MEDWIN PUBLISHERS
Committed to Create Value for Researchers

International Journal of Biochemistry & Physiology
ISSN: 2577-4360

Logistic Model of Credit Risk during the COVID-19 Pandemic

Bin Zhao[1]* and Jinming Cao[2]

[1]School of Science, Hubei University of Technology, China

[2]School of Information and Mathematics, Yangtze University, China

***Corresponding author:** Dr. Bin Zhao, School of Science, Hubei University of Technology, Wuhan, Hubei, China Tel: +86 130 2851 7572; Email: zhaobin@hbut.edu.cn

Research article

Volume 6 Issue 1

Received Date: March 25, 2021

Published Date: April 28, 2021

DOI: 10.23880/ijbp-16000193

Abstract

In this paper, the Markov Chain Monte Carlo (MCMC) method is used to estimate the parameters of Logistic distribution, and this method is used to classify the credit risk levels of bank customers. OpenBUGS is bayesian analysis software based on MCMC method. This paper uses OpenBUGS software to give the bayesian estimation of the parameters of binomial logistic regression model and its corresponding confidence interval. The data used in this paper includes the values of 20 variables that may be related to the overdue credit of 1000 customers. First, the "Boruta" method is adopted to screen the quantitative indicators that have a significant impact on the overdue risk, and then the optimal segmentation method is used for subsection processing. Next, we filter three most useful qualitative variable According to the WOE and IV value, and treated as one hot variable. Finally, 10 variables were selected, and OpenBU-GS has been used to estimat the parameters of all variables. We can draw the following conclusions from the results: customer's credit history and existing state of the checking account have the greatest impact on a customer's delinquent risk, the bank should pay more attention to these two aspects when evaluating the risk level of the customer during the COVID-19 pandemic.

Keywords: Data Analysis; Monte Carlo Model; OpenBUGS; Overdue Risk

Introduction

The Markov Chain Monte Carlo method (MCMC), originated in the early 1950s, is a Monte Carlo method that is simulated by computer under the framework of Bayesian theory. This method introduces Markov process into Monte Carlo simulation, and achieves dynamic simulation in which the sampling distribution changes as the simulation progresses, which makes up for the shortcoming that traditional Monte Carlo integral can only simulate statically. MCMC is a simple and effective computing method, which is widely used in many fields, such as statistics, Bayes problems, computer problems and so on. Credit business also known as credit assets or loan business, which is the most important asset business of commercial Banks. By lending money, the principal and interest are recovered, and profits are obtained after deducting costs. Therefore, credit is the main means of profit for commercial Banks during the COVID-19 pandemic.

By expanding the loan scale, the bank can bring more income, and inject more power into the social economy, so that the economy can develop faster and better. However, with the expansion of credit scale, it is often accompanied by risks such as overdue credit. Banks can reduce credit overdue risk from two aspects, one way is to increase the credit overdue penalties, such as lowering the personal credit, dragging into the blacklist, and so on. With the rapid development of Internet, personal credit registry has more and more influence on the individual. A bad credit report will bring much inconvenience for the individual, so, in order to avoid the adverse impact on your credit report, borrowers tend to repay the loan on time, but these means are all belong to afterwards, although reduced the frequency of overdue frequency, but still caused a certain loss to the bank.

Selectively lending to "quality customers" can reduce Banks' credit costs even more if they anticipate the likelihood of delinquency in advance, before the customer takes out the loan.

How to identify whether the client is the "good customer" will need to collect overdue related information about the customer in advance, through the establishment of probability model between relevant variables and overdue, thus to rank the customer's risk grade of overdue, if customers's risk level is extremely high, the bank should choose to increase loan interest or refuse to reduce the risk of credit bank loans [1].

This paper includes the following three parts: model introduction and data description, data preprocessing and OpenBUGS simulation, and summary.

Overview of Logistic Model

If we want to use linear regression algorithm to solve the problem of a classification, (for classification, y value equal to0 or 1), but if you are using the linear regression, then assumes that the function of the output value may be greater than 1, or much less than zero, even if all the training sample label y is 0 or 1 but if algorithms get value is greater than 1 or far less than zero, will feel very strange.

So the algorithm we're going to study in the next section is called the logistic regression algorithm, and this algorithm has the property that its output value is always between 0 and 1. So logistic regression is a classification algorithm whose output is always between 0 and 1.

First, let's take a look at the LR of the dichotomy. The specific method is to map the regression value of each point to between 0 and 1 by using the SIGmoid function shown in Figure 1.

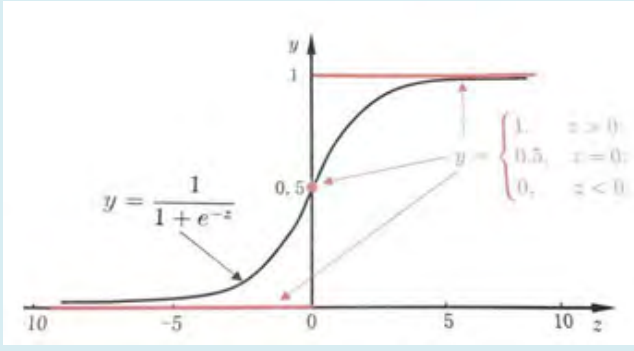

Figure 1: Sigmoid function.

