore, it is of great significance to control financial risks well.

Personal loan is an important part of financial risk control. With the increasing frequency and amount of personal loan, the application of measuring credit status and evaluating default risk is increasingly widespread, and it is also an indispensable business in personal credit. However, with the soaring demand for personal loans, the borrowers' solvency is uneven. If a large amount of money is lent to customers with weak solvency, the bank or enterprise may face great default risk, and the cash flow and capital chain of the lender will face high risk, which seriously endangers the operation of the enterprise. However, the traditional manual operation mode has many disadvantages such as low efficiency, low data processing dimension and deviation of previous experience, which makes it difficult to realize automatic decision-making of financial risks.

In order to solve these problems, we will introduce artificial intelligence and big data technology into the field of financial risk control. Due to the nature of the industry, banks have collected a large amount of structured data, which can be applied to big data analysis to solve the problem of financial risk. After obtaining the data set, the feature engineering is used to preprocess the original data, deal with the missing values and outliers, and normalize the original data set to complete the data cleaning. After that, data mining technology is adopted, and relevant algorithms are used to model according to various indicators of customer credit status. The corresponding model is obtained by training the large-scale data set, and the quality of the model is evaluated and improved by the test results. In this way, it analyzes and judges the customer's solvency, analyzes and predicts whether default risk will occur, and makes decisions on loan issuance according to the output results of the model.

The experimental results show that the financial risk control model proposed in this paper can make a good measure of the customer's solvency and make a very accurate prediction of the credit risk, which can be applied to the risk assessment of individual loans to improve the low efficiency of manual processing. To sum up, the financial risk control model proposed by using data mining algorithm in this paper has good reliability and innovation. Specific contributions are as follows:

(1) Acquired a large number of credit and loan information of individual bank customers from foreign websites.

(2) Feature engineering is adopted to clean the original data, including missing value processing, outlier processing, data conversion and other operations. Moreover, a variety of data mining algorithms are used to conduct repeated experiments and comparisons on the cleaned data, realizing the automation of personal credit loan evaluation, and simultaneously processing massive data sets efficiently and quickly.

(3) Experimental results show that the new algorithm proposed in this paper has achieved good results in financial risk control, and the prediction effect of default risk has a fairly high accuracy.

2. Related Work

With the popularity of personal credit and the development of artificial intelligence in China in recent years, many scholars have carried out research on the application of machine learning to the field of financial risk control. In "Research on Internet Financial Risk Control Model Based on Machine Learning Algorithm", Fan Jingjing started from users' consumption data on the Internet platform, operator data, Tongdun and other relevant data provided by third parties, using machine learning techniques such as logistic regression and support vector machine to build an effective risk control model, so as to help Internet financial enterprises better control lending risks; Liu Houqin combined the GBDT algorithm in the machine learning algorithm in the "Machine Learning Algorithm Credit Risk Prediction Model", and used the basic information, turnover records, user testing information and user testing scale of bank customers to conduct comprehensive evaluation, and divided customers into reputable customers and non-reputable customers; In "Data Mining of Lending Scenarios in the Field of Financial Risk Control", Guo Haoliang conducts risk control data mining based on the user's mobile phone application list and the user's previous lending behavior, and predicts the default risk of user lending through variable derivation and neural network model; In "Research on Internet Consumer Finance Risk Control Based on Feature Analysis", Xia Zhongxi used the Internet consumer finance industry data containing user consumption information, social information and credit information to establish a neural network model, a support vector machine model and a random forest model respectively, and the model evaluation showed that the neural network had the most ideal prediction effect on fraud risks.

In summary, there have been a large number of theoretical studies and practical results on the application of machine learning models to financial business scenarios, but there are still few literatures in the field of bank loan scenarios and customer credit status assessment, and there is a lack of relevant studies on the application effects of various ensemble learning algorithms. Therefore, based on the relevant information collected by foreign banks and through different ensemble learning algorithms, this paper trains the models for predicting default risk and tests its effect, and then compares and analyzes different models according to the test effect.

3. Model Framework

Financial risk control is an important means to ensure the security of banks and enterprises' funds. In the context of the gradual popularity of personal credit, the demand for financial risk control in the era is growing. In order to carry out financial risk assessment conveniently and efficiently, we obtained a large amount of information about customers' credit loans from foreign websites. After cleaning the data set, we used big data technology and intelligent algorithm to measure the credit

status of users from multidimensional indicators. And according to the result of the intelligent classification of the credit status of customers, the borrower's default risk is judged, so as to make the decision whether to lend money or not. Then the analysis results of the intelligent system are compared with the actual default situation to test the effect of the model and improve it.

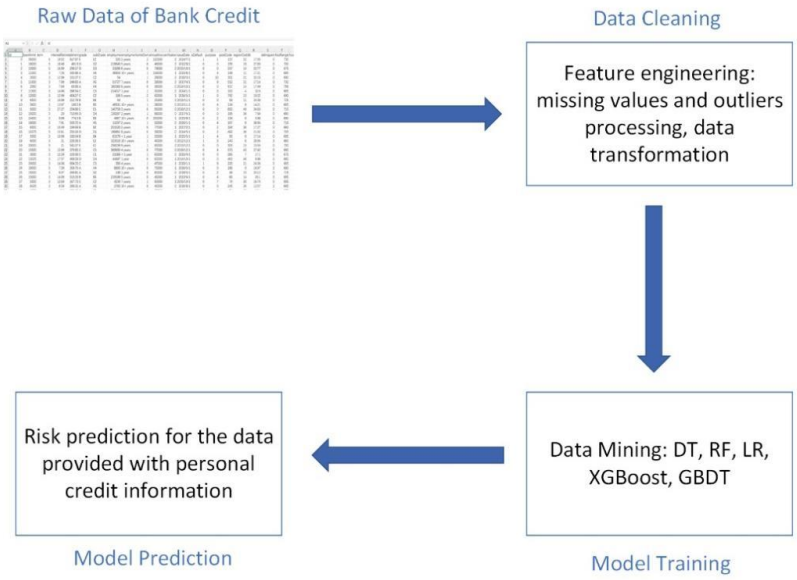

Figure 1. Model frame diagram

The framework of the entire model as shown in figure 1, which can be divided into the following three main links according to the different method of operation. (1) Data acquisition: obtain the dataset containing credit loan information about the lender's various asset status, loan transactions, maturity defaults and so on through foreign websites, and use them as training sets and test sets for models. (2) Data cleaning: preprocess data sets, fill missing values and correct outliers, and standardize corresponding data to facilitate subsequent modeling operations. (3) Model training: the training set is randomly selected from the complete data set, and the mathematical model of financial risk is constructed by using data processing or ensemble learning algorithms such as logistic regression, decision tree, random forest model and XGBoost model. (4) Model testing and prediction analysis: the remaining part of the data set is used as the test set to test the prediction effect of each model on credit default, and the accuracy of each model is compared horizontally to select the algorithm model with better effect.

