Filomat 32:5 (2018), 1513–1521
https://doi.org/10.2298/FIL1805513Z

Published by Faculty of Sciences and Mathematics,
University of Niš, Serbia
Available at: http://www.pmf.ni.ac.rs/filomat

Decision tree for credit scoring and discovery of significant features: an empirical analysis based on Chinese microfinance for farmers

Yajing Zhang[a], Guotai Chi[a], Zhipeng Zhang[a]

[a]*Faculty of Management and Economics, Dalian University of Technology, DaLian, Liaoning, 116024, P. R. China*

Abstract. For the tens of thousands of farmers' loan financing, it's imperative to find which features are the key indicators affecting the credit scoring of rural households. In this paper, C5.0, CHAID and C&RT three models are used to screen the key indicators affecting farmers' credit scoring, and 2044 farmers' microfinance data from 28 provinces in China are applied in the empirical study. The empirical results show the classification accuracy of C5.0 is better than CHAID and C&RT in both the training set and test set, thus finally use the feature subset selected by C5.0. Six key features screened from 44 attributes by C5.0, which have significant influence on credit scoring of farmers, namely, education level, net income each year/per capita GDP, education cost of children each year, Residence type, residential year, relationship with cosigners.

1. Introduction

The agricultural population accounts for 46.1% of the world's total population. Furthermore, in China, India and other developing countries, the proportion has reached 43.9% and 67.3%. For the tens of thousands of farmers' loan financing, the customer's credit rating directly influences bank's lending decision.

As the credit industry has been growing rapidly, credit scoring models have been widely used by the financial industry during this time to improve default prediction accuracy. However, a large amount of redundant information and features are involved in the credit dataset, which leads to lower accuracy and higher complexity of the credit scoring model. So, effective feature selection methods are required for credit dataset with huge number of features.

In this paper, C5.0 and CHAID decision tree models are utilized to do feature selection. The experimental result shows that C5.0 has a superior performance in improving classification accuracy compared with

2010 *Mathematics Subject Classification*. Primary 62P25.

Keywords. Decision trees, C5.0, CHAID, Data mining, Classification, Credit scoring.

Received: 2 September 2017; Accepted: 12 November 2017

Communicated by Hari M. Srivastava

Research supported by Key Program of National Natural Science Foundation of China (No. 71731003; 71431002), National Natural Science Foundation of China (No.71471027; 71171031), the National Social Science Foundation of China (No. 16BTJ017), Young Scientists Fund of the National Natural Science Foundation of China (No. 71503199; 71601041), Liaoning Province Social Science Planning Fund Project (L16BJY016), Liaoning Economic and Social Development Key Project (2015lslktzdian-05), Credit Risks Rating System and Loan Pricing Project of Small Enterprises for Bank of Dalian (No. 2012-01), and Credit Risks Evaluation and Loan Pricing For Microfinance Funded for the Head Office of Post Savings Bank of China (No. 2009-07).

Corresponding Author: Guotai Chi.

Yajing Zhang and Guotai Chi contributed equally to the paper as first authors.

Email addresses: yajing1990.08@163.com (Yajing Zhang), chigt@dlut.edu.cn (Guotai Chi), zhangzp517@126.com (Zhipeng Zhang)

CHAID. Therefore, the feature subset screened by C5.0 is more significant for Chinese farmers' credit scoring.

The rest of this paper is organized as follows. Section 2 gives a literature review about feature selection of credit risk. Section 3 introduces the process of feature selection with C5.0 and CHAID methods. Section 4 describes the data source. Section 5 presents the empirical results. Section 6 analyses the influence of the key features on the Chinese farmers' credit risk. Section 7 summarizes the contribution of this paper and specifies future work towards further improvements.

2. Literature review

Feature selection has attracted lots of research interests in the literature. Recent studies have shown that traditional statistical methods and Artificial Intelligence (AI) methods are usually applied to feature selection which can improve the accuracy of credit risk recognition.

