RESEARCH OF CREDIT RISK OF COMMERCIAL BANK PERSONAL LOAN BASED ON ASSOCIATION RULE

Zhang Zenglian
School of Economics and Management, University of Science and Technology Beijing, Beijing, China

Keywords: Commercial bank, Personal loan, Credit risk evaluation, Apriori.

Abstract: Guard against financial risks, reduce bad loans, increase the ability to identity risk of commercial banks, the key is risk warning. In view of the increasing proportion of personal loans in banking business, it is particularly important to warning personal loans credit risk. Commercial bank lending itself is a complex nonlinear system, using general linear theory is difficult to objectively reflect the laws of this, this paper uses association rule. Personal loan credit index first constructed, and then use apriori algorithm to extract rules. Results showed that apriori algorithm plays an important role in identifying risk in personal loans.

1 INTRODUCTION

With China's rapid economic development, financial markets are maturing and mature, the phenomenon of individuals as financiers to participate in financial markets has become increasingly common. Experience from the commercial banks, if we can properly address the credit risk management and control, personal loans will be a rewarding business. Individual loan risk management includes three aspects: risk assessment (ie assessment of the applicant's repayment ability and credit repayment to decide whether to grant the loan), repayment tracking, breach of contract, in which risk assessment is essential. At present, relevant information based on the applicant to rate, but also the relevant rules need to decide whether to grant loans. I let historical data of personal loans default, use Apriori algorithm to extract association rules for credit risk of personal loans.

For construction of early warning model of credit risk, current methods are widely used multiple discriminant analysis (MDA), logistic regression discriminant analysis, neural network analysis and other methods. Altman (1968) developed credit risk Z-score model determined by a number of variables (Altman et al., 1977), Martin (1977) used Logistic and the MDA method to predict bank failures (Martin D., 1977), Ohlson (1980) used Logistic analysis the relationship between two types of bankruptcy errors and split points, and achieved certain results (Ohlson J., 1980). In recent years, neural network technology as a self-organizing, parallel processing and fault tolerance, etc., more and more people's attention, Turban (1996) discussed the neural network in the bank loan credit risk management applications (Altman et al., 1994), most research results showed that neural network was superior to traditional statistical methods (Dan and Mark, 2000). In China, represented by Wang Chunfeng scholars used linear discriminant method, logistic discriminant analysis and neural network models and other methods to assess commercial bank credit risk (Tian Chunyan, 2006). However, these studies focused on business loans, commercial banks cannot meet our need for personal loan default warning.

Based on the above considerations, this paper uses Apriori algorithm for credit risk analysis of personal loans in commercial banks. Association rules for market basket analysis first, this article will introduce it to the field of risk assessment information in order to tap the personal loans of commercial banks, the most significant association rules, and analyzes the interaction between the factors.

Zenglian Z..
RESEARCH OF CREDIT RISK OF COMMERCIAL BANK PERSONAL LOAN BASED ON ASSOCIATION RULE.
DOI: 10.5220/0003413101290134
In *Proceedings of the 13th International Conference on Enterprise Information Systems* (ICEIS-2011), pages 129-134
ISBN: 978-989-8425-53-9

2 CONSTRUCTION OF WARNING MODEL FOR COMMERCIAL BANK PERSONAL LOAN DEFAULT RISK

Commercial bank personal loans default risk warning model is to analyze the characteristics of individual loans to impact on the probability of default. Dependent variable is whether the breach of contract, 1 default, 0 compliance. Default means an individual cannot repay the loan principal and interest on due, including subprime loans, doubtful loans and loss loans. Subprime loan is borrower's repayment ability apparent problem, totally dependent on their normal income cannot guarantee full repayment of loan principal and interest, even if the guarantee may also cause some loss; Doubtful loan is borrower cannot full repay the loan principal and interest, even if the collateral or security, but also sure to cause greater loss; Loss loan is taking all possible measures and all necessary legal procedures, the loan principal and interest is still not recovered, or can only recover a very small part. Compliance refers to a personal debt to maturity, including the normal loans and interest loans. Normal loan is the borrower to fulfill contract, there is not enough reason to doubt the timely and full repayment of loan principal and interest. The interest loan is although the borrower has ability to repay the loan principal and interest currently, but there are some possible negative impact factors on the repayment. To understand default standards, the key is to grasp the core of possibility. Different levels of inherent loans risk, the repayment probability is different. The classification of loan risk divided into the breach of default and undefault by the possibility of repayment, and in order to reveal the true value of the loan.

