Research on Credit Risk Assessment of Loan Customers

Ming Jin

Shanxi University of Finance and Economics, Taiyuan, 030006, China

1512540534@qq.com

Keywords: Credit evaluation, Chaid, Artificial neural network, Model evaluation

Abstract: With the continuous development and progress of China's market economy, a special economic mode has emerged gradually in the commodity economy, that is, credit economy. Credit has multiple levels and many side meanings. Credit in the economic sense is the activity of borrowing and lending. It is a special form of value movement under the condition of repayment. Under the condition of commodity exchange and currency circulation, The creditor borrows money or credit sales in the form of conditional transfer. The debtor pays the loan or pays the loan at the agreed date, and pays interest. The most important condition in the credit economy is to abide by the credit agreement. Otherwise, it will produce credit risk. In developed countries, risk has become a constituent factor of the business objective. A business item, a loan, often has high income and large risk, low income and low risk. Credit risk refers to the possibility that banks suffer losses because of the uncertainty in credit activities. In fact, it is all risks caused by customers' default. For example, the quality of assets caused by the failure of the borrowers in the asset business to deteriorate, and the depositors' large withdrawals of cash in the debt business and so on. In the field of credit evaluation, more and more mathematical methods and technologies have been used to achieve more precise and interpretable purposes. In this paper, based on the historical information of bank loan customers, the machine learning algorithms are introduced into the modeling of personal credit evaluation, and the data mining decision tree (CHAID algorithm) and neural network are used. According to the characteristics of different customers, we can distinguish the customers who are in good credit condition or those whose credit condition is not good. Then we compare the models obtained and choose an optimal model, so that banks can predict the possibility of future lenders default on loans.

1. Introduction

1.1Deicsion tree model: decision tree is a basic classification method. As the name implies, the decision tree model presents a tree structure to help solve the decision-making problem, showing a process of classifying instances based on features, consisting mainly of nodes and directed edges, where nodes are divided into internal nodes and leaf nodes. The decision is carried out along the directed edges to reach the leaf nodes according to these characteristics. The leaf nodes represent a class. The learning step of the decision tree model is to set up different rules from the root node, then divide the data set according to different rules, and put the already set of data sets on the sub nodes, then make a rule from the sub nodes. According to this rule, the data set of the sub node is divided into the leaf nodes. The leaf nodes correspond to the output results of the whole decision tree model. In the new dataset, we find the corresponding leaf nodes according to the characteristics of the data set and the rules of each sub node in the decision tree. Output result. So the whole decision tree learning is a step to find corresponding leaf nodes according to the rules of each sub node.

1.2CHAID algorithm: one of the algorithms of decision tree model. The predecessor of CHAID (chi-squared automatic interaction detection, chi square automatic interaction detection) is AID. The main feature is multidirectional bifurcation, forward pruning. Its standard is shown in the name. The core is chi square test. The deviation between the actual observation value and the theoretical inference value determines the size of the chi square. In addition, CHAID can only deal with class

  DOI: 10.25236/icemeet.2020.044

type input variables. Therefore, the continuous input variables must be discretized first.

1.3 Neural network: an algorithm mathematical model that mimics the behavior characteristics of animal neural network and distributive parallel information processing. This network relies on the complexity of the system and adjusts the connections among the large number of nodes to achieve the goal of processing information and has the ability of self-learning and adaptation. Neural network is an operation model.Neurons receive input signals from other neurons. These input signals are passed through a weighted connection. The total input value of neurons will be compared with the threshold of the nerve fish, and then processed by the "activation function" to generate the output of neurons, which is more suitable for forecasting.

2. Data Description

2.1 Data Field

The credit evaluation data used in this paper consist of 2464 samples and 6 fields. The method used is decision tree (CHAID algorithm) and neural network. In the research question, the dependent variable is the Credit rating, the independent variables are the customer age (age), the income level (Income), the number of credit cards held (Credit_ cards), the education level (Education), the number of vehicles borrowed (Car_ loans), as shown in Table 1.