About traditional statistical techniques, the existing researches explore factors which have impacted on credit risk of customers mainly through statistical methods such as Multiple Discriminant Analysis, Multiple Logic Regression, Markov Chain. Pinches [1] and McAdams [2] proposed Multiple Discriminant Analysis to study the influencing factors of the credit rating. Pishbahar [3] analyzed the data from 779 individual farmers by using Nested Logit Model (NLM), and revealed the key factors impacting on repayment. Karan et al. [4] screened credit evaluation indicators by building a logical regression model. Afolabi [5] analyzed some socioeconomic characteristics of 286 small scale farmers in Nigeria by using method of quantitative analysis. Karminsk [6], Geng et al. [7], Figby et al. [8] respectively applied Ordered Probit Regression, oneway ANOVA, and Survival Duration models to explore the indicators that influence credit risk. Shi [9,10] analyzed customers' credit qualification by the Fuzzy rough set method and F test method. Zhang [11] applied genetic algorithm to analyze the credit rating of customers. Petropoulos [12] proposed a hidden Markov model for credit rating predictions and yield significantly more reliable prediction. Hwang [13] made a comparison among traditional statistical methods and the results showed that the most successful methods are ordered logit regression and ordered probit model. Shi [14] proposed a novel technique to distinguish the customer's credit level by using fuzzy cluster analysis. Shi [15] combined logistic regression and correlation analysis to extract features. R cluster analysis and coefficient of variation were also applied to selection features [16].

About Artificial Intelligence (AI) techniques, recently researchers have proposed the hybrid data mining approach in the design of an effective credit scoring model. Neural network, support vector machine(SVM), genetic algorithm and other methods are investigated in feature selection for credit scoring. Akkoç [17] proposed a three stage hybrid Adaptive Neuro-Fuzzy Inference System credit scoring model. Huang [18] applied a neural network method to credit risk recognition, results showed that the total assets, total liabilities, operating profit margin are the key features for the US samples' credit risk. Kim [19], Cao [20] investigated SVM in credit risk evaluation, and an analysis of features shows that the generalization performance of SVM can be further improved by performing feature selection. Hájek [21] and Shin [22] used genetic algorithms to select input variables. Hajek [23] proved that wrappers performed better than filters in improving accuracy for both US and European datasets.

The above studies have made great strides in illustrating the depth and breadth of research on credit scoring issues. However, research on credit scoring issue is still insufficient. First of all, most of the credit risk features focus on financial indexes, or personal information of customers which ignore the macro economic factors. Secondly, most credit rating models only focus on the accuracy of classification, but they cannot be used to select the key factors which impact on the repayment willingness clients.

To fill in the above gaps, our study advances in three aspects. First, this paper selects 44 features which include 5 criterion layers, i.e. ”Basic information”, ”Repayment ability”, ”Repayment willingness”, “Guarantee and joint guarantee”, “Macro environment” to build the credit scoring model. The factors discussed in this paper are more extensive. Second, in this paper, the C5.0, CHAID and C&RT decision tree methods are used to construct the credit score model, and the key features influencing repayment willingness are determined by the decision tree with high precision. Then, we discussed how these key features influence the credit qualification of farmers in China, which can help the financing institutions to make or adjust credit policies.

3. Method

3.1. C5.0 decision tree

C5.0 is another new decision tree algorithm developed based on C4.5 by Quinlan [24]. The idea of construction of a decision tree in C5.0 is similar to C4.5. We introduce the construction of a decision tree based on C4.5 [25].

(1) Information Entropy

Where p_i is the proportion of S belonging to class i (default or not default).

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \tag{1}$$

If half of the customers are default, another half are non-default, the Entropy (S) would reach a maximum value. If Entropy (S)=0, means that all the customers are good or bad without uncertainty. Thus, the greater the difference among pi is, the smaller the information entropy is.

(2) Information Gain

Where *Values(a)* is the set of all possible values for attribute a, and S_v is the subset of S for which attribute a has value v.

$$Gain(S, a) = Entropy(S) - \sum_{v \in Values(a)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{2}$$

The greater the Gain value is, the stronger ability of an attribute to identify credit risk.

(3) Information Gain Ratio

$$GainRatio(S, a) = Gain(S, a)/Entropy(a) \tag{3}$$

The problem of information ratio is that if an attribute has many values, it would be biased towards tests. However, information gain ration Eq.(3) overcomes the bias of Information gain.

(4) C5.0 decision tree algorithm

C5.0 decision tree algorithm [26] is shown in Figure 1.

3.2. CHAID decision tree

The purpose of this section is to investigate in the utility of the chi-squared automatic interaction detection (CHAID) algorithm to identify the key features for credit scoring. According to Kass (1980), the CHAID algorithm operates using a series of merging, splitting, and stopping steps based on user-specified criteria as follows [27].