This paper makes an intelligent analysis of the credit situation of foreign countries, and effectively avoids the problems of lack of data sources and insufficient personal credit information in China. Through the application of machine learning and artificial intelligence algorithms, the problems of large amount of audit, slow business process and low efficiency caused by manual analysis of financial risks are largely solved, and the intelligent decision-making of financial risk control is realized. And through the comparison of the effects of various models, the adaptation degree of different algorithms to this kind of situation is demonstrated, which provides the selection of the optimal scheme for financial risk control.

4. Model

Based on the above situation, this paper proposes a mathematical model to measure the customer's solvency and estimate the default risk, and outputs the results of intelligent decision-making for the issuance of personal loans. The model is roughly divided into the following three modules: (1) Data acquisition and statistical analysis: understand the data dimensions, select the main features, combine new features, and eliminate the features with low correlation; (2) Data cleaning: solve the situation

of missing features, abnormalities, and the need for conversion; (3) Model training: Preprocessed data sets are used to build and train models according to regression analysis, decision tree, Bagging ensemble learning, boosting ensemble learning and other algorithms.

4.1 Data acquisition and analysis

Obtain the credit status of customers' personal loans and final default information from foreign websites, and explain and analyze the dimensional parameters of the dataset.

Table 1. Dimensional parameters of the dataset

id	the unique identification of each loan business
loanAmnt	Loan amount
term	Loan Term (year)
interestRate	Loan interest rate
installment	Installment amount
grade	Loan grade
employmentTitle	Employment title
employmentLength	Length of employment (years).
homeOwnership	The status of ownership of house provided by the borrower at the time of registration
annualIncome	Annual income
issueDate	The date the loan was issued
purpose	The type of loan purpose at the time of loan application
postcode	The first 3 digits of the zip code provided by the borrower in the loan application
regionCode	The region code of the place where the loan was applied
dti	Debt-to-income ratio
delinquency_2years	The number of default events in the borrower's credit file that are more than 30 days overdue in the past 2 years
ficoRangeLow	The lower limit to which the borrower's FICO falls at the time the loan is issued
ficoRangeHigh	The upper limit to which the borrower's FICO falls at the time the loan is issued
openAcc	The number of outstanding credit limits in the borrower's credit file
pubRec	The number of disparaging public records
pubRecBankruptcies	The number of public record purges
revolBal	Total credit turnover balance
revolUtil	Revolving line utilization, or the amount of credit used by the borrower relative to all available revolving credits
totalAcc	The total number of current credit limits in the borrower's credit file
initialListStatus	The initial list status of the loan, which indicates whether the user is a first-time borrower
applicationType	Indicate whether the loan is an individual application or a joint application with two co-borrowers
earliesCreditLine	The month in which the borrower first reported the credit limit was opened
title	The name of the loan provided by the borrower
n-series anonymous features	Anonymous characteristics n0-n14, which are the processing of some borrower behavior count features

(1) In the data set, some features are not related to the assessment of financial risks in terms of common sense and logic, so some variables need to be deleted in data cleaning. “id” is the unique identification of each loan business, is to distinguish each business number, does not belong to the financial risk-related characteristics; “issueDate” and “earliesCreditLine” indicate the time when the loan is released and credit limit is opened, which are also not related to credit risk; “regionCode” and “postCode” indicate the region and postcode where the loan is applied, which are independent of the characteristics of financial risk. “title” refers to the loan name provided by the borrower, which does not represent the credit status of the borrower. “n-series anonymous features” represent the processing of some borrower behavior count features, without referring to specific event meaning, and are not highly correlated with financial risk. The above features are not necessarily related to the project, so they have little effect and are deleted.

(2) “loanAmnt”, ”term”, ”interestRate”, ”installment” and “grade” respectively indicate the amount, term, interest rate, grade and other characteristics of the loan, which can represent the difficulty of loan repayment; “employmentTitle”, “employmentLength”, “homeownership” and “annualIncome” respectively indicate the borrower's employment status, real estate, income and other characteristics, which can represent the borrower's solvency; purpose indicates the type of loan purpose, which is related to the risk of the borrower's successful return of funds, which in turn affects the ability to repay the debt; “dti” indicates the ratio of the borrower's debt to income, which can reflect the borrower's asset structure and measure the solvency; “delinquency_2years”, “ficoRangeLow”, “ficoRangeHigh”, “openAcc”, “pubRec” and “pubRecBankruptcies” respectively indicate the borrower's default status, credit score, outstanding balances and other characteristics, showing the negative information of the borrower's credit status; “revolBal”, “revolUtil” and “totalAcc” indicates the borrower's credit limit, and the amount of credit can reflect the borrower's credit quality; “initialListStatus” indicates whether the user is a first-time loan, and if the customer has successfully performed the contract, it means that the user may have better solvency; “applicationType" indicates whether the loan is an individual application or a joint application with two co-borrowers, and if it is a joint application, the solvency may be higher. The above characteristics are highly correlated with credit risk and can be used for model training.