Table 1 Data Field Table

Field name	Description
Credit rating	credit rating: 0= is bad, 1= is excellent.
Age	customer age
Income	income level: 1= low, 2=middle, 3= high
Credit_ cards	the number of credit cards: 1= is less than five, 2= five or more.
Education	education level: 1= high school, 2= university
Car_ loans	the number of vehicles on loan: 1= no or one, 2= more than two vehicles.

2.2 Data Description

What I have studied is the question two classification problem. The goal is to distinguish whether the bank loan customers belong to the normal repayment (excellent credit) or the arrears of repayment (bad credit). There are 5 explanatory variables. I will use spss.modeler18 to set the measurement attribute of the field "Credit rating" as the mark, the role setting as the goal, and the other variables to measure the attribute as continuous, and the role as input. Auditing data shows that the valid values are 2464, and there is no invalid value, and the dependent variable is only classified variable, and the value is only two.

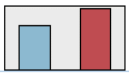
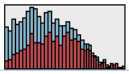
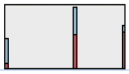
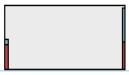
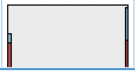
字段	样本图形	测量	最小值	最大值	平均值	标准差	偏度	唯一	有效
Credit rating		标记	0.000	1.000	--	--	--	2	2464
Age		连续	20.003	63.350	33.816	8.539	0.488	--	2464
Income		连续	1.000	3.000	2.091	0.729	-0.142	--	2464
Credit_cards		连续	1.000	2.000	1.676	0.468	-0.753	--	2464
Education		连续	1.000	2.000	1.501	0.500	-0.003	--	2464
Car_loans		连续	1.000	2.000	1.638	0.481	-0.573	--	2464

Fig.1 Data Audit Chart

From the graph, we can see that the average age of each variable is about 34 years old, the standard deviation is 8.539, the average income is 2.091, the standard deviation is 0.729, showing the left deviation; the average number of credit cards is 1.676, the standard deviation is 0.468, showing a left deviation, indicating that most customers have more than 5 credit cards; the average educational level is 1.501, the standard deviation is 0.5, showing the left skew. It shows that most of the customers are educated in universities. The average number of loans is 1.638, the standard deviation is 0.481, showing a left skew, which indicates that most customers have more than two vehicles.

3. The Establishment of the Model.

Before modeling, I categorized the data and divided them into training set and test set according to 7:3 ratio to generate models.