(1) *Merging step.*

χ^2-test for independence is performed for each pair of categories of the feature variable in relation to the binary target variable using the χ^2 distribution ($df = 1$) with significance (α_{merge}) set at 0.05. For non-significant outcomes, those paired categories are merged. For non-significant tests identified by $\alpha_{merge} > 0.05$ those paired categories are merged into a single category. For tests reaching significance identified by $\alpha_{merge} \leq 0.05$, the pairs are not merged.

(2) *Splitting step.*

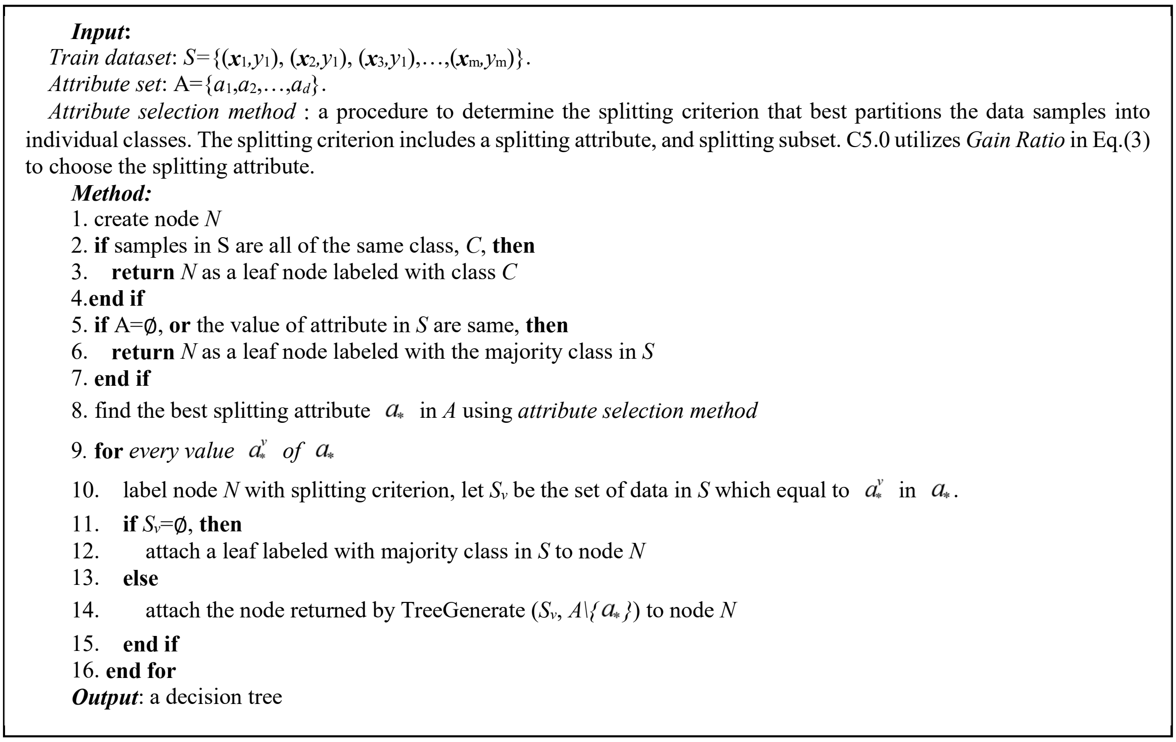
This step selects which predictor is to be used to "best" split the node using the following algorithm.

χ^2-test for independence is performed for each feature variable. The feature variable with the smallest adjusted value (i.e., most statistically significant) is split if the value less than 0.05 ($\alpha_{split} = 0.05$), otherwise the node is not split and is then considered a terminal node.

(3) *Stopping step.*

If the current tree reached the maximum tree depth level, or the size of a node is less than the user-specified minimum node size, the tree process stops.

The CHAID algorithm will continue until all the stopping rules are met.



```
Input:
Train dataset: $S=\{(\boldsymbol{x}_1,y_1), (\boldsymbol{x}_2,y_1), (\boldsymbol{x}_3,y_1),\ldots,(\boldsymbol{x}_m,y_m)\}$.
Attribute set: $A=\{a_1,a_2,\ldots,a_d\}$.
Attribute selection method : a procedure to determine the splitting criterion that best partitions the data samples into individual classes. The splitting criterion includes a splitting attribute, and splitting subset. C5.0 utilizes Gain Ratio in Eq.(3) to choose the splitting attribute.
Method:
create node $N$
if samples in S are all of the same class, $C$, then
    return $N$ as a leaf node labeled with class $C$
end if
if $A=\emptyset$, or the value of attribute in $S$ are same, then
    return $N$ as a leaf node labeled with the majority class in $S$
end if
find the best splitting attribute $a_*$ in $A$ using attribute selection method
for every value $a_*^v$ of $a_*$
    label node $N$ with splitting criterion, let $S_v$ be the set of data in $S$ which equal to $a_*^v$ in $a_*$.
    if $S_v=\emptyset$, then
        attach a leaf labeled with majority class in $S$ to node $N$
    else
        attach the node returned by TreeGenerate ($S_v$, $A\backslash\{a_*\}$) to node $N$
    end if
end for
Output: a decision tree
```