Personal loan risk including: qualification of unqualified borrowers; borrowers to apply the information false or non-compliance or not complete. Personal credit information system is not perfect now, banks cannot fully assess the borrower's credit and solvency, the borrower may be intentional fraud, fraudulent bank loans through forgery of personal credit information, the bank suffered financial loss; Risk posed by repayment ability of borrowers reduce. There are many personal loans are long-term loans, the borrower's repay ability decline is very likely to occur, could be transformed into the bank's loan risk. Personal loan default factor is related to characteristics of individual loans, including age(A), education level(B), length of service(C), residence(D), family income(E), loan income ratio(F), credit card debts (G), other debts(H), sex, the value of fixed assets, loan term, whether mortgage, the family structure. Education levels are usually inversely proportional to the individual loan default, the higher the education level, the smaller the likelihood of default; The relationship between age and the default may change with age; the higher length of service, the probability of default should be smaller; The longer the Living, the possibility of default should be smaller; the higher family income, the probability of default should be smaller; the higher loans income ratio, the probability of default should be higher; the more credit card debts, the higher probability of default; the higher other debt, the more likelihood of default; Women prefer stable, less likely to select default than men; The more value of fixed assets, the probability of default should be smaller; the longer the loan period, the greater the likelihood of default; The higher value of collateral, the smaller likelihood of default; usually the more stable family structure, the smaller likelihood of default.

3 APRIORI ALGORITHM OF ASSOCIATION RULES

Association rules is to identify the data at a time or things will appear: If item A is a part of the event, the item B also appears in the probability of event X%. Association rules with a specific set of conditions linked to conclusions. Association rules algorithm to automatically find those visualization techniques can be found through the association, such as web nodes, the advantage may exist in the data associated with any attribute, it tries to find out the number of rules, each rule can draw a appropriate conclusions; The disadvantage is that it can be very large, in an attempt to find model search space, will cost a long time. It uses a generation-test method to find the rules - simple rules originally generated, and was the data set proved to be effective. Good rules are stored, all the rules are subject to different constraints, and then be specialized. Specialization is a condition of accession to the rules of procedure. These new rules are then the data proved to be effective, then the process is repeated to store to find the best or the most significant rules. Users often a prerequisite for the number of rules that may make some restrictions. Or an effective indexing mechanism based on

information theory based on various technologies, the existence of many rules is often used to compress the search space. This process generated by the map displayed the best rules, but this set of rules cannot be directly used to predict, because there are many rules to different conclusions. Obtained by the association algorithm called association rule model is not refined.

Let I={i_1, i_2, ..., i_m,} is the set of all items, D for the services of a project database transaction is a subset of T($T \subseteq I$). Each transaction has a unique transaction identifier Tid. Let A be a set composed by the project, called the item set. Transaction T contains itemset A, if and only if $A \subseteq T$. Minimum support minsup association rules under which the user must satisfy the minimum support, it represents a set of items set in the statistical sense of the need to meet the minimum. Minimum confidence minconf association rules required that the user must meet the minimum confidence, which reflects the minimum reliability of association rules. Association rule mining is to find a transaction database D, given by the user's minimum support minsup and minimum confidence minconf of association rules. If the item set support given by the user more than the minimum support threshold (minsup), the set is said to frequent itemsets or large item set. Association rules in two steps: the minimum support threshold based on data set D, find all the frequent itemsets; under frequent item sets and minimum confidence threshold to generate all association rules. There are many association rule algorithm.

Apriori algorithm can only handle character variables and outcome variables. Because of its only character attributes, you can use a subset of intelligent technology to accelerate the search speed. It provides five methods for rules choice, using a complex index of programs to effectively handle large data sets. In the implementation of the node before the field type must be fully instantiated. It can keep the number of rules there is no specific limit, can handle up to 32 premises in the rules. Apriori algorithm uses hierarchical order of the search loop method (also called the search step by step iterative method) produces frequent itemsets, which uses frequent k-item set to explore generate (k+1)-itemsets. First, find out the length of one of the frequent itemsets, denoted L1, L1 is used to generate frequent 2-collection of itemsets L2, and used to generate frequent 3-itemsets L3, and so the cycle continues, until you could not find new frequent k-itemsets. Lk need to scan the database to find each one. Obtained using the following formula to calculate the confidence of association rules.