Table 2. Dimension selection

Feature	Operation: Keep/Delete	Reason
loanAmnt	**Keep**	represent the difficulty of loan repayment
term	**Keep**	represent the difficulty of loan repayment
interestRate	**Keep**	represent the difficulty of loan repayment
installment	**Keep**	represent the difficulty of loan repayment
grade	**Keep**	represent the difficulty of loan repayment
employmentTitle	**Keep**	represent the borrower's solvency
employmentLength	**Keep**	represent the borrower's solvency
homeownership	**Keep**	represent the borrower's solvency
annualIncome	**Keep**	represent the borrower's solvency
purpose	**Keep**	It is related to the risk of the borrower's successful return of funds, which in turn affects the ability to repay the debt
dti	**Keep**	reflect the borrower's asset structure and measure the solvency
delinquency_2years	**Keep**	show the negative information of the borrower's credit status
ficoRangeLow	**Keep**	show the negative information of the borrower's credit status

ficoRangeHigh	**Keep**	show the negative information of the borrower's credit status
openAcc	**Keep**	show the negative information of the borrower's credit status
pubRec	**Keep**	show the negative information of the borrower's credit status
pubRecBankruptcies	**Keep**	show the negative information of the borrower's credit status
revolBal	**Keep**	the amount of it can reflect the borrower's credit quality
revolUtil	**Keep**	the amount of it can reflect the borrower's credit quality
totalAcc	**Keep**	the amount of it can reflect the borrower's credit quality
initialListStatus	**Keep**	if the customer has successfully performed the contract, it means that the user may have better solvency
applicationType	**Keep**	if it is a joint application, the solvency may be higher
id	**Delete**	does not belong to the financial risk-related characteristics
issueDate	**Delete**	business time is not associated with credit risk
earliesCreditLine	**Delete**	business time is not associated with credit risk
regionCode	**Delete**	the region is not related to the characteristics of financial risk
postCode	**Delete**	the region is not related to the characteristics of financial risk
title	**Delete**	the loan name does not represent the creditworthiness of the borrower
n-series anonymous features	**Delete**	It does not refer to the specific meaning of the event, and the correlation with financial risk is not high

4.2 Data Cleansing

Data cleansing is the process of re-examining and validating data to remove duplicate information, correct errors, and provide data consistency. Leverage relevant techniques such as mathematical statistics, data mining, or predefined cleansing rules to transform dirty data into data that meets data quality requirements.

4.2.1 Consistency check

Consistency checking is based on the reasonable value range and interrelationship of each variable, checking whether the data meets the requirements and finding data that is beyond the normal range, logically unreasonable, or contradictory. For example, variables measured on scales 1-7 with 0 values and negative body weight should be considered to be outside the normal range. Answers with logical inconsistencies can come in many forms, such as: many respondents say they drive to work and report no cars; Or respondents report that they are heavy buyers and users of a brand, but at the same time give a low score on the familiarity scale. When inconsistencies are found, individual case numbers, variable names, error categories, etc. need to be recorded for further verification and correction.

4.2.2 Handling of outliers and missing values

Due to survey, coding and entry errors, there may be some outliers and missing values in the data that need to be handled appropriately. Commonly used processing methods are: estimation, whole case deletion, variable deletion and pairwise deletion.

(1) Estimation. Estimation is the easiest way to deal with outliers and missing values by replacing outliers and missing values with the sample mean, median, or mode of a variable. This approach is simple, but the information already in the data is not fully taken into account, and the error may be large. Another approach is to estimate based on the respondent's answers to other questions, through correlation analysis or logical inference between variables. In this dataset, there are missing values in the indicators named “employmentLength”, “dti”, “pubRecBankruptcies” and “revolUtil”, and the statistics are incomplete. Here, the average value of the variable where the missing value resides is used for data filling.

(2) Delete the whole sample. Whole delete is to remove samples that contain missing values. As many questionnaires may have missing values, the result of this practice may be a significant reduction in the effective sample size, which may not make full use of the collected data. Therefore, it is only suitable for cases where key variables are missing, or where the proportion of samples containing outliers or missing values is small. When collecting and counting customer information, banks often take the form of a combination of mandatory and optional items. Although some customers fail to give some optional information, the mandatory information in the dataset is complete, so each piece of data can basically reflect the customer's credit status. Therefore, the data cleaning does not adopt the practice of deleting the whole case.

(3) Deletion of variables. Variable deletion is when a variable can be considered to be deleted if there are many outliers and missing values, and the variable is not particularly important for the problem under study. This reduced the number of variables available for analysis, but did not change the sample size. As shown in Section 5.1, the variables of this dataset are selected and the variables with too low correlation with default risk are deleted in this paper.

(4) Deletion in pairs. Pairwise deletion uses a special code to represent outliers and missing values while preserving all variables and samples in the data set. However, only samples with complete answers are used in the specific calculation, so the effective sample size of different analyses will be different due to the different variables involved. This is a conservative approach that maximizes the amount of information available in the dataset.

4.2.3 Numerical encoding of categorical features

Categorical characteristics mainly refer to sex, blood type, and other features that are only valued within a limited number of options. The original input of categorical features is usually in the form of strings, except for a few models such as decision trees that can directly process inputs in the form of strings, for models such as logistic regression and support vector machines, categorical features must be processed and converted into numeric features to work correctly. Common conversion forms are sequence number encoding, one-hot encoding, binary encoding.

(1) Ordinal encoding. Ordinal encoding is typically used to process data with size relationships between categories. For example, customer deposits can be divided into three grades: small, medium and large, and there is a "large > medium > small" ranking relationship. The ordinal encoding assigns a numerical ID to a category feature based on the size relationship. For example, the large grade is 3, the medium grade is 2, and the small grade is 1. The size relationship is still preserved after conversion.

(2) One-hot encoding. One-hot coding is usually used to deal with features that do not have a size relationship between categories. For example, if the financing channel can be used in four ways (bank, stock, bond, fund), the one-hot coding will turn the type into a 4-dimensional sparse vector, which is represented by (1,0,0,0) for bank, (0,1,0,0) for stock, (0,1,0,0) for bond and (0,0,0,1) for fund.

(3) Binary encoding. Binary encoding is mainly divided into two steps, first using the ordinal encoding to assign a category ID to each category, and then the corresponding binary encoding of the category ID as the result. Taking the financing channel as an example, the ID of the bank is 1, and

the binary representation is 001; the ID of the stock is 2, and the binary representation is 010; the ID of the bond is 3 and the binary representation is 011; the ID of the fund is 4 and the binary representation is 100. It can be seen that binary encoding essentially uses binary to hash map the ID, and finally obtains a 0/1 feature vector, and the dimensionality is less than the one-hot encoding, which saves storage space.

4.2.4 Numeric conversion and normalization

In order to eliminate the dimensional influence among data features, we need to normalize the features to make different indicators comparable. In the analysis of the influence of different independent variables on the dependent variable, the results of the analysis will obviously be inclined to the features with large numerical differences due to the influence of different dimensions of each feature. In order to get more accurate results, it is necessary to carry out feature normalization processing, so that all indicators are in the same numerical magnitude, so as to conduct analysis.