As shown in the figure, let $z = w \times x + b$, When $z > 0$, the greater z is, the closer the sigmoid returns to 1 (but never more than 1). On the contrary, when $z < 0$, the smaller z is, the closer the sigmoid return value is to 0 (but never less than 0).

This means that when you have a binary classification task (positive cases corresponding labeled 1, counterexample corresponding labels 0) and samples of each of the sample space for linear regression $z = w \times x + b$, then the mapping using sigmoid function of g = sigmoid (z), and finally output the corresponding class label each sample (all value between 0 and the one greater than 0.5 is marked as positive example), then, two classification is completed. The final output can actually be regarded as the probability that the sample points belong to the positive example after the model calculation.

Thus, we can define the general model of the dichotomous LR as follows:

$$X \in R^n, Y \in \{0,1\}, w \in R^n, b \in R$$

$$p(\mathrm{Y}=1 \mid X)=\frac{\exp(w.x+b)}{1+\exp(w.x+b)}$$

$$p(\mathrm{Y}=0 \mid X)=\frac{1}{1+\exp(w.x+b)}$$

For a given input x, $p(\mathrm{Y}=1 \mid X)$ and $p(\mathrm{Y}=0 \mid X)$ can be obtained, and the instance x will be classified into the category with high probability value.

Odds of an event refer to the ratio between the probability of its occurrence and the probability of its non-occurrence. If the probability of its occurrence is P, the probability of the event is P/(1-P), and the log odds or logit function of the event is

$$\log it(p)=\log \frac{p}{1-p}$$

logistic regression can be obtained

$$\log \frac{p(Y=1 \mid X)}{1-p(Y=1 \mid X)}=w.x$$

That is, the logarithmic probability of output Y=1 in the logistic regression model is a linear function of input X.

When learning logistic regression models, for a given data set

$$T = \{(x_1, y_1), (x_2, y_2), ...(x_N, y_N),\}$$

$$x_i \in R^n, y_i \in \{0,1\}$$

The maximum likelihood estimation method can be used to estimate the model parameters, and then the logistic regression model can be obtained. Set

$$p(Y = 1 \mid x) = \pi(x), p(Y = 0 \mid x) = 1 - \pi(x)$$

The likelihood function is

$$\prod_{i=0}^{N} [\pi(x_i)]^{y_i} [1 - \pi(x_i)]^{1-y_i}$$

The logarithmic likelihood function is

$$L(w) = \sum_{i=1}^{N} [y_i \log \pi(x_i) + (1 - y_i) \log(1 - \pi(x_i))]$$

$$= \sum_{i=1}^{N} [y_i \log \frac{\pi(x_i)}{1 - \pi(x_i)} + \log(1 - \pi(x_i))]$$

$$= \sum_{i=1}^{N} [y_i (w . x_i) - \log(1 + \exp(w . x_i))]$$

By gradient descent algorithm and newton method can get the maximum value in the L(w) and the estimates of w : $\hat{w}$, then the logistic regression model:

$$p(Y = 1 \mid X) = \frac{\exp(\hat{w} . x + b)}{1 + \exp(\hat{w} . x + b)}$$

$$p(Y = 0 \mid X) = \frac{1}{1 + \exp(\hat{w} . x + b)}$$

MCMC

The formula of Markov Chain is as follows

$$p(X_{t+1} = x \mid X_t, X_{t-1,....}) = p(X_{t+1} = x \mid X_t)$$

That is, the state transition probability value is only related to the current state. Let *P* be the transition probability matrix, where P_{ij} represents the probability of the transition from *i* to *j*. So we can prove that $\lim_{n \to \infty} P_{ij}^n = \pi(j)$

$$\lim_{n \to \infty} p^n = \begin{bmatrix} \pi(1)\pi(2).....\pi(m) \\ \pi(1)\pi(2).....\pi(m) \\ \vdots \\ \pi(1)\pi(2).....\pi(m) \end{bmatrix}$$

Where π is the solution to $\pi p = \pi$ Since the probability of x obeys $\pi(x)$ after each transfer, it is possible to sample from $\pi(x)$ by transferring the different bases to this probability matrix. Then, given $\pi(x)$, we can construct the transition probability matrix by the Gibbs algorithm.

Gibbs Algorithm:

Random initialization $\{x_i : i = 1, 2, ... n\}$

for $t = 0, 1, 2, ...$ Cycle *Sampling*

$x_1^{(t+1)} \sim p(x_1 \mid x_2^{(t)}, x_3^{(t)}, ..., x_n^{(t)})$

$x_2^{(t+1)} \sim p(x_2 \mid x_2^{(t+1)}, x_3^{(t)}, ..., x_n^{(t)})$

.....

$x_j^{(t+1)} \sim p(x_j \mid x_1^{(t+1)}, ..., x_{j-1}^{(t+1)}, x_{j+1}^{(t+1)}, ..., x_n^{(t)})$

.....