$$confidence(A \rightarrow B) = P(A \mid B) = \frac{\textbf{support\_count(A} \cup \textbf{B)}}{\textbf{support\_count(A)}}$$

Which, support_count(A∪B) that contains the transaction itemsets A∪B number of records, support_count(A) is the set A contains the item number of records of transactions. Rules of frequent item sets generated by the algorithm described as follows:

for all frequent k itemset l_k, k⩾2 do begin
H1={l_k in the rules after the pieces, after the pieces of rule that only one item};
Call ap_genrules(l_k,*H*1);
end;
Procedure ap_genrules(l_k:frequent itemsets, Hm: m projects of pieces after collection)
if(k>m+1) then begin
*H*m+1=apriori_gen(*H*m)
for all hm+1∈H_{m+1} do begin
conf=support(l_k)/support(l_k -h_{m+1});
if(conf≥minconf) then
output rules l_k-*h*m+1→*h*m+1 with confidence=conf and support=support(l_k);

For the existence of a large number of frequent patterns, long-closed mode or the value of the minimum support is small, Apriori algorithm will face the following deficiencies: algorithm will spend a large overhead to deal with particularly large number of candidate sets; multiple scans the transaction database, requires a lot of I/O load.

4 ASSOCIATION RULES IN COMMERCIAL BANK CREDIT EVALUATION OF PERSONAL LOANS

Data from a commercial bank committed to reducing the loan default rate data, including 700 former customers of the financial and demographic information. Credit personal loans will target variable I, the A to H as explanatory variables, its use Clementin 12.0 Apriori algorithm of association rules. Part of the original data in Table 1, the Apriori algorithm can only handle character variables and outcome variables, the first of the original discrete data, discrete criteria in Table 2, the discrete part of the data in Table 3. By Clementin 12.0 after the discrete data of the Apriori algorithm, so that the independent variables in the foregoing paragraph, the dependent variable for the latter, 10% minimum support, minimum confidence of 80%, resulting in the 39 association rules in Table 4.

Table 1: Original data of part commercial banks credit personal loans.

Records	age	ed	employ	address	income	debtinc	creddebt	othdebt	default
1	27	1	0	1	16.00	1.70	0.18	0.09	0
2	24	1	2	1	21.00	0.60	0.03	0.10	0
…	…	…	…	…	…	…	…	…	…
700	30	1	0	11	17.00	3.70	0.45	0.18	1

Table 2: Discretization criteria of commercial banks credit risk data of personal loan.

age		ed		employ		address		income		debtinc		creddebt		othdebt		defau	
20-25	A1	Pr	B1	0	C1	0	D1	1.3-1.9	E1	0.1-1.9	F1	0.01-0.99	G1	0.05-0.99	H1	U	I1
26-29	A2	Hi	B2	1-3	C2	1-3	D2	2.0-2.9	E2	2.0-3.9	F2	1.00-1.99	G2	1.00-1.99	H2	D	I2
30-35	A3	U	B3	4-6	C3	4-6	D3	3.0-3.9	E3	4.0-6.9	F3	2.00-2.99	G3	2.00-2.99	H3		
36-39	A4	P	B4	7-10	C4	7-10	D4	4.0-5.9	E4	7.0-9.9	F4	3.00-4.99	G4	3.00-4.99	H4		
40-49	A5			11-15	C5	11-15	D5	6.0-9.9	E5	10.0-15.9	F5	5.00-9.99	G5	5.00-9.99	H5		
50-56	A6			16-33	C6	16-34	D6	10.0-44.6	E6	16.0-41.3	F6	10-20.56	G6	10-20.56	H6		

Table 3: Part of discrete data of commercial banks credit personal loans.

Records	age	ed	employ	address	income	debtinc	creddebt	othdebt	default
1	A2	B1	C1	D3	E1	F1	G1	H1	I1
2	A2	B1	C1	D2	E1	F1	G1	H1	I1
…	…	…	…	…	…	…	…	…	…
700	A4	B2	C1	D5	E3	F1	G1	H1	I2

(1) Age has greater impact on credit risk. Rules 1 to 5 by the rules shows that, with age, increase the solvency of individual credit risk reduction. 4 shows the 36-39 rule, the customer, with 81.98% degree of confidence that it will not default on their loans. We can see from the rule 5, 40-49-years-old client, a 80.84% degree of confidence that it will not default on their loans. If it is older, and credit card loans is very low (rule 1), a higher level of confidence is not in arrears (94.52%).