Normalizing the features of the numeric type can unify all the features into a roughly the same numeric interval. The most commonly used methods are mainly the following two.

(1) Min-Max Scaling. It performs a linear transformation of the original data so that the result is mapped to the range of [0,1], achieving equal proportional scaling of the original data. The normalization formula is as follows

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

Where X is the original data, Xmax and Xmin are the maximum and minimum values of the data respectively.

(2) Z-Score Normalization. It maps the raw data to a distribution with a mean of 0 and a standard deviation of 1. Specifically, assuming that the mean of the original feature is μ and the standard deviation is σ, the normalization formula is defined

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

Of course, data normalization is not essential. In practical applications, the modulus solved by gradient descent usually needs to be normalized, including linear regression, logistic regression, support vector machine, neural network and other models. However, the decision tree model is not applicable, taking C4.5 as an example, the decision tree is mainly based on the information gain ratio of the data set on the feature when it is divided, and the information gain ratio is independent of whether the feature is normalized, because the normalization does not change the information gain of the sample on the feature.

In this data set, due to the large number of data dimensions and the large difference in the numerical magnitude of different indicators, the range of order of magnitude among different indicators covers 10^0~10^5, the magnitude of the values varies greatly, and the comparability between different indicators is low, so the normalization operation is adopted for the dataset. The data of each indicator were normalized by Min-Max scaling, and all values were mapped to the interval of 0~1, which was convenient for subsequent model training and testing.

4.2.5 Combined features

Combinatorial features are also called feature crosses, that is, cross-combinations between different types or different dimensions of features, its main purpose is to make up for the early model in the CTR scene which cannot effectively combine features. With the progress of algorithm models, although some machine learning sorting models (GBDT + LR, FM, etc.) and deep learning sorting models can capture the connection between features, the early stage will still generate some combined features, for example, user behavior statistics under certain categories, data statistics under gender, etc.

Features in daily work scenes usually start from users and their corresponding attributes, and then cross combine with attributes or context features corresponding to items, such as: The combination of user identity and time characteristics, students on Saturday, Sunday and holidays, will have more abundant behavior in an entertainment APP, office workers in the morning and evening peak, will have more abundant behavior in an entertainment APP; In the news platform, the click times of young users under entertainment news will be higher, and the click times of adult users under social news will be higher; For the mutual combination of different financial characteristics of users, for example, to judge the customer's solvency, not only his asset status should be examined, but also his debt level should be considered. Only when the two are balanced can the disposable property level of customers be measured and their solvency be judged.

4.3 Model Training

4.3.1 Decision Tree model

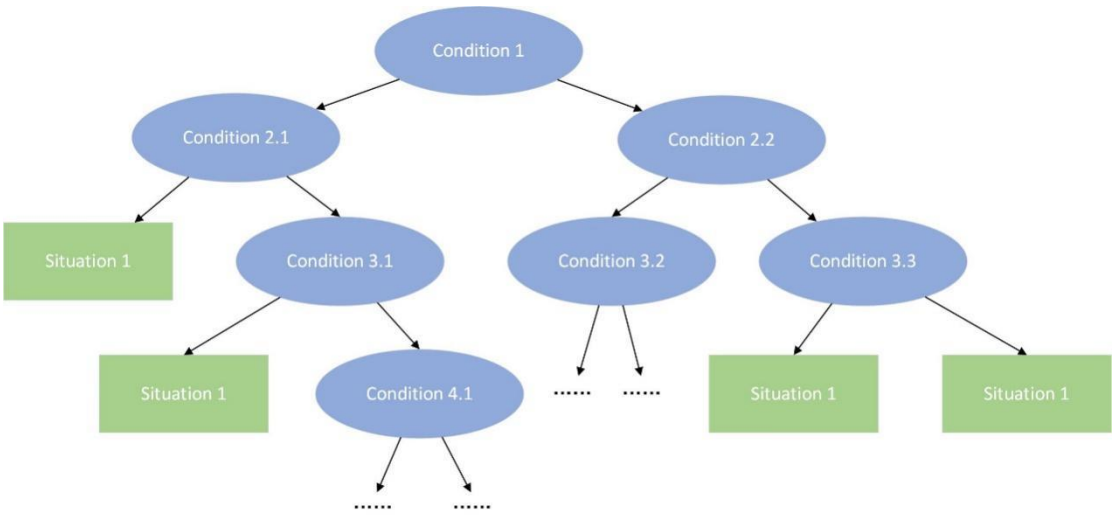

Figure 2. Schematic diagram of a decision tree

Decision tree is a very common classification method. It is a kind of supervised learning, given a bunch of samples, each sample has a set of attributes and a category, these categories are determined in advance, then through learning to get a classifier, this classifier can give the correct classification of the new object. A decision tree is a tree structure in which each internal node represents a test on an attribute, each branch represents a test output, and each leaf node represents a category. Decision tree represents a mapping relationship between object attributes and object values. Each node in the tree represents an object, each fork path represents a possible attribute value of, and each leaf represents the value of the object represented by the path from the root node to the leaf node. Decision tree is a frequently used technique in data mining, which can be used to analyze data as well as make predictions.

A decision tree represents a tree structure whose branches classify objects of that type by their attributes. Each decision tree can rely on a split of the source database for data testing. This process will recursively prune the tree. The recursion is completed when the split is no longer possible or when a separate class can be applied to a branch. The mainstream of decision tree algorithm is C4.5 decision tree, which inherits the advantages of traditional ID3 algorithm, but improves its disadvantages: (1) It uses information gain rate to select attributes, which overcomes the disadvantage of choosing attributes with more values when using information gain; (2) It is pruned in the process of tree construction; (3) It can complete the discrete processing of continuous attributes; (4) It can handle the missing data.