$x_n^{(t+1)} \sim p(x_n \mid x_1^{(t+1)}, x_2^{(t+1)}, ..., x_{n-1}^{(t+1)})$

Data Description and Preprocessing

The German credit card data set is adopted in this paper, which contains 20 variables, including 7 quantitative variables and 13 qualitative variables. The details are shown in Table 1.

The data set includes 20 variables, the influence of different variables on credit overdue is different, adopting too many variables will not only increase the cost of collecting data, and wast customer's, also increases the complexity of the model, reduce the accuracy of prediction, so before to fitting the model we need to screen all the indicators which have a significant effect. The following content will be introduced from the screening of quantitative indicators and the screening of qualitative indicators.

"Boruta" Screening of Quantitative Indicators

The goal of Boruta is to select all feature sets related to dependent variables, which can help us understand the influencing factors of dependent variables more comprehensively, so as to conduct feature selection in a better and more efficient way.

Algorithm Process

1. Shuffle the values of various features of feature matrix X, and combine the post-shuffle features and the original real features to form a new feature matrix.
2. Using the new feature matrix as input, training can output the feature_importance model.
3. Calculate Z_score of real feature and shadow feature.
4. Find the maximum Z_score in the shadow features and mark it as $Z_{\max}$
5. The real feature whose Z_score is greater than Zmax is marked as "important", the real feature whose Z_score is significantly less than $Z_{\max}$ is marked as "unimportant", and is permanently removed from the feature set.
6. Delete all shadow features.
7. Repeat 1 to 6 times until all features are marked as "important" or "unimportant" The importance order of quantitative variables using the Boruta package of R software is shown in Figure 1.

Quantitative variable	qualitative variable	
duration	Purpose	property
credit_amount	credit_history	housing
installment_rate	checking_account_status	other_installment_plans
present_residence	savings	job
age	other_debtors	telephone
existing_credits	personal	foreign_worker
people_liable	present_employment	

Table 1: Data specification.

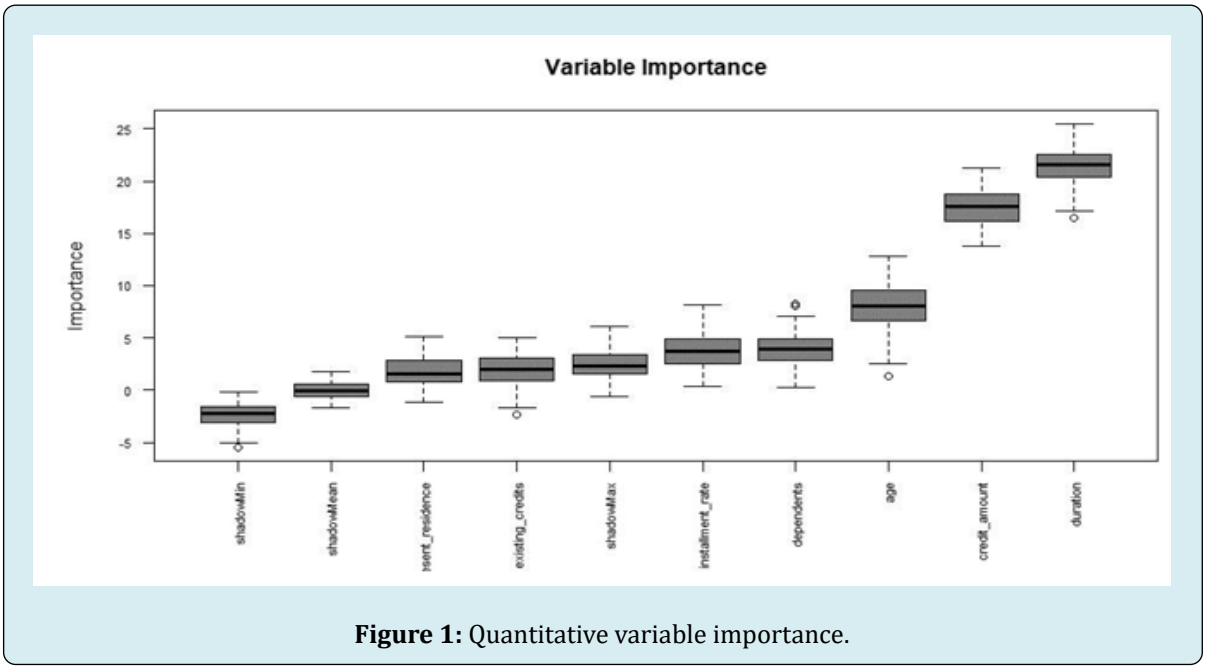

Figure 1: Quantitative variable importance.