3.1 Decision Tree (Chaid) Model

(1)Importance of variables

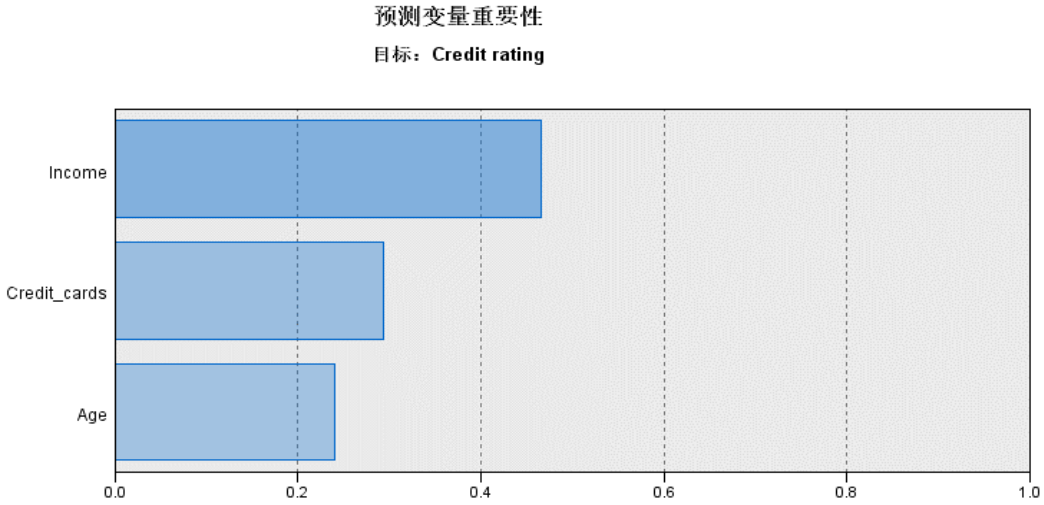


Fig.2 Importance Map of Prediction Variables

According to the importance map of prediction variables, we can see that income, credit card number and age are selected as the important variables of decision tree, and the highest degree of income is 0.46. The number of credit cards is 0.28 and the importance of the latter is 0.23. Therefore, income is the first important basis for division, that is, the first node. The number of credit cards held is second important divisions, and age is the third important basis.

(2) Decision tree model

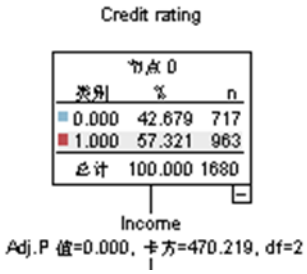


Fig.3 Branch Diagram of Node 0

Node 0 provides all records in the training set, and 42.679% of the customers in the training group are relatively poor in credit rating.

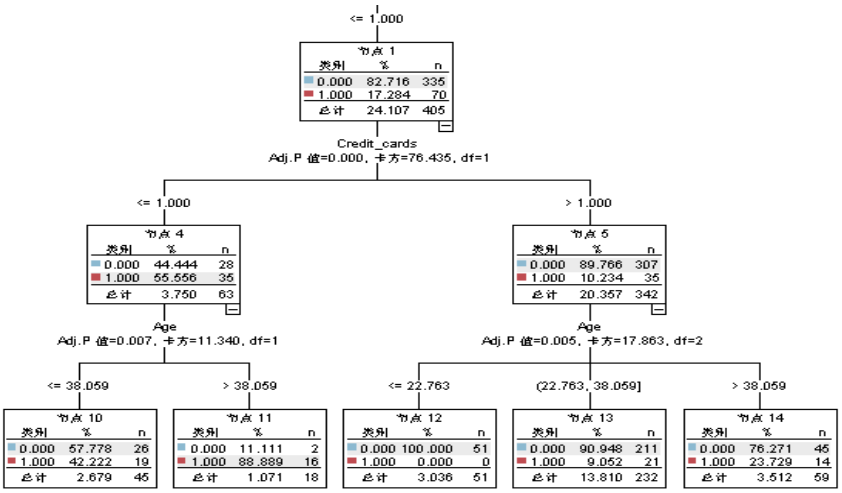


Fig.4 Branch Diagram of Node 1

From the above variables, we can see that the decision tree is based on the Revenue Branch. Node 1 is a low income customer, of which 17.284% of credit is good, and 82.715% of credit is bad. According to the number of credit cards we hold, customers are divided into two categories. The number of customers who hold fewer than 5 credit cards is 55.556%, while that of bad credit accounts for 44.444%.The number of customers who hold more than 5 credit cards account for only 10% of the customers who have good credit, while the proportion of credit bad accounts for nearly 90%. Next, we further divide them according to the age of customers. Although the proportion of customers with bad credit is higher than that of good credit at each age group, we can still observe that the proportion of credit good customers increases with age.