4.3.2 Random Forest model

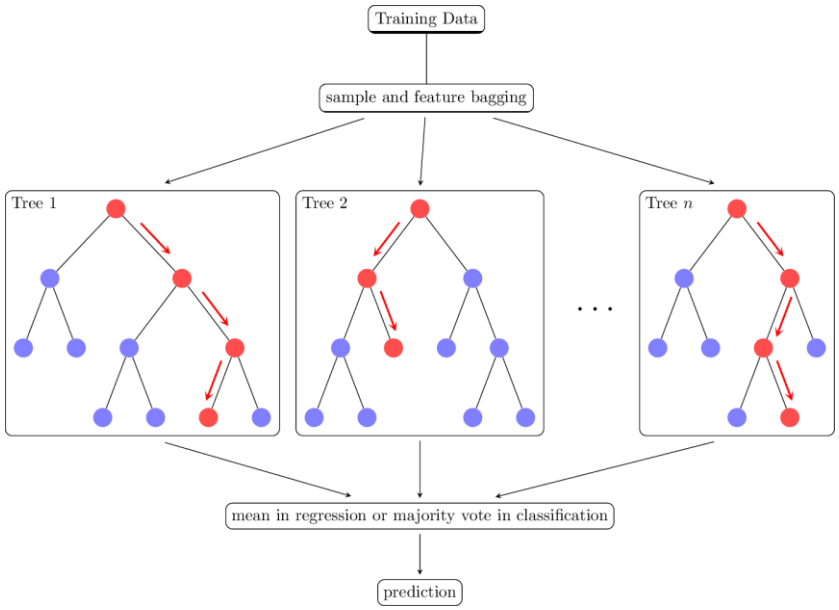

Figure 3. Schematic diagram of random forest

Random forest is a type of bagging machine learning algorithm. Bagging is a technique that reduces generalization error by combining several models. The main idea is to train several different models individually and then have all the models vote through the "model average" to test the output of the sample. For a given weak learning algorithm and training set, due to the low accuracy of a single weak learning algorithm, the learning algorithm is used multiple times to obtain a sequence of prediction functions, vote, and the final result accuracy will be improved.

Random Forest is a classifier that uses multiple decision trees to train and predict samples, and the category of its output is determined by the mode of the category output of each decision tree. In the process of algorithm model building, the sampling with put back is first taken from the original dataset, and the subdata set is constructed, and the amount of data in the subdataset is the same. Elements of different subdatasets can be duplicated, as can elements within the same subdataset. The sub-decision tree is then constructed using sub-datasets, and the data is placed in each sub-decision tree, and each sub-decision tree outputs a result. Finally, if there is new data that needs to be classified through a random forest, the classification result with the highest number of votes is the output of the random forest by voting on the judgment result of the sub-decision tree.

Similar to the random selection of a dataset, each splitting process of a subtree in a random forest does not use all the features to be selected, but randomly selects certain features from all the features, and then selects the optimal features from the randomly selected features. This allows decision trees in random forests to differ from each other, increasing the diversity of the system and improving classification performance.

4.3.3 Logistic Regression model

Logistic Regression is a machine learning method for solving binary classification problems to estimate the likelihood of something. For example, the possibility of a user buying a certain product, the possibility of a patient suffering from a certain disease, and the possibility of an advertisement being clicked by the user. Taking default risk analysis as an example, two groups of people are selected, one group is a default borrower, one group is a performance borrower, and the two groups of people must have different assets and credit status. Therefore, the dependent variable is whether it is in default, the value is "yes" or "no", and the independent variables can include income, savings, debt, credit history and many other factors, which can be either continuous or categorical. Then, through logistic regression analysis, the weights of the independent variables can be obtained, so that

it is possible to get a rough idea of which factors are the key factors for default. At the same time, according to this weight, the probability of default of a person can be predicted according to key factors.

Both Logistic Regression and Linear Regression are generalized Linear models. Logistic regression assumes that the dependent variable Y follows a Bernoulli distribution, while linear regression assumes that the dependent variable Y follows a Gaussian distribution, so logistic regression has much in common with linear regression. Without the Sigmoid mapping function, the logistic regression algorithm is a linear regression. It can be said that logistic regression is theoretically supported by linear regression, but logistic regression introduces nonlinear factors through Sigmoid function, so it is more suitable to deal with 0/1 classification problems.

4.3.4 XGBoost model

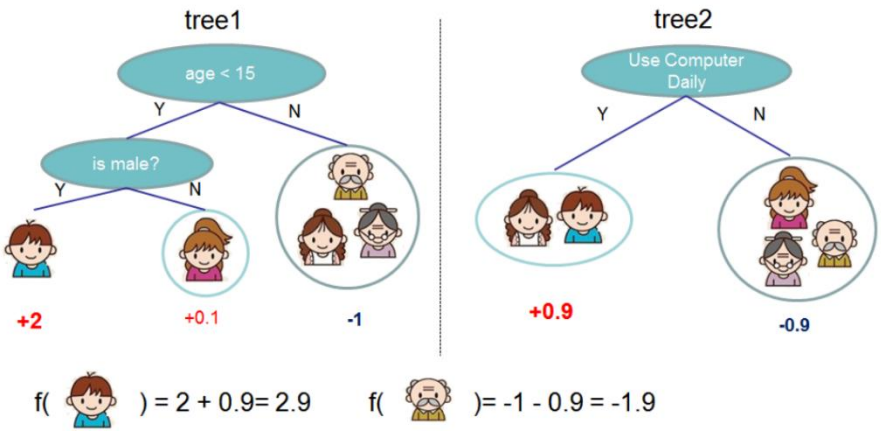

Figure 4. Schematic diagram of the Boosting algorithm

Boosting algorithm is a machine learning algorithm that can be used to reduce the bias in supervised learning. The sample subset is obtained by the operation of the sample set, and then the weak classification algorithm is used to train the sample subset to generate a series of base classifiers. It can be used to improve the recognition rate of other weak classification algorithms, that is, other weak classification algorithms are put into the Boosting framework as the base classification algorithm, and the training sample set is operated by the Boosting framework to get different training sample subsets, which are used to train and generate the base classifier. After each sample set is obtained, the base classification algorithm is used to generate a base classifier on the sample set. In this way, n base classifiers can be generated after a given number of training rounds n. Then Boosting framework algorithm will perform weighted fusion of these n base classifiers to produce a final result classifier. The recognition rate of each single classifier is not necessarily high, but their combined results have a high recognition rate, which improves the recognition rate of the weak classification algorithm.

XGBoost is an improvement on the gradient boosting algorithm, using Newton's method to solve the extremes of the loss function, expanding the loss function to the second order with Taylor expansion, and adding regularization terms to the loss function. The objective function at training consists of two parts, the first part is the gradient boosting algorithm loss, and the second part is the regularization term. The loss function is defined as

$$L(\phi) = \sum_{i=1}^{n} l(y_i', y_i) + \sum_k \Omega(f_k) \tag{3}$$

where n is the number of training function samples, l is the loss of a single sample, assuming it is a convex function, the predicted value of the model for the training sample, and the true label value of the training sample. Regularization terms define the complexity of the model:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|w\|^2 \quad (4)$$

where, γ and λ are artificially set parameters, w is the vector formed by the values of all leaf nodes in the decision tree, and T is the number of leaf nodes.