The first three quantitative variables duration, creat_amount and age were selected into the model in order of importance, and the continuous variables were divided into boxes, with WOE (weight of evidence) and IV (Importance Value) values dividing the variables with the best predictive ability into groups

$$WOE_i = \ln(\frac{p_{good}}{p_{bad}}) = \ln(\frac{good_i / good_T}{bad_i / \text{bad}_T})$$

Where $good_i$ stands for the number of good tags in each group, good_T for the total number of good tags; The same for bad.

$$\text{IV} = \sum_{i=1}^{N} (\text{p}_{\text{good}} - \text{p}_{\text{bad}}) * WOE_i$$

Where N is the number of grouped groups, and IV can be used to represent the grouping ability of a variable, as shown in Table 2.

IV	Strength
<0.03	Extremely low
0.03~0.09	low
0.1~0.29	medium
0.3~0.49	high
>0.5	extremely high

Table 2: IV VS. Ability to predict.

To make the difference between groups as large as possible, smbinning package in R softwear is used to segment the continuous variable duration, credit_amount and age using the optimal segmenting method. The result of segmenting is shown in Figure 2.

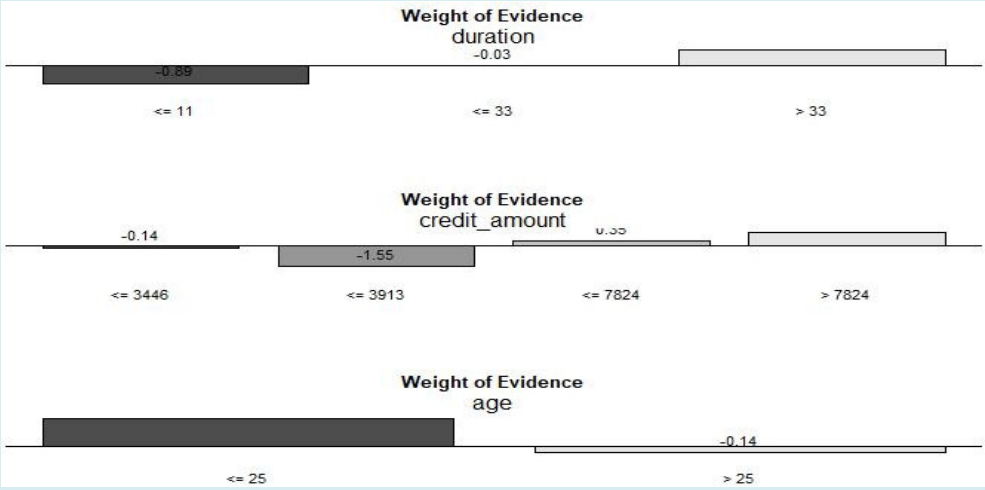

Figure 2: Optimal segmentation of continuous variables.

The Duration loan variable was divided into three sections: [0,11], (11,33) and (33,+ ∞); Credit_amount loan amount (a continuous variable) was divided into four paragraphs:

[0,3446] and (3446,3913], (3913,7824], (7824,+∞]; The Age of the Age applicant is divided into [0,25] and (25,+∞). All segments of the variable are corresponding to the WOE value with a large difference, indicating a large difference between groups. The IV value calculated according to the WOE value of the group are respectively duration: 0.225, Credit_amount: 0.229, age: 0.073.

Screening of Qualitative Indicators

IV values were calculated for all types of variables and sorted from high to low. Since the data only contained 1000 rows and the sample size was relatively small, only variables with large IV values, namely those with obvious classification effect, were selected in this paper. Three variables with IV values greater than 0.15 were selected, and the results were shown in Table 3.

Vars	IV
Account_status	0.666
Credit_history	0.2932
Savings	0.196

Table 3: Screening of qualitative indicators.

For each of the three selected qualitative indicators, the possible values of the variables are matched to a 0-1 variable. For example, the variables checking_account_status are treated as

$$A11=\begin{cases}1, \text{status}=A11\\ 0, \text{status}=\text{others}\end{cases} \quad A12=\begin{cases}1, \text{status}=A12\\ 0, \text{status}=\text{others}\end{cases}$$

$$A13=\begin{cases}1, \text{status}=A13\\ 0, \text{status}=\text{others}\end{cases} \quad A14=\begin{cases}1, \text{status}=A14\\ 0, \text{status}=\text{others}\end{cases}$$

variables A11, A12, A13, and A14, Variables may have values as shown in the table4.

Vars	Accuracy rate
Account_status	A11,A12,A13,A14
Credit_history	A30,A31,A32,A33,A34
Savings	A61,A62,A63,A64,A65

Table 4: Variables values.

Step Forward Likelihood Ratio Test

After preprocessing, there are 17 variables, not all of which have significant influence on the overdue risk. Before fitting the model, variables with small correlation and more significant significance are screened out, which can not only improve the classification accuracy of the model, but also simplify the model and reduce the cost of collecting customer information for Banks [2].

Variables with significant influence on overdue risk were selected by using the forward likelihood ratio test. The simulation results of SPSS are shown in Table 5.

Vars	LL	-2LL Change	DF	Sig.
A12	-500.591	4.050	1	.044
A14	-531.725	66.318	1	.000
A13	-503.329	9.526	1	.002
A21	-515.948	34.765	1	.000
A34	-503.507	9.882	1	.002
A30	-501.945	6.758	1	.009
A31	-503.137	9.141	1	.002
A61	-506.506	15.879	1	.000
A62	-500.883	4.634	1	.031
A131	-501.842	6.551	1	.010

Table 5: Step forward likelihood ratio test result.