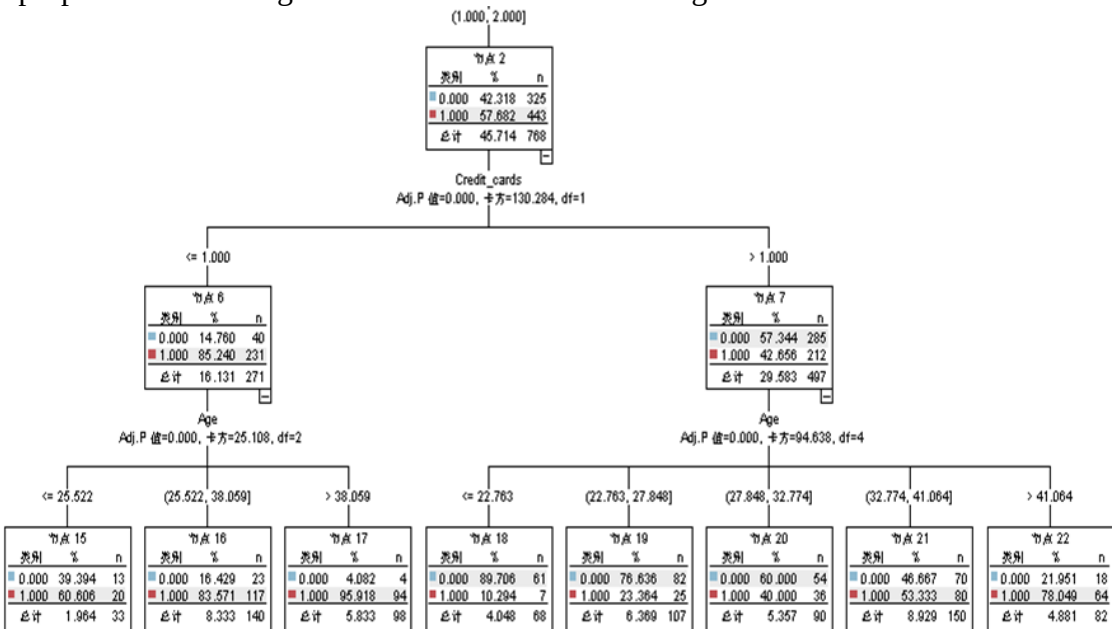


Fig.5 Branch Diagram of Node 2

Node 2 is a middle-income customer with a good credit rating of 57.682%, which is higher than that of the low income group. It also divides the customers into two categories according to the number of credit cards held. The number of customers who hold fewer than 5 cards is 85.24%, while the proportion of credit cards is 14.76%.Credit only accounts for 42.656%, while bad credit accounts for 57.344%. Next, according to the age of customers, we can further see that with the growth of age, the proportion of customers with good credit has also increased.