XGBoost is an optimized distributed gradient enhancement library designed to achieve efficiency, flexibility, and portability by implementing machine learning algorithms under the Gradient Boosting framework, providing parallel tree boosting that can solve many data science problems quickly and accurately.

4.3.5 GBDT model

Gradient Boosting Decision Tree is an additive model based on the idea of boosting integrated learning, which uses a forward distribution algorithm for greedy learning during training, and learns a regression tree at each iteration to fit the residual difference between the prediction results of the previous tree and the true value of the training sample. GBDT model can be explained strongly, the application effect is good, and it has been widely used in data mining, computational advertising, recommendation system and other fields.

Gradient Boosting is a large class of Boosting algorithms. Its basic idea is to train a new weak classifier according to the negative Gradient information of the loss function of the current model, and then combine the trained weak classifiers into the existing model in the form of accumulation. In each iteration, the negative gradient of the current model on all samples is calculated firstly, and then a new weak classifier is trained with this value as the target to fit and calculate the weight of the weak classifier, and finally the model is updated.

The Gradient Boosting algorithm that uses a decision tree as a weak classifier is called GBDT, and the decision tree used in GBDT is usually CART. Since GBDT is trained with residuals, in the process of prediction, we also need to add up the predicted values of all trees to get the final prediction result.

Gradient boost decision trees use additive models and forward step-by-step algorithms to implement the optimization process of learning, estimating in initialization the constant value that minimizes the loss function, which is a tree with only one root node. The value of the negative gradient of the loss function in the current model is then calculated as an estimate of the residual, and the regression leaf node region is estimated to fit an approximation of the residual. Then linear search is used to estimate the value of leaf node region, minimize the loss function, update the regression tree, and get the final output model.

In the construction of the model, this paper starts from the original data of bank credit, completes the cleaning of the original data and the refinement of the characteristics of the data set through feature engineering, and screens out the training set to provide a good foundation for model training. Through a variety of ensemble learning and data mining algorithms, the datasets are modeled and processed, and then the test sets are used to test different models, and the prediction effects of different algorithms on default risk are compared horizontally.

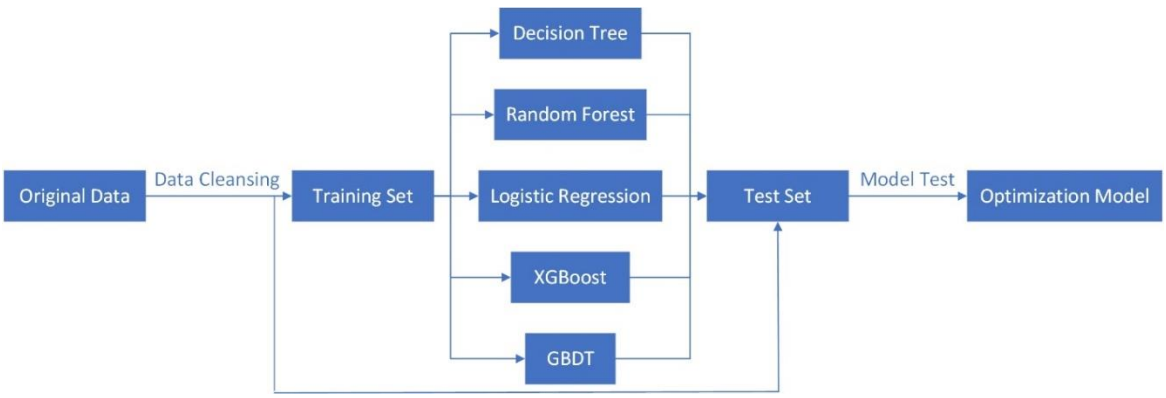

Figure 5. Model structure

5. Experimental Results and Analysis

The experiment uses the data set of personal credit status collected by foreign banks to conduct data mining and data analysis after feature engineering. Five models, Decision Tree, Random Forest, Logistic Regression, XGBoost and GBDT, were used to process the dataset, and the characteristics and advantages of the classifier models composed of single tree, multiple trees in parallel, multiple trees in series and generalized linear models were compared. Through the repeated debugging of the algorithm and parameters, the parameters such as the number of trees and the maximum number of depth features in the classification model are set to be in a reasonable state to ensure the normal effect of the model. Then the training set and the test set are divided by the ratio of 8:2 to train the model and evaluate the effect respectively. In view of the large gap in the number of positive and negative examples in the original data set, the role of positive and negative examples in the data set was balanced by adjusting the training weight. Then a new round of training was conducted on the model to compare the effects of the two rounds of models. Finally, the accuracy, precision, recall and F1 values obtained by model testing were used to evaluate the performance and characteristics of each model, and the different models were compared and analyzed.

5.1 Dataset

This paper obtains a data set containing 800,000 personal loan information from foreign websites, including 22 selected dimensions of credit status and the final default results. Out of 800,000 credit records, 159,610 loan transactions, or about 20 per cent of the total, were in default. From this dataset, machine learning is used to train the models to judge credit risk.

Table 3. Dataset characteristics

Total amount of data	Training set	Test set	Number of features	Amount of compliance data	Amount of default data
800000	640000	160000	22	640390	159610
Proportion	80%	20%	/	80.05%	19.95%

5.2 Experimental Setup

In the experiment, five models including decision tree, random forest, logistic regression, XGBoost and GBDT were used to process the dataset. According to the empirical value and multiple debugging results, the parameters of each model were set as follows: due to the large data dimension and amount, Gini coefficient was used to calculate the information entropy of decision tree to save computational power, the selection standard of feature cut-off point is chosen as best, and the depth of tree is set as 6. In the random forest model, samples are extracted without replacement, and there is only one opportunity for selection in each sample unit. The maximum number of features considered in constructing the optimal model of decision tree is set as 8, and the number of weak estimators in the integration, namely, the number of trees, is set as 100. In the logistic regression model, the value of C is 0.5 to improve the regularization effect and enhance the generalization ability of the model. The optimization method of loss function is chosen as liblinear. In the XGBoost and GBDT models, the depth of the tree is set to 6 and the number of trees is set to 100.

