RESEARCH ARTICLE

Research on Credit Customer Management Based on Customer Classification and Classification Preference

Li Chen[*,1]**, Xiaoling Luo**[1]**, Sijia Tang**[1] **& Meina Xie**[1]

[1]*Changsha Normal University, China*

Abstract: Accurate classification of credit customers is the premise of providing personalized credit services to them. According to customers' credit needs, we collect customer sample data, and then use users' repayment ability and repayment willingness to mark the samples. Bayesian classifier is constructed by constructing probability distribution function. By using test training and testing classification algorithm, it is found that Gaussian Bayesian algorithm can classify and predict data well. In the process of classifying samples, funds are allocated in combination with classification preferences. Experiments show that credit rating parameters have a significant impact on the optimization of resource allocation. By properly setting the values of credit rating parameters and classification preferences, it has reference value for reducing credit risks.

Keywords: credit risk; naive Bayes; classify; classification preference

1. Introduction

Credit policy is the embodiment of national economic policy in the supply of credit funds in a certain period. It consists of two parts: loan supply policy and loan interest rate policy. The loan supply policy specifies the investment, scale, support priorities, restricted objects of loans, and the overall goal of promoting national economic development; The loan interest rate policy defines the principle of the general level of the loan interest rate and the differential interest rate. The two are interrelated, complementary and play a role together. Combined with the current stage of China's loan policy, focus on supporting the development of energy and transportation enterprises; Give priority to supporting the textile industry to produce famous products and all kinds of marketable products; We will actively support the procurement of agricultural and sideline products and the production of products that earn exchange rates for export. For operating loss-making enterprises, enterprises with high product costs, poor

quality and no sales, and enterprises on the edge of closure and suspension, loans must be strictly controlled, and enterprises must be urged to seriously deal with the backlog of materials until loans are compressed and stopped. As of the end of 2017, the calibre is differentiated according to the five credit levels, and the resulting set of comments V= (normal, secondary, concern, doubtful, loss). Credit risk refers to the possibility that small, medium and micro enterprises will have a negative impact on the credit assets of commercial banks due to their own changes in various factors, resulting in the loss of bank credit assets, and eventually cause the value of credit assets or even the value of the whole bank to decline (Liu et al., 2019). The most common is the possibility that the debtor's credit enterprise will default due to its deteriorating financial situation. Therefore, when the credit officer receives a loan application, he will analyze the size of the credit risk from the two aspects of the lender's repayment willingness and repayment ability. There are many factors to be considered in the analysis of credit risk, and some commonly used methods have gradually formed in

Corresponding Author: Li Chen
Changsha Normal University, China
Email: 845019859@qq.com

the practice of credit risk management, among which the most commonly used are the 5C method and 7C method of credit analysis (Yan & Wu, 2009). 5C factor analysis method mainly focuses on the borrower's moral character, repayment ability, capital strength, guarantee, and business environment conditions to conduct a comprehensive qualitative analysis, in order to judge the borrower's repayment willingness and repayment ability. The method is to score each element one by one, so that the credit number is quantified, so as to determine its credit rating as whether it is a loan. 7C factor analysis is based on 5C to increase the sensitivity of controlling and dealing with risk factors. The expert method is mainly a management method for credit officers to judge and measure loan risks and make corresponding decisions by analyzing the main factors that may affect the repayment of the borrower's capital and interest. The comprehensive analysis of financial ratio is a problem that financial institutions usually transform the measurement of credit risk into the measurement of enterprise financial status. Therefore, the comprehensive analysis method of financial ratio is to analyze, explain and evaluate the financial situation and operation of the industry as a whole, systematic, comprehensive and comprehensive financial analysis index. Based on the financial data and experience information provided by enterprises, this paper generates the operating expenses and sales income indicators of each enterprise, classifies the enterprises applying for loans according to the existing data sample information, and provides reference value for reducing credit risks.