Figure 1: A pseudo code of C5.0 decision tree algorithm

3.3. Classification And Regression Trees(C&RT)

C&RT is the commonly used decision tree in data mining which was developed by Breiman et al. (1984) [28]. C&RT partitions the data into two subsets so that the clients within each subset are more homogeneous than in the previous subset.

In this paper, the Gini index is the splitting criterion

(1) *Gini index*.

The Gini index $G(S)$ at a node S in a C&RT tree, is defined as:

$$G(S) = 1 - \sum_{i=1}^{c} p_i^2 \tag{4}$$

Where p_i is the proportion of S belonging to class i (*default or not default*). Therefore, when the clients in a node have no difference in default status, the Gini index takes its maximum value. When all clients are in a node belong to the same class, the Gini equals to 0.

(2) Gini criterion function $\Delta G\,(S)$.

Where N_r is the number of clients in S sent to the right child node, and N_l is the number of clients in S sent to the left child node, and N is the number of clients in S node, and $G(S_r)$ is the gini index of left subtree, $G(S_r)$is the gini of right subtree.

$$\Delta G(S) = G(S) - \frac{N_r}{N} G(S_r) - \frac{N_l}{N} G(S_l) \tag{5}$$

The Eq.(5) stands for the decrease in impurity relative to the impurity of the node being split.

(3) C&RT algorithm

Find each feature's best split. Identify the best split t point at S node which maximizes the value $\Delta G\,(S)$ in Eq.(5).

Find the best split for the node. Select the feature whose best split provides the most significant decrease in impurity for the node, then utilize that feature's best split as the best overall split for the node.

Check stopping rules, and recurse. Similarly with the stopping step of CHAID, check the maximum tree depth level, and the size of each parent node and each child node. If no stopping rules are triggered, implement the algorithm again to each child node.

4. Data source

This paper selects 44 features of loans for farmers from a Chinese national commercial bank, which includes 5 criterion layers, i.e. "Basic information", "Repayment ability", "Repayment willingness", "Guarantee and joint guarantee", "Macro environment", as shown in Table 1.

The dataset consists of 2044 customers, 1816 customers belong to good credit and 228 belong to bad credit. For each customer, contains 44 attributes (numeric and nominal), and for each application, no missing value.

All of the data is divided into a training data set and testing data set, including 1406 training samples (1239 good ones, 167 bad ones) and 638 testing samples (577 good ones, 61 bad ones), shown in Table 2.

Table 1: Index Set of Farmers Credit Scoring

Criterion layers	Indicator layers	Index type	Indicator layers	Index type
Basic information	X_1 Age	numeric	X_6 Number of laborers	numeric
	X_2 Education level	nominal	. . .	. . .
	X_3 Marital status	nominal	X_9 Loan purpose	nominal
	X_4 Gender	nominal	X_{10} Value of house owing	numeric
	X_5Number of family members	numeric	X_{11} House value	numeric
Repayment abilit	X_{12} Job skills of borrower	nominal	X_{18} Expenses/incomes	numeric
	X_{13} Net income of household business	numeric	X_{19}Total property	numeric
	X_{14} Net income each year/per capita GDP	numeric	X_{20} Net agricultural incomes	numeric
	X_{15} Net income of borrower	numeric	. . .	. . .
	X_{16} Expense of family's daily life	numeric	X_{23}Non-agricultural incomes/total incomes	numeric
	X_{17} Total expenses	numeric	X_{24} Education cost of children each year	numeric
Repayment intention	X_{25} Residence type	nominal	X_{30}Have private loan or not	nominal
	X_{26} Residential year	numeric	X_{31} Record of overdue loans	nominal
	X_{27} Outstanding loan balance in bank	numeric	$X_3$2 Number of loan applications	nominal
	X_{28} Have outstanding loan in bank or not	nominal	X_{33}Loan records of borrower	nominal
	X_{29} Bank deposit	numeric	X_{34} Social reputation status	nominal
Guarantee joint guarantee	X_{35} Have guarantee or not	nominal	X_{37} Joint guarantee state	nominal
	X_{36} Guarantor's monthly income	nominal	X_{38} relationship with cosigners	nominal
Macro environment	X_{39} Net income per capita for a rural household	numeric	X_{42} CPI	numeric
	X_{40} Per capital agricultural output value	numeric	X_{43} Residents' deposit balance	numeric
	X_{41} Increasing rate of regional GDP	numeric	X_{44} Engel's coefficient	numeric