Model Training and Prediction

The significance level was set as 0.05. A total of 10 variables were screened to establish the model:

$$\log it(\mathrm{p}_i) = \beta_0 + \beta_1 X_1[i] + \beta_2 X_2[i] + \ldots.\beta_{10} X_{10}[i]$$

Where β_o is the constant term, $\beta_i(\mathrm{i} = 1, 2, \ldots 10)$ is the partial regression coefficient of independent variable;
Parameters of the model have been given independent "non-informative" prior distribution, and OpenBUGS software is used for modeling and sampling, as well as Doodle modeling through OpenBUGS, to specify the distribution type and logical relationship of various parameters, as shown in the figure3:

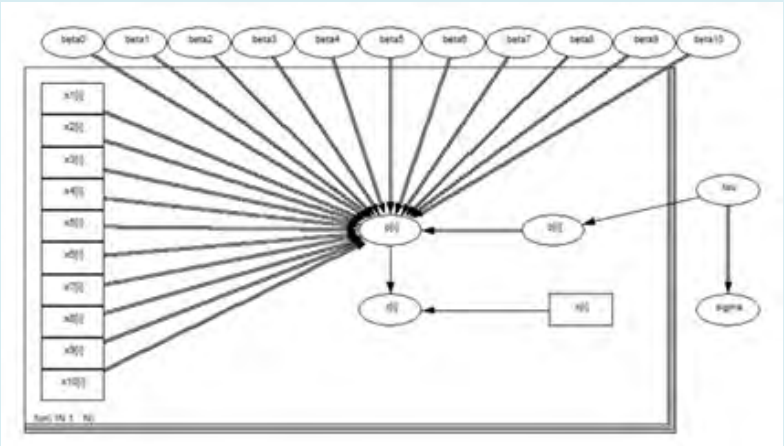

Figure 3: Doodle model.

Each ovals represent a node IN the graph, rectangle with constant node, single arrow from the parent node to the random child nodes, hollow double arrows indicate the parent node to the logical type child nodes, the rectangular outside for tablet, the lower left corner "for (I IN 1: N)" said for loop, is used to calculate the likelihood function of all samples, and the overall likelihood function is obtained [3].

The posterior distribution statistics for each parameter were obtained using OpenBUGS software, as shown in Table 6.

Where, MC_error represents the error of monte Carlo simulation and is used to measure the variance of the mean value of parameters caused by simulation.Val2.5 PC and VAL97.5 PC represent the lower and upper limits of the 95% confidence interval of the median, respectively; Median is usually more stable than mean; Start represents the starting point of Gibbs sampling. In order to eliminate the influence of initial value on sampling, sampling is started after 1001 times. Sample represents the total number of samples extracted. A total of 10,000 samples were extracted in this paper [4].

	mean	sd	MC_error	Val2.5pc	median	Val97.5pc	start	sample
beta0	-1.7150	0.5688	0.0077	-2.8060	-1.7320	-0.6263	1001	10000
beta1	-0.4220	0.0409	0.0022	-0.8879	-0.4183	-0.0158	1001	10000
beta2	-1.5840	0.0363	0.0022	-2.0860	-1.5780	-1.1000	1001	10000
beta3	-0.7002	0.0598	0.0025	-1.6370	-0.6884	-0.1802	1001	10000
beta4	0.8689	0.0218	0.0019	0.5103	0.8637	1.2330	1001	10000
beta5	-0.7936	0.0286	0.0009	-1.2660	-0.7879	-0.3396	1001	10000
beta6	0.8922	0.0759	0.0017	-0.2776	0.8786	2.1490	1001	10000
beta7	1.6560	0.0703	0.0020	0.0646	1.5850	3.6720	1001	10000
beta8	0.6818	0.0285	0.0015	0.2262	0.6755	1.1460	1001	10000
beta9	0.5608	0.0466	0.0016	-0.1907	0.5603	1.3200	1001	10000
beta10	-0.3846	0.0309	0.0029	-0.8548	-0.3755	0.0714	1001	10000
tau	270.9	0.0062	0.0000	5.065	105.0	1524.0	1001	10000

Table 6: Parameter estimation result of MCMC.

According to the parameters of Bayesian estimation, the error of model Colot simulation is generally relatively small, which indicates that the model has a good effect. With each parameter of the Gibbs sampling sample mean as a parameter to estimate, from the point of the results, the variable whether checking_account_status values for A13 (greater than 200 DM) and A14 (no checking account), variable credit_history whether values for A30 (not credit) and A31 (have to pay all the bank's credits) have bigger influence on the overdue risk, relative variable savings for A61 values (<100 dm) and A62 (100<x<500DM) has little impact on the overdue risk, indicating that the customer's historical credit history and the current check status have a greater impact on the overdue risk, that is, the customer's historical credit and current economic status have a greater impact on the overdue risk. Banks should focus on these two aspects when judging the customer's credit risk level [5].