2. Related Research

Bayes algorithm has outstanding uncertainty expression ability and can make comprehensive use of prior knowledge and data sample information, so it is a common technical method for classification problems (Sinaga & Sinaga, 2020). Bayes classification algorithm has many branches, including naive Bayes classification algorithm, tree expansion naive Bayes classification algorithm, Bayesian network classification algorithm and Bayesian neural network classification algorithm. Despite the emergence of new Bayes classification algorithms, naive Bayes classification algorithm is still widely used in practical work because of its simple operation and relatively high classification accuracy, especially in the case of fewer feature attributes. Naive Bayes classification technique is widely used in various fields of finance and reliability analysis. Jin Shaohua et al. use Bayes theorem to determine the subsample, estimate the failure rate of the mixed Weibull population, and give a new analytical method to estimate the parameters of the mixed Weibull distribution. Bayesian Reliability Analysis, published by Martz et al., is the earliest monograph to study reliability from the perspective of Bayes. Kelly et al. discussed in detail the application of Bayesian inference in probabilistic risk assessment in their work . In the reference, naive Bayes, K nearest neighbor algorithm, support vector machine and other algorithms are used to properly classify stock market information, and the prediction results are compared with manual labeling. The experiment shows that support vector machine has the highest prediction accuracy, followed by naive Bayes.

Farughi et al. (2016) took the financial indicators of listed companies as the research object, built a naive Bayes classifier as a stock selection method, and confirmed that the cumulative return on investment was better than the benchmark return rate. In terms of text research, Fang Xiaoyu conducted data preprocessing on the collected messages and established a multi-label classification model for the public messages on the network platform by using naive Bayes algorithm, thus realizing automatic classification of massive messages. Customer classification research started in the 1950s. In recent years, with the development of big data technology, it has been widely used in personalized recommendation technology in e-commerce system, social system and we-media system (Fu et al., 2021). In the financial industry, customer classification is becoming more and more important to meet the

needs of personalized service. However, due to the customer-intensive service industry and the simultaneous existence of business complexity and regulatory complexity, data mining and machine learning technologies, including Bayesian technology, are widely used in the fields related to financial user classification. Shahbazi (2020) proposed a feature selection algorithm for the risk of the financial guarantee circle. They selected the feature subset from the feature set of financial customers and transactions that could cover most of the original data information without redundancy, and established the risk identification model of the guarantee circle by using logistic regression, RF, decision tree, Bayesian network and other technologies respectively, and found that logistic regression has the best risk identification accuracy. In the face of small and micro companies and self-employed households with limited time limit for capital injection, low amount of capital, high frequency, fast demand and the actual situation of lack of strong capital pledge, in the case of paying attention to the customer's first repayment source, that is, the borrower's own operating conditions and cash flow, the tool effectively evaluates the borrower's repayment willingness and repayment ability, and matches the appropriate loan amount according to the operating conditions. It mainly issues guaranteed loans that do not need to provide collateral, fully meets the needs of small and micro enterprises for funds, and has practical operational significance for the development mode of small and micro enterprises' credit business

3. Naive Bayes Credit Customer Classification Method

3.1. Naive Bayes classification

In machine learning, a naive Bayes classifier is a series of simple "probabilistic classifiers." They use Bayes' theorem to complete classification based on the assumption of strong independence between features. Have strong model representation, learning and reasoning skills, while showing high efficiency and high accuracy in learning small data sets (Wang, 2016).

There are k populations G1, G2,... Gk, and their prior probabilities are respectively. The density function of each population is, where x is an observed sample, and the posterior probability of the sample coming from the KTH population is:

$$P(j/x) = \frac{q_j p(x/j)}{p(x)} = \frac{q_j p(x/j)}{\sum_{i=1}^{k} q_i p_i(x)}, \quad j=1,2,\cdots\cdots,k$$

When, judge x to be from population j. Class-independent constants, the classification of samples depends on a prior distribution known as the class, which can be calculated, where is the number of training samples by analogy, is the total number of training samples, naive Bayes classification is a classification method based on the independent assumption of feature conditions, from which it can be assumed that, for discrete features, it can be statistically calculated, for continuous features, You can construct a probability distribution function.