Table 2: The number of customers in total sample, training sample and testing sample

	Number of customers	Number of good customers/Frequency	Number of bad customers/Frequency	Proportion of good and bad customers
Total dataset	2044	1816(88.85%)	228(11.15%)	7.96:1
Training dataset	1406	1239(88.12%)	167(11.88%)	7.42:1
Testing dataset	638	577(90.44%)	61(9.56%)	9.46:1

5. Emprical analysis

(1) *Credit default prediction accuracy of C5.0, CHAID, and C&RT*

We use SPSS Clementine to construct C5.0, CHAID, and C&RT model based on 1406 training samples. 44 variables in table 1 are the independent variables, and default state is target variable. If the customer is default, the default state is 1. If the customer is non-default, the default state is 0. During modeling the decision trees, all parameters relative to C5.0, CHAID and C&RT model are set to default. 638 testing samples are used to verify the models.

Classification error rates with C5.0, CHAID and C&RT models are presented in Table 4.

Table 3: Classification confusion matrices of C5.0, CHAID and C&RT

C5.0	Training set	0(classified as good)	1(classified as bad)
	0 (true good)	1156	83
	1 (true bad)	122	45
	Test set	0(classified as good)	1(classified as bad)
	0 (true good)	527	50
	1 (true bad)	51	10
CHAID	Training set	0(classified as good)	1(classified as bad)
	0 (true good)	1065	174
	1 (true bad)	90	77
	Test set	0(classified as good)	1(classified as bad)
	0 (true good)	474	103
	1 (true bad)	47	14
C&RT	Training set	0(classified as good)	1(classified as bad)
	0 (true good)	1091	148
	1 (true bad)	111	56
	Test set	0(classified as good)	1(classified as bad)
	0 (true good)	495	82
	1 (true bad)	44	17

The Type I error (false-positive) means the model assigns a bad credit quality when, in fact, the quality is good. Potential losses resulting from this Type I error refer mainly to opportunity costs and lost potential profits. The Type II error (false-negative) corresponds to the assignment of good quality to a customer who is default. The cost to the investor can be the loss of principal and interest. For banks, the Type II error is more important than Type I error. Therefore, we set that the ratio of misclassification costs, associated with Type I and Type II, is 1:2.

According to the confusion matrix of C5.0 based on training set in table 3, calculate total error, Type I error, and Type II error as follow: total error rate = (122+83)/1406=14.58%, Type I error rate =83/1406 =5.90%, and Type II error rate =122/1406=8.68%, Cost of misclassification=C_{01}×Type I error rate + C_{10}×Type II error rate = 1×5.90%+2×8.68% =23.26%. In the same way, other models' error rates can be calculated, the results are shown in table 4. As shown in table 4, the total error rate and the type I error rate of C5.0 are the smallest in both the training sample and the test sample compared to CHAID, and C&RT. However, for the type II error rate, CHAID and C&RT have better performance than C5.0. Then, compared the costs of misclassification among three decision models, C5.0 still has the best performance. Overall, C5.0 is superior to CHAID and C&RT. Therefore, empirical results of C5.0 are finally being selected.

Table 4: Classification error rates of C5.0 and CHAID methods

Method	C5.0	CHAID	C&RT
Training set			
Total error	14.58%	18.78%	18.42%
Type I error	5.90%	12.38%	10.53%
Type II error	8.68%	6.40%	7.89%
Cost of misclassifica-tion	23.26%	25.18%	26.31%
Test set			
Total error	15.83%	23.51%	19.75%
Type I error	7.84%	16.14%	12.85%
Type II error	7.99%	7.37%	6.90%
Cost of misclassification	23.82%	30.88%	26.65%

Note: set the cost of Type I error: the cost of Type II error=1:2, that is, C_{01}: C_{10}=1:2, thus Cost of misclassification= $C_{01} \times$ TypeI $+C_{10} \times$ TypeII =1 × Type I error rate+2 × Type II error rate.