The logistic regression equation can be obtained

$$\log it(p_i) = -1.732 - 0.4183X_1[i] - 1.578X_2[i] - 0.6884X_3[i] + 0.8637X_4[i] - 0.7879X_5[i] + 0.8786X_6[i] + 1.585X_7[i] + 0.6755X_8[i] + 0.5603X_9[i] - 0.3755X_{10}[i]$$

When dividing the overdue risk level of customers, there may be two wrong divisions, that is, dividing "high-quality customers" into high-risk customers and high-risk customers into "high-quality customers". Generally speaking, the economic costs of these two wrong divisions are different. For Banks, the cost matrix is shown in the table7 (0=Good, 1=Bad) [6].

	0	1
0	0	1
1	5	0

Table 7: Cost matrix.

The rows represent the actual classification and the columns the predicted classification. It is worse to class a

customer as good when they are bad (cost=5), than it is to class a customer as bad when they are good (cost=1). Define the loss function as

$$Loss = \sum_{i=1}^{1000} f(pre(i) - real(i))$$

$pre(i)$ and $real(i)$ is The classification results of the i t h sample and The actual category, respectively, and $f(x)$ is a piecewise function:

$$f(x) = \begin{cases} 5 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases}$$

Each sample input the results of logistic regression model as a probability value, if the probability value is greater than a given probability value is the sample classification is 1, otherwise the classification of 0, due to the loss of the two types of error, according to the different probability of loss matrix can be calculated at a given value under the condition of overall losses, the results as shown. It can be seen that when the given probability value is 0.21, the overall loss is the smallest. The Confusion matrix are shown in the table 8. The precision of the model is 85%. The model identifies the vast majority of high-risk customers.

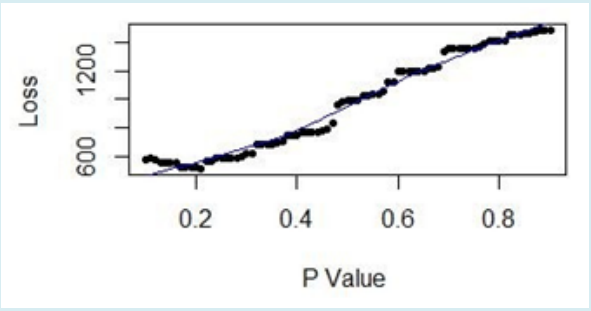

Figure 4: Loss diagram.

	0	1
0	513	187
1	45	255

Table 8: Confusion matrix.

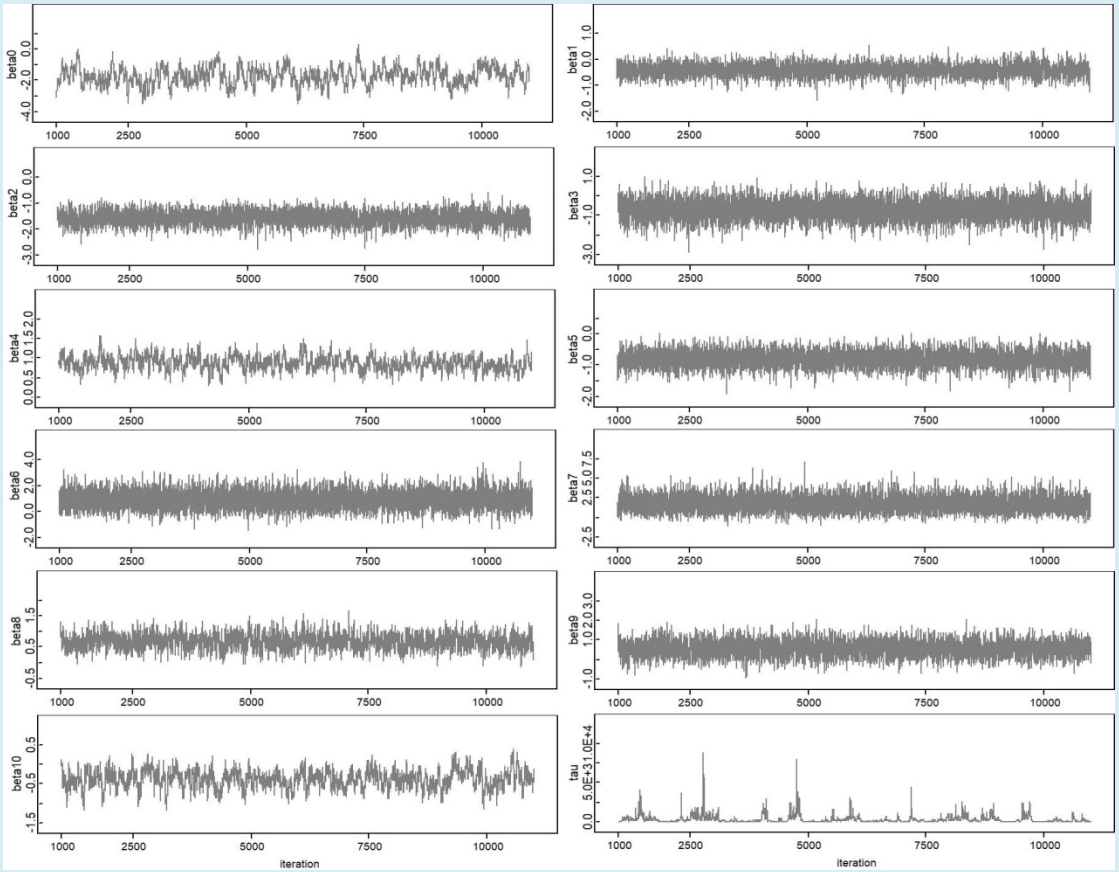

Figure 5: Iteration history of OpenBUGS.