3.2. Sample data collection and preprocessing

This paper is based on the analysis of A bank's credit business. First of all, the input and output invoices of the bank's existing credit customers are preliminarily sorted out, and whether the enterprise repays on time is registered. Then, the credit rating of the enterprise is classified into four categories: CLASS=(A, B, C, D) according to the loan supply policy and the operation of the enterprise, with A indicating normal and B indicating secondary. C represents concern, D represents loss, and CLASS is the target variable. Therefore, the original data sample set is constructed, in order to find the law from it, classify future credit customers, and finally realize the purpose of optimizing fund allocation and reducing bank credit risk.

Therefore, the classification of credit customers here is a multi-objective classification problem. Obviously, the classification process should pay more attention to "normal" customers (CLASS = A), who are the focus of consideration for lending, based on the information about whether they are paying on time. Banks first conduct preliminary screening of enterprises in need of loans according to the loan

supply policy, and take key support, priority support, active support, and other as the target objects. With the help of the financial invoices and repayment records of each enterprise, they classify and label each enterprise. Feature selection is the process of selecting the most important, task-relevant subset of features from the set of available features. Here, we choose four characteristics: ENTERPRISE, loan DEFAULT, EXPENSE and sales INCOME. On the one hand, it is because of the limitation of the data of loan enterprises that credit banks can obtain. On the other hand, the correlation between enterprise characteristics and credit rating classification is considered more, and the correlation between characteristics and categories can be observed by using data visualization technology.

Algorithm 1: indicates the enterprise type algorithm, which is used to mark the enterprise type

Input: Enterprise set U, type standard keyword K

Output: Enterprise type classification vector c

Begin:

1. c← [] / / Initializes the enterprise type class vector

2. for (u in U) {if(u.q $\in k_q$) or (u.q$\in k_t$) or(u.q $\in k_h$)or(u.q $\in k_b$):

/ / c← "Focus" / / Marked as "Focus"

3. else: c← "Priority" / / Marked as "Priority"

4. else: c← "Positive" / / Marked as "Positive"

5. else: c← "Not Supported" / / Marked as "Not Supported"

6. return c

End

3.3. Model building

Quantitative research on credit risk mainly focuses on the discriminant analysis of the financial predicament of loan enterprises; The analysis tool used is mainly Gaussian naive Bayes classification method in machine learning, and the data used is mostly static data (Wang, 2017). This paper mainly establishes the enterprise credit risk assessment index system from two aspects: quantitative index and qualitative index, the quantitative index includes:

operating expenses, sales income; Qualitative indicators include repayment intention and enterprise type. According to the enterprise financial invoice information and enterprise credit information, the sample data are analyzed and the characteristic attributes are obtained.

Assuming that the data of enterprise operating expenses and sales income meet the Gaussian distribution, equation is established

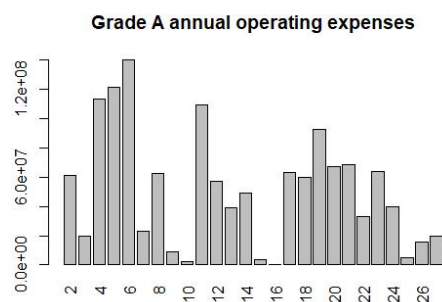$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(x-u)^2}{2\sigma^2} \right)$$