(2) *Feature subsets of C5.0, CHAID and C&RT*

C5.0 decision tree screens six features, CHAID screens nine features, and C&RT selects five features, shown in table 5.

Table 5: Feature subset with different methods

Method	Feature subset
C5.0	X_2 Education level X_{14} Net income each year/per capita GDP (%) X_{24} Education cost of children each year (¥) X_{25} Residence type X_{26} Residential year X_{38} Relationship with cosigners
CHAID	X_8 Supporting population X_{11} House value (¥) X_{15} Net income of borrower (¥) X_{19} Total property (¥) X_{25} Residence type X_{29} Bank deposit (¥) X_{35} Have guarantee or not X_{40} Per capita agricultural output value (¥) X_{44} Engel's coefficient
C&RT	X_2 Education level X_{11} House value (¥) X_{14} Net income each year/per capita GDP (%) X_{36} Guarantor's monthly income (¥) X_{39} Net income per capita for a rural household (¥)

(3) *Decision tree diagram*

Because the overall performance of C5.0 is better than CHAID and C&RT, for the sake of brevity herein, only gives a decision tree diagram of C5.0. Through 1406 training samples, the decision tree generated by C5.0 is illustrated in figure 2.

6. Discussion

(1) About *residence type*, credit quality of the farmers with owner occupied housing, mortgage loan housing, shared-ownership housing, relatives housing, is better than the farmers with rental housing and other situations. The reasons may come from two aspects. First, residence type can reflect the economic situation of the borrower, if an obligor has its own house, etc., the borrower has a good economic situation. In China, if it's possible, people will first improve their housing conditions. Second, the residence Type Indirectly reflects the obligor's mobility. Once the borrower leaves the location of the lending institution, it's difficult to collect the corresponding loan. For obligors with rental housing or other situations, who live in a certain region for a short time, there is an objective possibility of default because of the mobility. For obligors with stable living environment, they have more social relationship in the local, and will pay greater attention to maintain their credit image, therefore, their repayment willingness will be greater.

(2) *Education type* affects the credit of obligors who have own housing, etc. Strangely, customers with "Bachelor degree or above" have higher default rate than customers with low-levels of education. In the training samples, for 14 customers with "Bachelor degree or above", 6 customers are default, default rate reaches 48.96%(6/14). Ignoring the repayment willingness, the customers with high-level education know more investment channel and invest more money in the project with high risk and high return, which also lead to high default possible. Relatively, obligors with low-level education are more inclined to invest in low-risk projects.

(3) *Education cost of children each year* affects the customer who satisfies the following conditions that "residence Type Is own housing etc., education level is junior college degree and others". Once the children's education costs more than 2200 RMB. The default possibility of Borrowers will increase.

This may be due to the fact that education expenditure is one of the major expenditures of the family, the higher the cost of education for the children, the greater the pressure on the family, and more likely customers are to default.

(4) *Residential year* affects the credit quality of the customer who satisfies the following conditions that "residence Type Is own housing etc., education level is high school or below". Farmers whose residential year is less than one year have a high default rate. It's similar to "residence type". Customers with longer residential years will pay more attention to maintain their credit image. After all, the trust of the financial institutions and the neighborhood will significantly improve the quality of the obligors in the local life.