(2) Education had little influence on credit risk. Shows that by the rules 6, middle and the following culture, and credit card loans is very small (less than 100 yuan) customers are 85.31% confidence level will not default on their loans. This is as expected, generally considered higher education, the better the credit. But the data shows that low levels of education, other debts or credit card loans with little or less seniority or length of residence, the client's credit better. This suggests that bank lending, it does not have to select the highly educated customers.

(3) Years of service have a greater impact on credit risk. Rules 7-14 shows that with the length of service growth, increased personal solvency, credit risk reduction. Rule 9 shows that length of service among the customers in the 16-33, with 84.48% degree of confidence that it will not default; shows by rule 12, length of service among the customers in the 11-15, with 91.23% confidence level that the Not in arrears; by the rule 13 shows, length of service between customers in 7-10, with 81.25% degree of confidence that it will not default. If the relatively long length of service, and low levels of education or credit card loans is very low, no higher degree of confidence in arrears.

(4) Residence has greater impact on credit risk. Showing by rule 15-19, as the residence of growth, increased personal solvency, credit risk reduction. Rule 17 shows that, when the residence time in between 16-34 years, with 87.61% degree of confidence that it will not default; by rule 18 shows, when the residence time in between 7-10 years, with 85.42% of the degree of confidence that it will not default. If you live a relatively long period, and the low level of education or credit card loans is very low, there are substantial grounds for believing that it will not default.

(5) Income has greater impact on credit risk. Rules 20-24 show that, as income levels increase, individuals enhance the solvency and credit risk reduction. Rule 23 shows that, when the income of 0.06-0.099 million, 81.55% degree of confidence that it will not default; year income 0.03-0.039 million, 81.25% degree of confidence that it will not

default. If the income is relatively large, and the low level of education or credit card loans is very low, there are substantial grounds for believing that it will not default.

(6) Loan income ratio has greater impact on credit risk. We can see by rules 25-31, loan income is relatively low, less pressure on individuals to pay, credit risk reduction. Rule 27 shows, when the loans than in the 2-3.9 times revenue, 88% confidence level that it does not default; by rule 29 shows, when the loans than in the 4-6.9 times revenue, with 84.85% confidence level I believe it will not default; by rule 31 shows, when the loans than in the 7-9.9 times revenue, with 80.65% degree of confidence that it will not default.

Table 4: Association rules generated by Apriori algorithm.

Rules	After items	Before items	Support	Confidence
1	default = I1	age = A5 and creddebt = G1	10.43	94.52
2	default = I1	age = A5 and ed = B1	11.71	89.02
3	default = I1	age = A3 and creddebt = G1	13.57	86.32
4	default = I1	age = A4	15.86	81.98
5	default = I1	age = A5	23.86	80.84
6	default = I1	ed = B1 and creddebt = G1	30.14	85.31
7	default = I1	employ = C6 and ed = B1	10.14	92.96
8	default = I1	employ = C4 and creddebt = G1	12.86	92.22
9	default = I1	employ = C6	16.29	91.23
10	default = I1	employ = C5 and ed = B1	10.29	87.5
11	default = I1	employ = C3 and creddebt = G1	10.71	85.33
12	default = I1	employ = C5	16.57	84.48
13	default = I1	employ = C4	20.57	81.25
14	default = I1	employ = C4 and ed = B1	11.43	81.25
15	default = I1	address = D4 and ed = B1	11.29	91.14
16	default = I1	address = D4 and creddebt = G1	11.29	88.61
17	default = I1	address = D6	16.14	87.61
18	default = I1	address = D4	20.57	85.42
19	default = I1	address = D3 and creddebt = G1	10.71	81.33
20	default = I1	income = E3 and creddebt = G1	11.29	89.87
21	default = I1	income = E2 and ed = B1 and creddebt = G1	11.43	83.75
22	default = I1	income = E4 and ed = B1	10.43	82.19
23	default = I1	income = E5	14.71	81.55
24	default = I1	income = E3	18.29	81.25
25	default = I1	debtinc = F3 and ed = B1	12.71	91.01
26	default = I1	debtinc = F3 and ed = B1 and creddebt = G1	11.0	90.91
27	default = I1	debtinc = F2	10.71	88.0
28	default = I1	debtinc = F4 and ed = B1	10.14	85.92
29	default = I1	debtinc = F3	23.57	84.85
30	default = I1	debtinc = F3 and creddebt = G1	19.71	84.06
31	default = I1	debtinc = F4	17.71	80.65
32	default = I1	creddebt = G1	54.43	81.89
33	default = I1	othdebt = H2 and ed = B1 and creddebt = G1	10.43	87.67
34	default = I1	othdebt = H1 and ed = B1 and creddebt = G1	11.57	85.19
35	default = I1	othdebt = H1 and creddebt = G1	20.43	83.22
36	default = I1	othdebt = H2 and ed = B1	15.29	83.18
37	default = I1	othdebt = H1 and income = E2	10.14	83.10
38	default = I1	othdebt = H2 and creddebt = G1	17.71	83.06
39	default = I1	othdebt = H3 and ed = B1	10.71	80.0