x is business operating expenses or sales revenue. The parameters of equation are determined by μ being the sample mean and σ being the sample standard deviation. Based on non-discretized data, a classification model is constructed using Gaussian Bayes algorithm and labeled GNB(GNB, GaussianNaive Bayesian). By using the test training and testing classification algorithm, it is found that the Gaussian naive Bayes classification algorithm has a good performance to evaluate and classify the predicted enterprises, and then optimize the resource allocation according to the credit rating parameters and classification preferences.
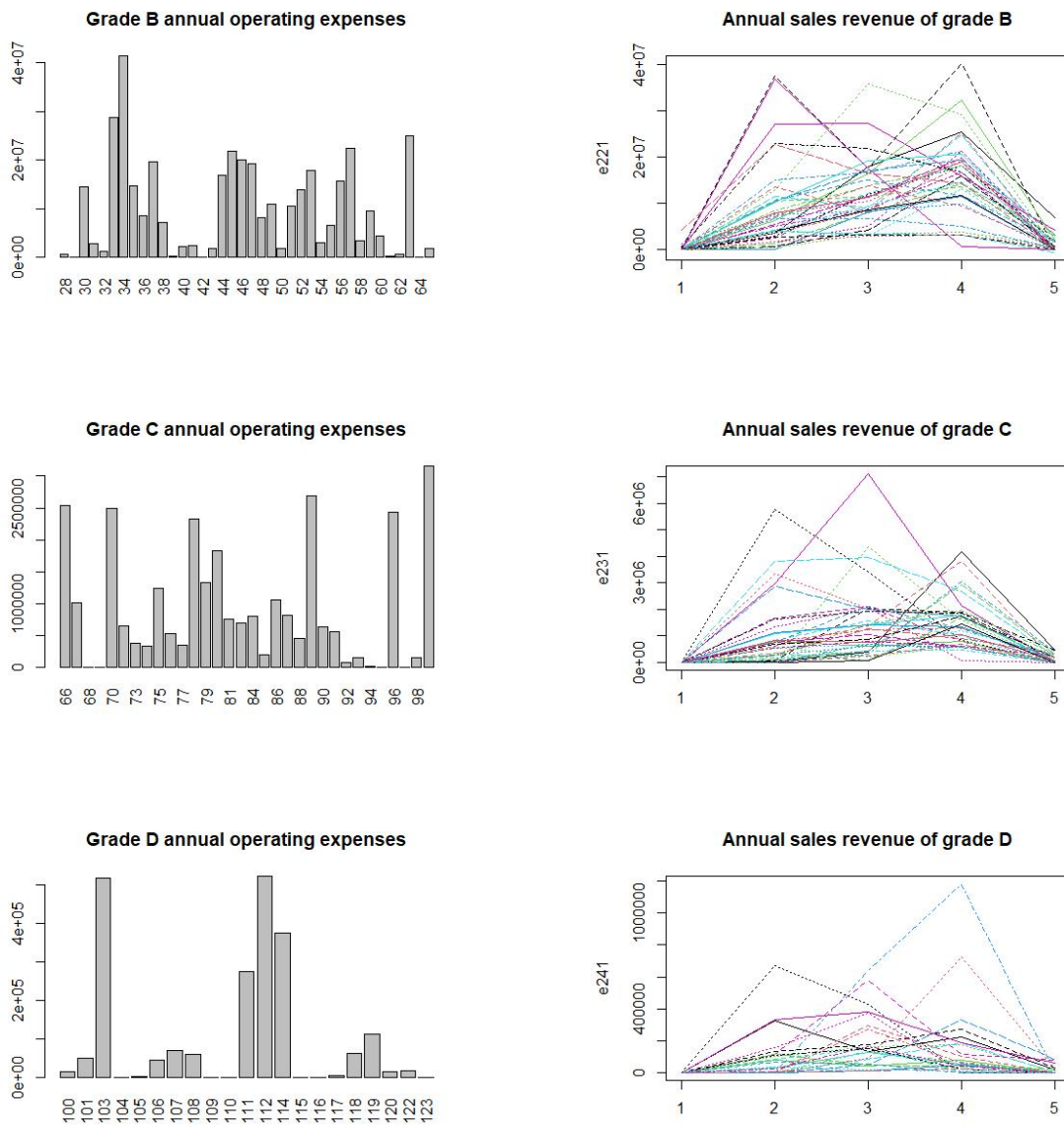
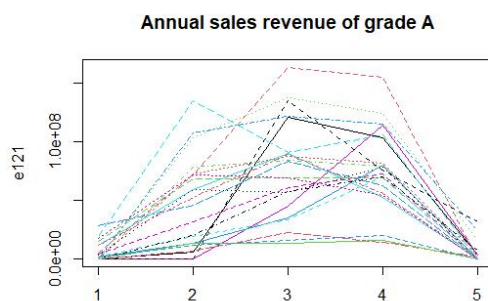## 4. Experiment