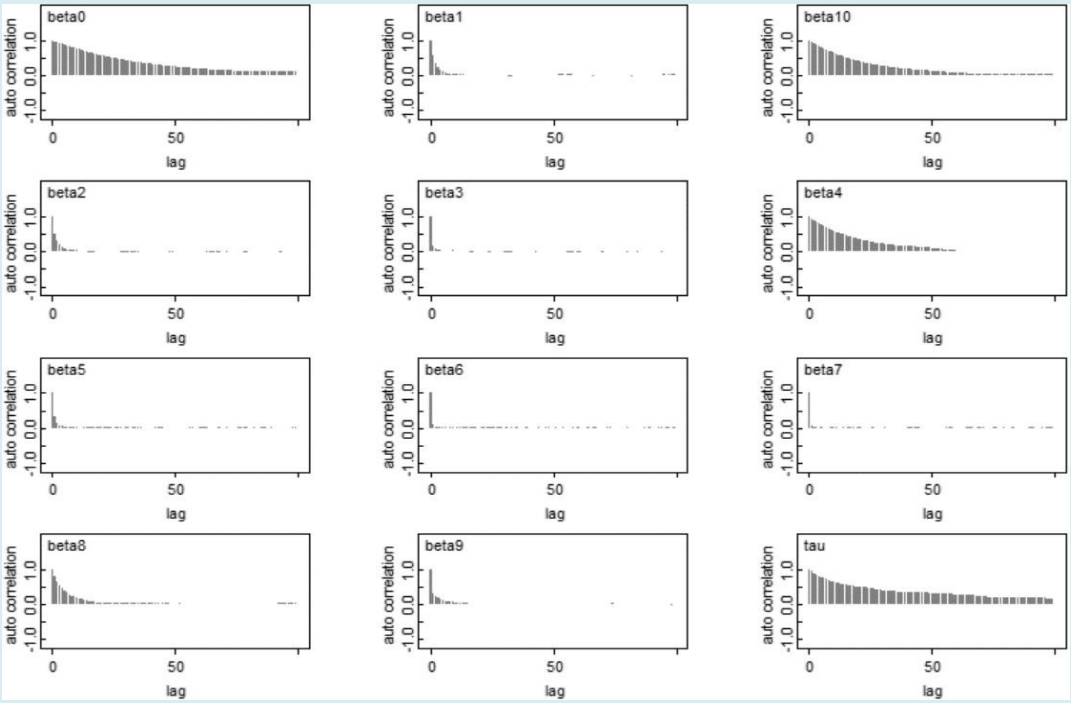

Figure 6: The Autocorrelation diagra of OpenBUGS.