5.3 Model Training

The dataset is divided into training set and test set in an 8:2 ratio, and the training set is trained with the five algorithm models above. Then the data of the test set is input into the trained model to test the quality of the model, and the indicators of accuracy, precision, recall, F1 value and other indicators of the five models are obtained, and the advantages and disadvantages of each model are compared. In addition, since the defaulted data in the dataset only accounts for about 20% of the total data volume, the number of positive and negative cases is in an imbalanced state, and the resulting model is deformed. Therefore, through the adjustment of feature importance, the overall weight of

positive and negative examples is adjusted to a balanced state of 1:1, and the model is retrained and tested. In addition, the quality indicators of the five algorithm models after the feature importance adjustment are obtained, and the values of accuracy, precision, recall and F1 are compared with the original model to test the training effect of the new model.

5.4 Experimental Results and Analysis

Among the quality indicators of the model, accuracy represents the proportion of correct predictions in all prediction samples. Although this indicator can reflect the overall accuracy of model prediction, for data sets with severe imbalance between positive and negative examples, fairly high accuracy does not necessarily represent excellent application effect and may not meet the requirements of "mine clearance". Therefore, two indexes, precision and recall, should be introduced. From the perspective of prediction results, precision describes how many of the positive examples predicted by the binary classifier are true positive examples, that is, how many of the positive examples predicted by the binary classifier are correct. From the perspective of real results, recall describes how many true positive examples in the test set are selected by the binary classifier, that is, how many true positive examples are recognized and recalled by the binary classifier. Therefore, precision and recall are the key indicators to show the effectiveness of model training for dichotomous classification of deformed datasets. Precision and recall are usually two indexes, which are difficult to get both, and they restrict each other in large-scale data sets. Therefore, F1 value, namely the harmonic average of precision and recall, is introduced to comprehensively consider the results of precision and recall. When F1 is higher, the test method is more effective.

Before the loan, the bank made a rough survey on the assets, credit, income and other aspects of the customers and filtered out most of the customers who did not meet the loan conditions, which resulted in the problem of data imbalance. If we ignore the data imbalance and directly train the model, it will cause the deformity of the training model and affect the prediction ability. To solve this problem, we adjusted the data with imbalance as follows:

(1) The ratio of default data and non-default data is about 1:4;

(2) Resample operation was performed on the dataset to adjust the weight of positive and negative examples in the process of model training, so that the effect of the two was equivalent to a ratio of 1:1.

5.4.1 Performance evaluation of decision tree model

Table 4. Evaluation indicators of decision tree model

Dataset status	accuracy	precision	recall	F1
Before weight adjustment	80.60%	56.40%	2.28%	4.39%
After weight adjustment	62.45%	29.69%	67.65%	41.26%

As can be seen from the table, before the weight adjustment of positive and negative examples, although the model trained on the original data set has a high accuracy, the recall effect of the model is poor due to the deformity of the data set, so the F1 value is at a very low level, and the avoidance effect of financial risks is poor. After adjusting the weights of positive and negative examples, although accuracy decreased, recall rate and F1 value were greatly improved, and the practicability of the model was greatly improved.

5.4.2 Performance evaluation of random forest model

Table 5. Evaluation indicators of random forest model

Dataset status	accuracy	precision	recall	F1
Before weight adjustment	80.51%	50.15%	9.90%	16.54%
After weight adjustment	64.09%	30.21%	64.20%	41.08%

As can be seen from the table, under the joint action of multiple decision trees trained by randomly selected data from the total data set, accuracy before weight adjustment of positive and negative examples is also at a high level, and F1 value also shows better performance than decision tree. After the weight adjustment, the accuracy of random forest also decreased and F1 value also increased, but the change range of evaluation index was smaller than that of single decision tree, and the model showed greater robustness.

5.4.3 Performance evaluation of logistic regression model

Table 6. Evaluation indicators of logistic regression model

Dataset status	accuracy	precision	recall	F1
Before weight adjustment	80.61%	53.22%	4.90%	8.98%
After weight adjustment	65.99%	31.28%	62.16%	41.62%

Different from other algorithms, logistic regression algorithms are not based on decision trees for model building and testing. From the perspective of model effect, before the weight adjustment of positive and negative examples, the accuracy of logistic regression is similar to that of decision tree, and the F1 value is at a higher level. After the weight adjustment, both accuracy has decreased and the F1 value has also risen, and the F1 value has reached a similar level, but the accuracy of the logistic regression model is relatively better.

5.4.4 Performance evaluation of XGBoost model

Table 7. Evaluation indicators of XGBoost model

Dataset status	accuracy	precision	recall	F1
Before weight adjustment	80.82%	55.04%	9.08%	15.58%
After weight adjustment	64.50%	31.04%	67.14%	42.45%

According to the evaluation index of XGBoost, the performance of the model trained on the original data set is similar to that of the random forest model under the sequential approximation of multiple decision trees. After the weight adjustment, the model shows better optimization effect, and at the appropriate sacrifice of accuracy, the F1 value is greatly improved, which greatly improves the ability to avoid financial risks.

5.4.5 Performance evaluation of GBDT model

Table 8. Evaluation indicators of GBDT model

Dataset status	accuracy	precision	recall	F1
Before weight adjustment	80.83%	55.82%	8.03%	14.04%
After weight adjustment	64.50%	31.12%	67.58%	42.61%

Since GBDT and XGBoost are both Boosting ensemble learning algorithms, the experimental effects of the two models are similar, and it can be seen that the overall effect of Boosting classification algorithm is better than other types of algorithms.

5.4.6 Model effect comparison

Before the data set was generated, the bank staff had made a preliminary review on the credit status of the customers, resulting in a small proportion of positive examples in the data set. As a result, the trained model is more inclined to output the performance prediction result of "0" in order to meet higher accuracy, while the default result of "1" is extremely unlikely to output. That makes it difficult to successfully identify customers with default risk. This is evident when the ratio of positive and negative examples in the test set is adjusted to 1:1:

Table 9. Comparison of test effects of deformed models

	Decision Tree	Random Forest	Logistic Regression	XGBoost	GBDT
accuracy(Before the test set is adjusted)	80.60%	80.51%	80.61%	80.82%	80.83%
accuracy(After the test set is adjusted)	50.94%	53.70%	51.96%	53.67%	53.29%

It can be seen that when the ratio of positive and negative examples of the test set is adjusted from about 4:1 to 1:1, the accuracy of the model output result also drops from slightly more than80% to slightly more than 50%, which is enough to show that the model trained in the deformed state has an unreasonable tendency, and the output result is basically "0" ", the judgment on whether there is a risk of default is very weak.