If the loan is relatively low income and lower education level, or credit card loans is very low, there are substantial grounds for believing that it will not default.

(7) Credit card lending is inversely proportional to credit risk. Rule 32 shows, credit card loans of less than 100 yuan, a 81.89% degree of confidence that it will not default.

(8) Other debts have greater impact on credit risk. we can see by rules 33-39, the less other liabilities, credit risk is smaller. When other less debt, and credit card debt low level of education or less, the customer usually does not breach of contract.

5 CONCLUSIONS

Unrefined 39 rules by a representative of association rules modeling the rules found in the node, which contains the rules extracted from the data information, these rules are not designed to directly predict, but to provide useful information on bank loans . We can see by these rules, the older (more than 30 years of age), length of service longer, living longer lives, higher income, low income loans, credit cards, small loans, other debt a few customers are high-quality customers. Limited to the index system of loans and the lack of sample data, this paper considers only the impact of individual loan default the main factors. Limited to the index system of loans and the lack of sample data, this paper considers only the impact of individual loan default the main factors. In reality, the credit rating personal loans banks will also be required to consider some other factors, such as use of the loans and so on, to get more objective results. Commercial banks should be based on their own circumstances, such as market positioning, risk tolerance and other factors of risk and establishing early warning model to their own circumstances, to update the sample database, on the one hand can provide a more accurate assessment of the borrower's credit risk, on the other hand can also be integrated master The Bank's operating status, for subsequent development banks, the overall planning and layout to provide information support.

In order to guard against default risk of personal loans, banks need to spare "three investigations work", effort to reduce losses caused by information asymmetry. Asymmetric information easily lead to adverse selection and moral hazard, therefore, to do "loans before, loads middle, loans after" three examinations work, effectively reducing the information asymmetry caused by the credit default risk. Investigation stage of the loan, to strengthen the borrower's eligibility review to ensure the qualifications of the borrower of the main provisions of commercial banks; Meanwhile, according to the borrower to submit proof of income, books, tax return and other materials, combined with its occupation, job security and other information on the solvency of the borrower income and comprehensive analysis and use of personal credit information database accurately The borrower's overall credit evaluation. Strengthen dynamic monitoring, early warning. Customer's credit rating is a dynamic change with a variety of factors, dynamic monitoring of these factors can not only advance warning, it is important to have sufficient time to seek countermeasures.

REFERENCES

Altman E. I. et al. Zeta analysis: Anew model to identify bankruptcy risk of corporations[J]. *Journal of Banking and Finance*, 1977, 1(1):29-54.

Martin D., Early warning of bank failure: A logit regression approach[J]. *Journal of Banking and Finance*, 1977, 1(3):249-276.

Ohlson J., Financial ratios and the probabilistic prediction of bankruptcy [J]. *Journal of Accounting Research*, 1980, 18(1):109-131.

Altman E. I., Marco G., Varettl F., Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks [J]. *Journal of Banking and Finance*, 1994,18(3):505-529.

Dan M. C. & Mark G. R., A comparative analysis of current credit risk models [J]. *Journal of Banking and Finance* 2000, 24(1): 59-117.

Zhang. On the personal loan business in the commercial banking risk prevention [J]. *Financial Development*, 2008 (8).

Tian Chunyan and so on. Bank loan approval personal knowledge of the rule base refinement [J]. *Management*, 2006 (9).

Xue Feng. Rough Sets - neural network system in the commercial bank loan classification application [J]. *Systems Engineering Theory & Practice*, 2008 (1).

Xie Bangchang. Clementine data mining application practices [M]. *Beijing: Mechanical Industry Press*, 2008.