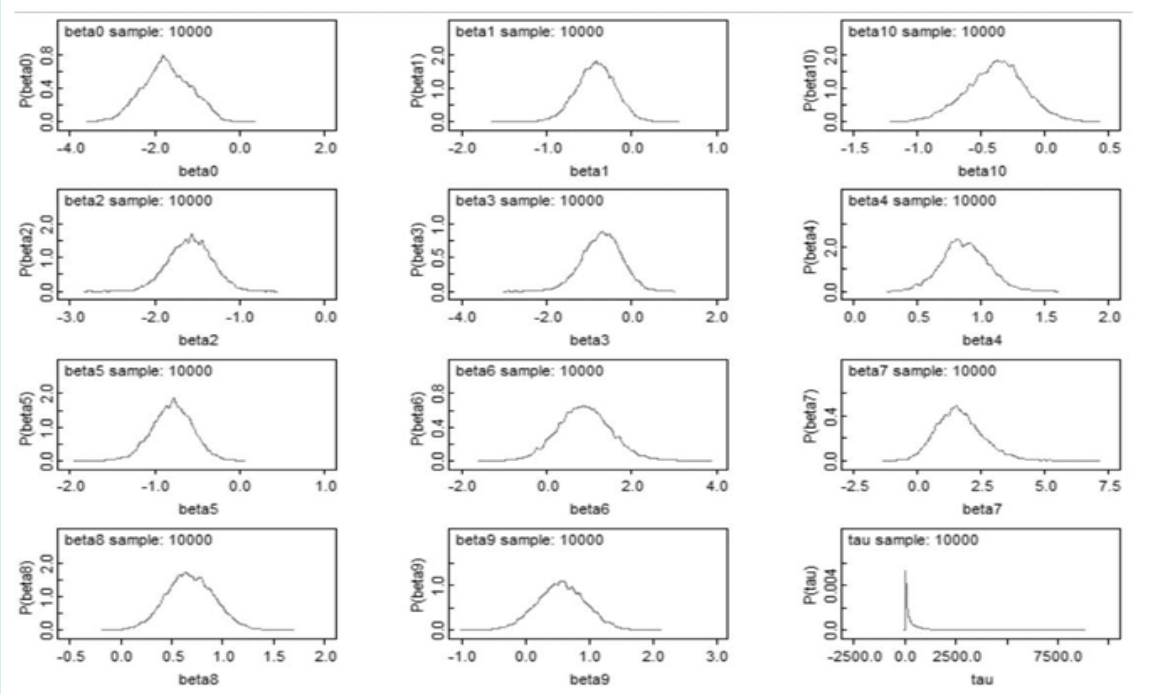

Figure 7: The Nuclear density of OpenBUGS.

Figure 5~7 shows the iteration history diagram, autocorrelation function diagram and kernel density diagram of all parameters.

Monte Carlo simulation starts from the initial value given for each parameter. Due to the randomness of extraction, the first part of extracted value is used as an independent

sample obtained by annealing algorithm. Therefore, we must judge the convergence of the extracted Markov Chain. The convergence of Markov chains can be analyzed according to the results of parameter extraction [7].

Iteration History Diagram

From the graphs in Fig. 5 we can safely conclude that the chains have converged as the plots exhibits no extended increasing or decreasing trends, rather it looks like a horizontal band.

Nuclear Density Figure

According to the distribution density of extracted samples, it can be seen that the samples extracted by Gibbs algorithm are mostly concentrated in a small area, which can also explain the convergence of Markov chain [8].

Autocorrelation Diagram

Autocorrelation plots clearly indicate that the chains are not at all autocorrelated. The later part is better since samples from the posterior distribution contained more information about the parameters than the succeeding draws. Almost negligible correlation is witnessed from the graphs in Fig. 6. So the samples may be considered as independent samples from the target distribution,
i.e. the posterior distribution [9].

Conclusion

This paper constructs a binomial logistic regression model based on the customer characteristic data of Banks.

Content mainly includes two parts, the first is the part of data pretreatment, the original data contains 20 variables, in order to make the model more concise, and improve the accuracy of classification model, reduce the cost of information collection and the time cost of customers, using "Boruta" method of screening of three quantitative indicators, and use the optimal segmentation method will be treated as continuous variable section.

Then, three qualitative variables were selected into the model by calculating the IV value of the variable, and the qualitative variables were treated with a unique heat type.

Two logIST-IC regression of SPSS software was used to screen out 10 variables with significance less than 0.05 into the model.

All the selected variables were brought into OpenBUGS software to obtain the parameter Bayesian estimation of the binomial logistic regression model. From the estimation results, it can be seen that the customer's historical credit (Credit_history) and current economic status (Checking_account_status) have the greatest impact on credit delinquency. Banks should pay more attention to these two aspects when evaluating the customer's credit risk level [10].

Conflict of Interest

We have no conflict of interests to disclose and the manuscript has been read and approved by all named authors.

Acknowledgement

This work was supported by the Philosophical and Social Sciences Research Project of Hubei Education Department (19Y049), and the Staring Research Foundation for the Ph.D. of Hubei University of Technology (BSQD 2019054), Hubei Province, China.

References

1. Carroll R, Lawson A, Faes C, Kirby R, Aregay M, et al. (2015) Comparing INLA and OpenBUGS for hierarchical Poisson modeling in disease mapping. Spatial and Spation-temporal Epidemiology 14(15): 45-54.

2. Lyle Konigsberg W, Susan Frankenberg R (2013) Bayes in biological anthropology. American Journal of Physical Anthropology 152(S57): 153-184.

3. Gamerman D, Lopes H (2006) Markov chain Monte Carlo: Stochastic simulation for Bayesian inference, 2nd (Edn.), CRC Press.

4. Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gassian models using integrated nested Laplace approximations. J R Stat Soc Series B 71(2): 319-392.

5. Chen M, Shao Q, lbrahim JR (2000) Monte Carlo methods in Bayesian computation. Springer.

6. Lunn DJ, Andrew A, Best N, Spiegelhalter D (2000) WinBUGS—a Bayes modeling framework: concepts, structure extensibility. Stat Comput 10: 325-337.

7. Kingma DP, Ba J Adam (2015) A Method for Stochastic Optimization. 3rd International Conference for Learning Representations. San Diego.

8. Srivastava AK, Kumar V (2011) Software reliability data analysis with Marshall–Olkin extended Weibull model using MCMC method for non- informative set of priors. Int J Comput Appl 18(4): 31-39.

9. Srivastava AK, Kumar V (2019) Markov Chain Monte Carlo methods for Bayesian inference of the Chen model. Int J Comput Inf Syst 2(2): 7-14.

10. Hang Li (2012) Statistical Learning Methods. Beijing: Tsinghua University Press.