From the above analysis, it can be seen that the performance of the five algorithm models is significantly improved after the positive and negative examples of the data set are adjusted to the equilibrium state. Therefore, the performance indicators of the five models adjusted by weights are compared to better judge the actual effect of each model.

Table 10. Model performance comparison

	accuracy	precision	recall	F1
Decision Tree	62.45%	29.69%	67.65%	41.26%
Random Forest	64.09%	30.21%	64.20%	41.08%
Logistic Regression	65.99%	31.28%	62.16%	41.62%
XGBoost	64.50%	31.04%	67.14%	42.45%
GBDT	64.50%	31.12%	67.58%	42.61%

As can be seen from the above table, due to the multi-tree parallel and mode voting method, the random forest model has better accuracy than the single decision tree. However, in terms of the actual utility of default risk aversion, the F1 value does not improve. The logistic regression model has a higher level of accuracy, but the improvement degree of F1 value reflecting risk aversion performance is still limited. Boosting model connects multiple weak classifiers in series to generate a strong classifier with stronger learning ability, so as to improve the recognition degree of classification algorithm. Moreover, the model is better than decision tree and random forest model both in accuracy and F1. XGBoost and GBDT models have shown better performance in both default prediction and risk aversion, and have shown better results overall, which can further avoid the occurrence of nearly 70% of default events under the action of the original credit screening mechanism, and ensure a low false positive rate, so as to avoid too many customers with solvency cannot apply for loans. It can better escort the development of inclusive finance.

6. Summary and Prospect

In the era of vigorous development of inclusive finance, personal credit has become an increasingly important form of loan, but if a large number of borrowers cannot repay their debts on time, the resulting bad debt losses will pose a great threat to the capital chain of banks or enterprises, which will lead to financial risks, so the assessment of borrowers' solvency and credit risk has become a necessary means to ensure the operation of financial industry and economy. In this paper, the decision tree, random forest, logistic regression, XGBoost, and GBDT algorithms are used to train the risk assessment model through data mining technology, and the binary classification model is used to intelligently analyze and make decisions on whether to lend. We can also further improve the prediction accuracy of the model through the optimization of algorithms and parameter settings, strengthen the learning and classification ability of the model, increase the precision and recall to

improve the identification and early warning ability of default risk, and minimize the occurrence of bad debts. In addition, we will conduct relevant research on the latest natural language processing model called BERT in the future, explore its application in the field of financial risk control, and further prevent and avoid financial risks.

References

[1] Fan Jingjing. Research on Internet financial risk control model based on machine learning algorithm[D]. Hangzhou Dianzi University,2019.

[2] Liu Houqin. Machine learning algorithm credit risk prediction model[J]. Microcomputer applications,2019,35(02):70-73.

[3] Guo Haoliang. Data mining of lending scenarios in the field of financial risk control[D].Dalian University of Technology,2021.DOI:10.26991/d.cnki.gdllu.2021.003316.

[4] Xia Zhongxi. Research on Internet consumer finance risk control based on feature analysis[D].South China University of Technology,2019.DOI:10.27151/d.cnki.ghnlu.2019.000982.

[5] Long Huihui. Research and prototype implementation of credit risk control early warning method based on machine learning[D]. University of Electronic Science and Technology of China, 2020.DOI: 10.27005/d.cnki.gdzku.2020.002889.

[6] Li Wei. Research on data-driven default risk prediction method of consumer finance[D]. Hefei University of Technology, 2019.DOI: 10.27101/d.cnki.ghfgu.2019.000509.

[7] Yang Zhigang. Analysis and forecasting of bank customer data based on interpretable machine learning models[D]. Lanzhou University, 2020.DOI: 10.27204/d.cnki.glzhu.2020.000358.

[8] Zhao Jinyang. Internet financial default prediction system based on feature engineering and multi-model integration[D]. Chongqing University of Posts and Telecommunications, 2020.DOI: 10.27675/d.cnki.gcydx.2020.000843.

[9] Hu Qingfeng.Credit risk prediction model of financial companies based on machine learning algorithm[J]. Electronics and Software Engineering,2021(10):146-147.

[10] Feng Jing. Research on personal credit risk assessment model based on BP neural network[D]. Taiyuan University of Technology,2017.

[11] Lin Yuli. Credit risk assessment model research[D]. Chongqing University,2014.

[12] HE Zijiao, OUYANG Hao, LIU Zhiqi, FU Junning, CHEN Zhuoting, XU Yue.Individual credit risk assessment model based on decision tree[J]. Information Technology and Informatization,2021(07):122-124.

[13] Niu Xiaojian, Ling Fei.Research on individual credit risk assessment model based on portfolio learning[J]. Fudan Journal (Natural Science Edition),2021,60(06): 703-719.DOI: 10.15943/j.cnki.fdxb-jns.2021.06.003.

[14] Zhao Jingda. Research on credit risk model of online loan based on machine learning algorithm[D]. University of International Business and Economics, 2021.DOI: 10.27015/d.cnki.gdwju.2021.000234.

[15] Wang Xiaoxiao, Wang Tingwen, Ma Yuling, Fan Jiayi, Cui Chaoran.Credit risk assessment method for P2P online loan borrowers based on deep forest[J]. computer science,2021,48(S2):429-434.

[16] Bian Lingzhi. Financial risk prediction and interpretability research based on enhanced deep forest[D]. Donghua University, 2022.DOI: 10.27012/d.cnki.gdhuu.2022.001518.

[17] He Xiaolong. Research on the effect of default prediction of machine learning models[D]. University of Electronic Science and Technology of China, 2021.DOI: 10.27005/d.cnki.gdzku.2021.003402.

[18] Zhang Lijuan.Research on individual credit risk assessment based on GBDT+LR model[J]. Network security technology and application,2021(07):40-42.

[19] He Haolong. Personal credit assessment methodology based on XGBoost and LR integration[D]. Shantou University, 2021.DOI: 10.27295/d.cnki.gstou.2021.000443.

[20] Zhou Qi. Research on personal credit risk assessment algorithm for categorical imbalance data[D]. Hebei University, 2020.DOI: 10.27103/d.cnki.ghebu.2020.001318.