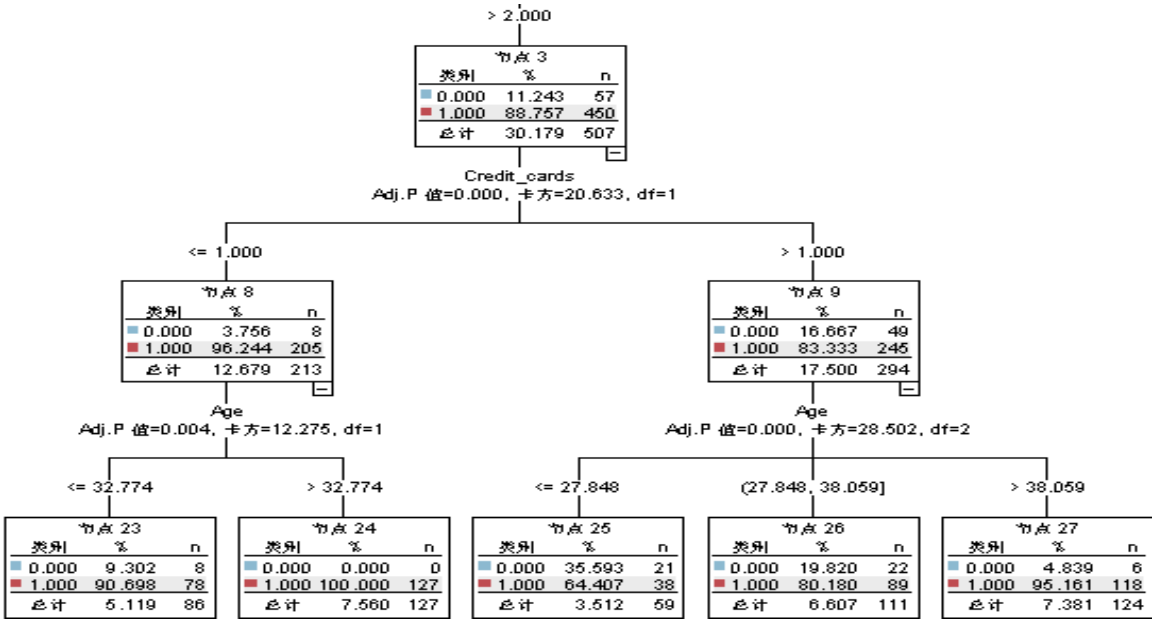


Fig.6 Branch Diagram of Node 3

Node 3 is a high-income customer with a good credit rating of nearly 88.757%. It is the highest proportion of all income types. It also divides customers into two categories according to the number of credit cards held. The number of customers who hold fewer than 5 cards is 96.244%, while those with bad credit account for only 3.756%.Credit accounted for 83.333%, while credit accounted for 16.667%. Next, according to the age of customers, we could further see that the proportion of credit good customers increased with age.

Therefore, we can observe that through the decision tree model, we can see that the higher the income is, the less the number of credit cards we hold, the better the credit condition of the older customers is.

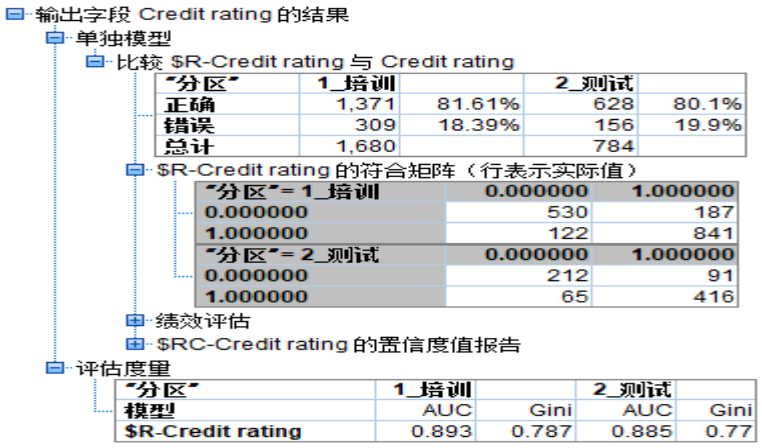


Fig.7 Analysis Results of Chaid

From the above, we can conclude that the correct rate of the test set is 80.1%, the error rate is 19.9%, and the AUC is 0.885.

The confusion matrix of classification results of test sets can be obtained from the analysis results.

Table 2 Confusion Matrix Table for Classification Results of Test Sets

Real situation	Prediction results	
	Customers with good credit standing	Customers with poor credit standing
Customers with good credit standing	416	65
Customers with poor credit standing	91	212

Precision rate $P = \frac{TP}{TP+FP} = \frac{416}{416+91} = 82.05\%$

Recall rate R $= \frac{TP}{TP+FN} = \frac{416}{416+65} = 86.49\%$

Overall accuracy rate $F1$ by

$$F1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{\text{Total sample number } + \text{tp } * \text{ FN}} = \frac{2 \times 416}{784 + 416 - 65} = 73.30\%$$

3.2 Neural Network Model

First of all, the importance of forecasting variables is still being analyzed.