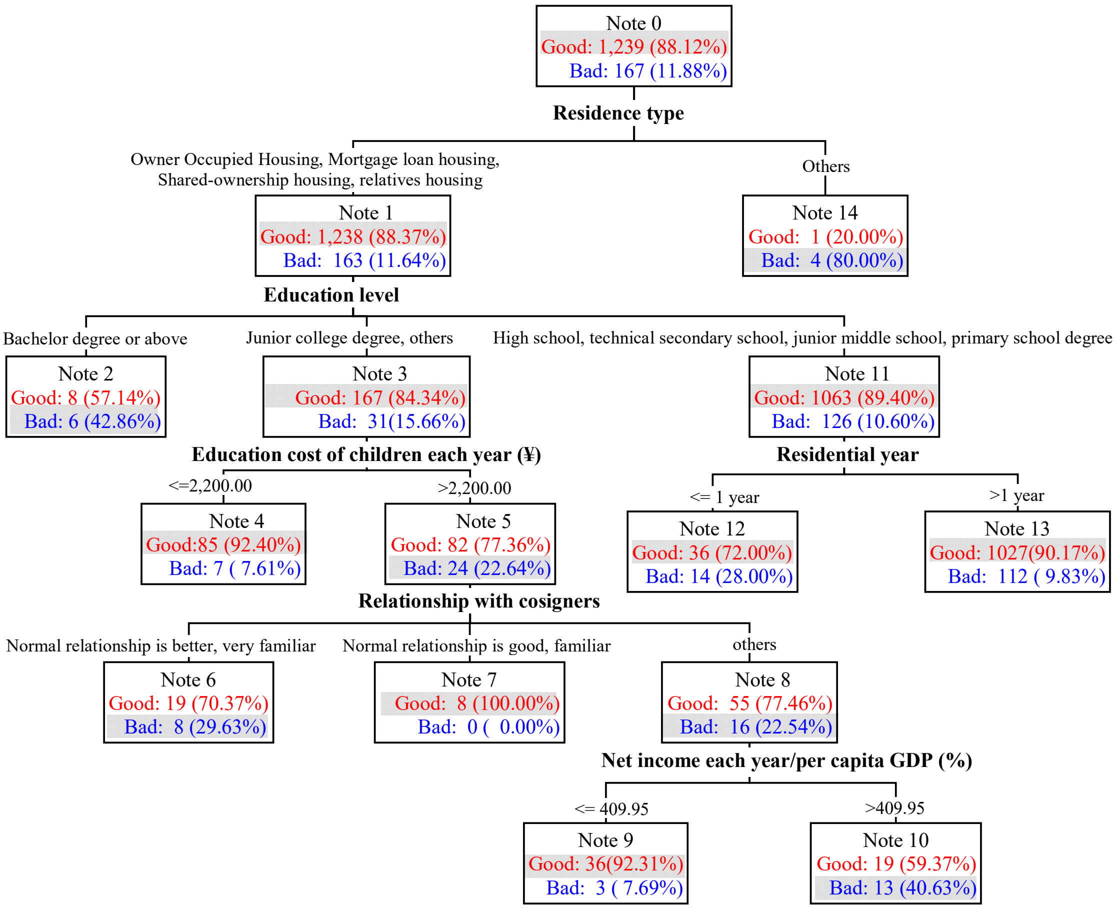


Figure 2: Decision tree of C5.0 method

(5) *Relationship with cosigner*s influences on the default state of customers who satisfy the following conditions that"residence Type Is owner occupied housing etc., and education level is junior college degree, and education cost of children each year is more than 2000 RMB". Obligors who have a "very familiar" relationship with his consigners is less likely to default. It's due to the fact that if the cosigners are obligor's good friends, the possibility of default is even higher, because people with close relationships tend to take risks for fraud together. On the contrary, if the relation with cosigners is not so close, the obligor and the cosigners are less likely to conspire to fraud.

(6) "*Net income each year/per capita GDP* affects the customer who satisfies the following conditions that "education level is junior college degree, and education cost of children each year is more than 2000 RMB. The relationship with cosigner is unclear". Customers with higher net income likely invest in high risk project, and the possibility of default is also high.

7. Conclusion

(1) In this paper, 2044 agricultural microfinance customers in 28 provinces in China are empirical samples. Through stratified sampling, 70% are used as training samples and 30% are test samples. C5.0

and CHAID and C&RT decision tree models are used to establish the credit scoring model and select the key feature affecting credit scoring of farmers, the results show that the credit prediction accuracy of C5.0 is obviously higher than CHAID and C&RT. Therefore, we use the result of C5.0 to determine the key features.

(2) From the 44 primary features, six key features are selected by C5.0 models, which have significant influence on credit scoring of farmers, namely, education level, net income each year/per capita GDP, education cost of children each year, Residence type, residential year, relationship with cosigners.

(3) These key features selected by C5.0 method can help in improving the credit scoring system for Chinese farmers with a certain degree of significance. At the same time, this paper makes up for the lack of credit scoring studies of Chinese farmers in the existing researches.

References

[1] G. E. Pinches, K. A. Mingo, A multivariate analysis of industrial bond ratings. The journal of Finance, 28 (1973) 1-18.

[2] L. McAdams, How to anticipate utility bond rating changes. The Journal of Portfolio Management, 7 (1980) 56-60.

[3] E. Pishbahar, M. Ghahremanzadeh, M. Ainollahi, et al., Factors Influencing Agricultural Credits Repayment Performance among Farmers in East Azarbaijan Province of Iran. Journal of Agricultural Science and Technology, 17 (2015) 1095-1101.