4.1. Data visualization analysis

By visual analysis of the data, the distribution characteristics of the feature variables and the correlation with the target variables can be directly analyzed.

Figue1

Annual operating expenses

Grade A annual operating expenses

Grade B annual operating expenses

Annual sales revenue of grade B

Grade C annual operating expenses

Annual sales revenue of grade C

Grade D annual operating expenses

Annual sales revenue of grade D

Figur2

Annual sales revenue

Annual sales revenue of grade A

It is obvious that operating expenses and sales income have better Gauss-like distribution characteristics. Therefore, the assumption of Gaussian distribution is made for the sample data.

4.2. Model training and testing

The experimental environment of this paper is operating system WIN10, python3 language and Pycharm environment. Due to the small sample size, this paper adopts a 5-fold cross-test, that is, 20% of the data samples are randomly selected each time as the test set, and the rest 80% of the sample data is used as the training set. The Gaussian naive Bayes classification model GNB was trained by using sample set data. accuracy is the core evaluation index

of classification algorithm. According to the problems studied in this paper, it is obvious to pay more attention to the classification accuracy of enterprises with "normal" loans (CLASS = A), that is, the recall rate of such customers. According to the confusion matrix, the recall rate is 0.95.

5. Conclusion

According to the demand of specific micro-enterprise loans, sample data of loan enterprises are collected. The probability distribution function is constructed, and four characteristics of ENTERPRISE, loan DEFAULT, EXPENSE and INCOME are selected to build a Bayesian network classifier. The performance of naive Bayes classification algorithm NBBDD is found to be the best. The next research will focus on the feature selection and other classification algorithms of loan enterprises. At the same time, the classification preference of various classifiers is also a meaningful research direction.

Conflict of Interest

The authors declare that they have no conflicts of interest to this work.

References

Liu, Y., Li, H, ..., & Zheng, B. (2019). Risk identification method of guarantee circle combined with feature selection and logistic regression. *Minicomputer Systems*, *40*(8), 1781–1787.

Yan, W., & Wu, W. (2009). *Data structure*. Tsinghua University Press.

Sinaga, J., & Sinaga, B. (2020). Data mining classification of filing credit customers without collateral with k-nearest neighbor algorithm (case study: PT. BPR diori double). *Journal of Computer Networks, Architecture and High Performance Computing*, *2*(1), 89–102.

Farughi, H., Alaniazar, S., & Mousavipour, S. hamed. (2016). An empirical study on customer risk management in banking industry: Applying k-means and RFM methods (evidence from two iranian private banks). *International Journal of Risk Assessment and Management*, *19*(4), 315–330.

Fu, Z., Zhang, B., Ou, L., Sun, K., Sun, X., & Chen, N. (2021). Research on enterprise financial customer classification method and preference based on intelligent algorithm. *Wireless Communications and Mobile Computing*, *15*(1), 115–120.

Shahbazi, F. (2020). Using decision tree classification algorithm to design and construct the credit rating model for banking customers. *IOSR Journal of Business and Management (IOSR-JBM)*, *21*(3), 123–130.

Wang, B. (2016). *Multivariate statistical analysis and R language modeling*. Jinan University Press.

Wang, M. (2017). *Bayesian statistical model and its application research*. Economic Science Press.