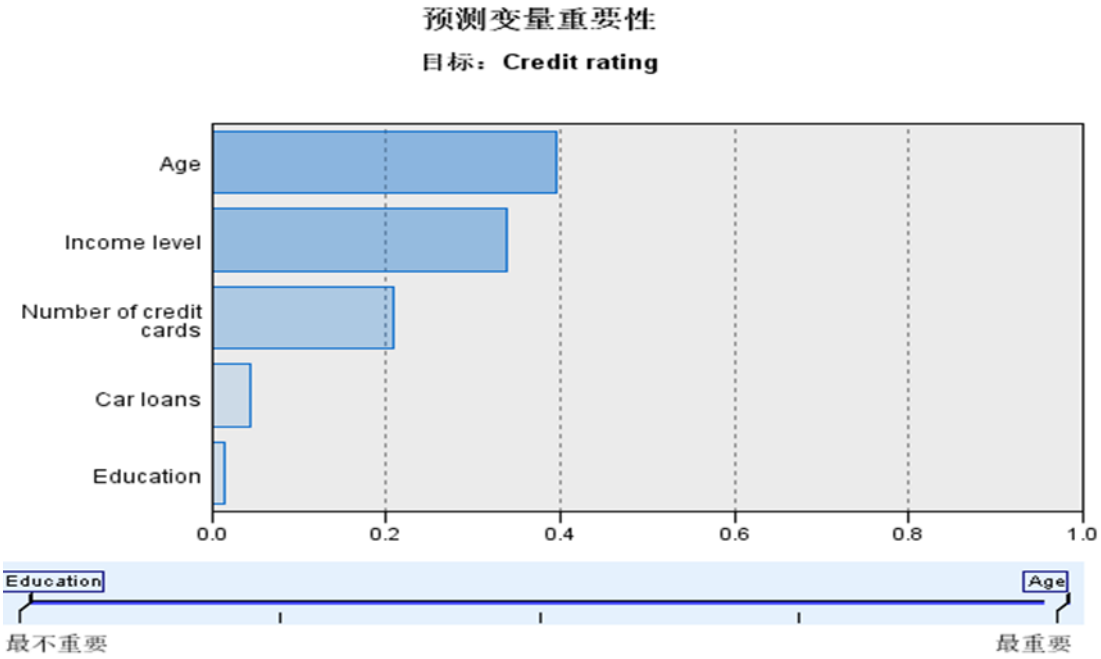


Fig.8 Importance Map of Prediction Variables

The most important variables in these five variables are age, the next is the income level, then the number of credit cards held, then the number of loans, and finally the level of education. We choose three variables which are age, income level and credit card number.

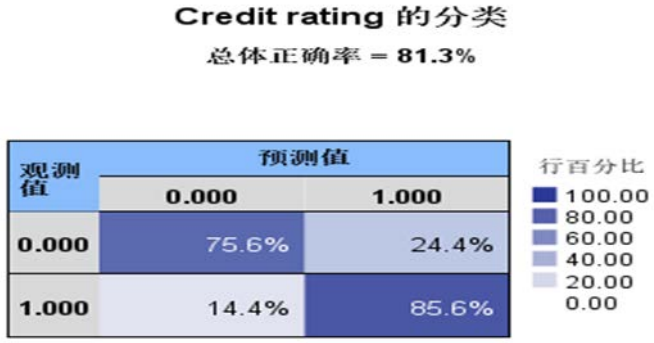


Fig.9 Diagram of Classification Results

The “0” stands for bad credit. The “1” stands for good credit. The observed value is 0. The forecast value is 0, accounting for 75.6%. The forecast value is 1, accounting for 24.4%. The observed value is 1, and the predicted value is 1, 85.6%. The prediction value is 0, and the overall accuracy rate is 0, which shows that the model is very good.

The neural network model established is shown in the following figure.