[4] M. B. Karan, A. Ulucan, M. Kaya, Credit risk estimation using payment history data: a comparative study of Turkish retail stores . Central European Journal of Operations Research, 21 (2013) 479-494.

[5] J. A. Afolabi, Analysis of loan repayment among small scale farmers in Oyo State, Nigeria. Journal of Social Sciences, 22 (2010) 115-119.

[6] A. M. Karminsky, E. Khromova. Extended Modeling of Banks' Credit Ratings. Procedia Computer Science, 91(2016) 201-210.

[7] R. Geng, I. Bose, X. Chen. Prediction of financial distress: An empirical study of listed Chinese companies using data mining . European Journal of Operational Research, 241 (2015) 236-247.

[8] S. Figini, P. Giudici. Statistical merging of rating models . Journal of the Operational Research Society, 62 (2011) 1067-1074.

[9] C. Bai, B. Shi, F. Liu , et al., Banking Credit Worthiness: Evaluating the Complex Relationships, Omega, (2018) 1-13. Article in Press. doi: 10.1016/j.omega.2018.02.001.

[10] B. Shi, B. Meng , H. Yang, J. Wang , et al., A Novel Approach for Reducing Attributes and Its Application to Small Enterprise Financing Ability Evaluation, Complexity, 2018 (2018) 1-17. doi:10.1155/2018/1032643.

[11] Y. Zhang, G. Chi, A credit rating model based on a customer number bell-shaped distribution, Management Decision, 56 (2018) 987-1007. doi: 10.1108/md-03-2017-0232.

[12] A. Petropoulos, S. P. Chatzis, S. Xanthopoulos, A novel corporate credit rating system based on Student's-t hidden Markov models. Expert Systems with Applications, 53 (2016) 87-105

[13] R. C. Hwang, Forecasting credit ratings with the varying-coefficient model. Quantitative Finance, 13 (2013) 1947-1965.

[14] B. Shi, N. Chen, and J. Wang, A credit rating model of microfinance based on fuzzy cluster analysis and fuzzy pattern recognition: Empirical evidence from Chinese 2157 small private businesses. Journal of Intelligent & Fuzzy Systems, 31(2016) 3095-3102.

[15] B. Shi, J. Wang, J. Qi, et al., A novel imbalanced data classification approach based on logistic regression and Fisher discriminant. Mathematical Problems in Engineering, 2015 (2015).

[16] B. Shi, H. Yang, J. Wang, et al., City green economy evaluation: Empirical evidence from 15 sub-provincial cities in China. Sustainability, 8 (2016), 1-39.

[17] S. Akkoç. An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data . European Journal of Operational Research, 222 (2012) 168-178.

[18] Z. Huang, H. Chen, C. J. Hsu, et al., Credit rating analysis with support vector machines and neural networks: a market comparative study. Decision support systems, 37 (2004) 543-558.

[19] K. Kim, H. Ahn, A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach. Computers & Operations Research, 39 (2012) 1800-1811.

[20] L. Cao, L. K. Guan, Z. Jingqing, Bond rating using support vector machine. Intelligent Data Analysis, 10 (2006) 285-296.

[21] P. Hájek, Municipal credit rating modeling by neural networks. Decision Support Systems, 51 (2011) 108-118.

[22] K. Shin, I. Han. Case-based reasoning supported by genetic algorithms for corporate bond rating. Expert Systems with applications, 16 (1999) 85-95.

[23] P. Hajek, K. Michalak. Feature selection in corporate credit rating prediction. Knowledge-Based Systems, 51 (2013) 72-84.

[24] Wu, Xindong, et al. Top 10 algorithms in data mining. Knowledge and information systems, 14 (2008): 1-37.

[25] S. Pang, J. Gong, C5. 0 classification algorithm and application on individual credit evaluation of banks. Systems Engineering-Theory & Practice, 29 (2009): 94-104.

[26] G. P. Siknun , I. S. Sitanggang, Web-based Classification Application for Forest Fire Data Using the Shiny Framework and the C5. 0 Algorithm. Procedia Environmental Sciences, 33 (2016) 332-339.

[27] B. Miller, M. Fridline, P. Y. Liu, et al., Use of CHAID decision trees to formulate pathways for the early detection of metabolic syndrome in young adults. Computational and mathematical methods in medicine, 2014 (2014).

[28] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen,Classification and regression trees, CRC press, 1984.