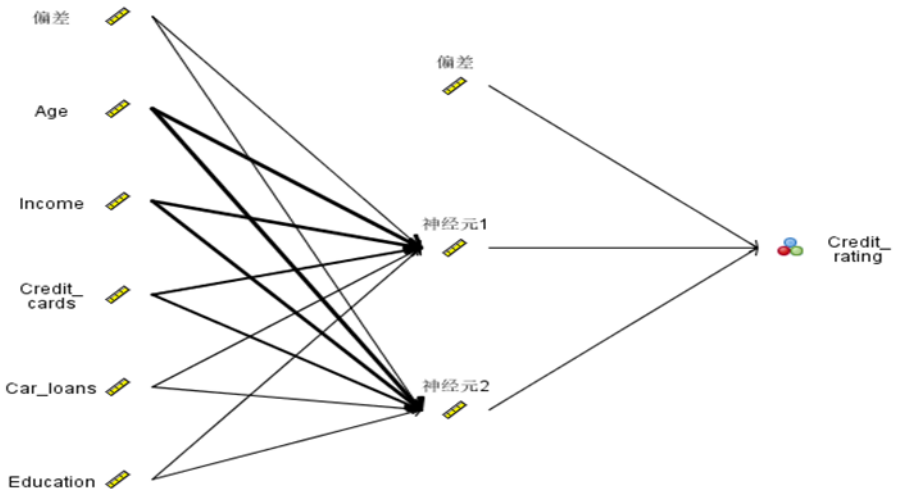


Fig.10 Neural Network Diagram

This is a single hidden layer feedforward neural network. There are 5 neurons in the input layer, 2 neurons in the hidden layer, and 1 neurons in the output layer. The input layer neurons accept external input, and the hidden layer and output layer neurons process the signal. The final result is output from the output layer neurons. The thickness of the line indicates the weight of the input layer. The age of the graph shows the maximum number of credit cards. Therefore, it can be concluded that its weight is large, so it has a larger proportion in deciding the final output, indicating that it plays an important role in deciding whether the customer credit is good or not.

According to this model, the analysis results are as follows.

输出字段 Credit rating 的结果

单独模型

比较 $N-Credit rating 与 Credit rating

"分区"	1_培训		2_测试	
正确	1,366	81.31%	635	80.99%
错误	314	18.69%	149	19.01%
总计	1,680		784	

$N-Credit rating 的符合矩阵（行表示实际值）

"分区"= 1_培训	0.000000	1.000000
0.000000	542	175
1.000000	139	824
"分区"= 2_测试	0.000000	1.000000
0.000000	220	83
1.000000	66	415

绩效评估

$NC-Credit rating 的置信度值报告

评估度量

"分区"	1_培训		2_测试	
模型	AUC	Gini	AUC	Gini
$N-Credit rating	0.892	0.784	0.887	0.775

Fig.11 Results of Neural Network Analysis

The correct rate of the test set is 80.99%, the error rate is 19.01%, and the AUC is 0.887.

And the confusion matrix can be obtained.

Table 3 Confusion Matrix Table for Classification Results of Test Sets

Real situation	Prediction results	
	Customers with good credit standing	Customers with poor credit standing
Customers with good credit standing	415	66
Customers with poor credit standing	83	220

Precision rate$P = \frac{TP}{TP+FP} = \frac{415}{415+83} = 83.33\%$

Recall rate$R = \frac{TP}{TP+FN} = \frac{415}{415+66} = 86.29\%$

Overall accuracy rate$F1 = \frac{2 \times P \times R}{P+R} = \frac{2 \times TP}{\text{Total } sample\ number\ + tp\ * FN} = \frac{2 \times 415}{784+415-66} = 73.26\%$

4. Model Assessment

4.1 Recall and Precision:

For the two classification problem, the results obtained from the data analysis of the cases are divided into the following categories: the true rate (TP), the true category of the category is a positive case, and the category obtained by prediction is a positive case; the false positive case (FP) refers to the true category of the category is negative.The categories obtained by prediction are positive cases; false negative cases (FN) refer to the true category of the category as a positive case; the category obtained by prediction is negative; true negative case (TN) refers to the true category of the category is negative, and the category obtained through prediction is a positive case, thus constructing the confusion matrix.

Recall ratio and precision rate are a pair of contradictory measures. Generally speaking, when the recall rate is high, the precision rate is often low. When the precision rate is high, the recall rate is always low. So we need to find a balance point (BEP), which is the value of recall ratio = precision rate. But BEP is too simplistic, and more commonly used is F1 metric: F1= (2 x p x r) / (p+r) = (2 x TP) / (sample total +TP-FN), in general,The greater the value of F1, the better the classification.

	Table 4 Comparison of two model parameters	Decision tree model
Prexison ratio P	82.05%	83.33%
Recall ratio R	86.49%	86.29%
Overall accuracy F	73.30%	73.26%
AUC	0.885	0.887

From the above table, we can see from the two aspects of precision, recall and F1, the decision tree model is better.

But in actual classification, it is very one-sided to rely solely on recall rate, precision rate and overall accuracy rate. The requirement of model classification assessment will be different because of different scenes. At this time, the ROC curve can be introduced to represent the classification effect of the model synthetically. The ordinate of the.Roc curve is TPR (real case rate) and the abscissa is fpr (false positive rate).The essence of ROC curve is to show the trend of real case rate with the change of false positive rate.The ROC curve is made up of a series of (FPR, TPR) points. When a general classifier predicts each sample, it can output the probability value of the sample belonging to the positive class (that is, 1). The range of the probability value is between (0-1), and the general threshold is 0.5. That is, the probability value greater than or equal to 0.5 is considered to be a positive class, otherwise it is a negative class.Then the predicted probabilities are used as thresholds in turn. Each time the threshold is obtained, the number of samples is positive and negative, and then a set of (FPR, TPR) values is generated, so that one point on the ROC curve can be obtained. Finally, the ROC curve is generated by connecting all the points. Obviously, the more times the threshold is set, the more (FPR, TPR) values will be generated.The ROC curve is also smoother. The characteristic of.Roc is for different test samples. The ROC curve does not change with the positive and negative samples in the dataset. The data in the test set are prone to data imbalance. The number of positive samples is much higher than the number of negative samples or the number of negative samples is much higher than the number of positive samples.The number of positive samples and negative samples may vary with time. Each model can output a ROC curve when comparing the classification results of different models for the same data.

But using naked eye to watch the ROC curve to compare the advantages and disadvantages of the model can only get ambiguous results, so it will be better if we can quantify the advantages and disadvantages of different classifiers with a single value, and then AUC appears. AUC refers to the

area surrounded by the ROC curve and the X axis and the Y axis. The larger the area and the higher the AUC value, the better the classification effect of the model.

AUC is not only a quantitative indicator, but also has a good measurement of the classification effect of the model, and it can also play a good role in data imbalance. Therefore, based on AUC, the AUC value of the neural network is larger than that of the decision tree. Finally, the neural network model is selected to evaluate the credit of the loan customer, so as to decide whether to provide loans for the customers.

5. Summary

By establishing the credit evaluation model and selecting the optimal model, it provides the security for the bank to issue the loan to the customer, which is helpful for the bank to avoid the risk brought by the personal credit, thus reducing the loss to the bank. Personal credit business is mainly faced with natural persons, which is characterized by the small size of single business funds and complex business and large number. Credit risk is the risk in the credit activities of economic subjects, which exists more in bank credit. For banks, loans are the largest and most obvious source of credit, so it is necessary to evaluate and predict the credit excellence of customers by constructing models.

References

[1] Zhou Zongfang, Zhang Ying, Chen Lin. A Study on the Evolution Mechanism and Evaluation Method of Credit Risk in Emerging Technology Enterprises. Science Press, 2010.

[2] Handsome. Development of Theory and Methods of Personal Credit Risk Assessment.2015.

[3] Xiao Li. A Study on Personal Credit Assessment of Bank. Nanjing University of Information Engineering, 2008.